

NATIONAL CONFERENCE ON
RECENT ADVANCEMENTS IN COMPUTER SCIENCE
(CONRACS 2019)

26 – 28 July 2019

Sponsored by DST - NewDelhi.
Technically Co-Sponsored by CSI Hyderabad.

CONFERENCE PROCEEDINGS

(ISBN No: 978-93-88808-31-6)

Chief Editor

Dr. SambasivaRao Baragada

M.Sc., M.Phil., Ph.D (Computer Science)

Assistant Professor of Computer Science

BJR Government Degree College, Narayanguda, Hyderabad.



Organized by

**Mahatma Gandhi National Institute of Research
and Social Action (MGNIRSA)**

HYDERABAD

Telangana State, INDIA – 500034.



**Mahatma Gandhi National Institute
of Research and Social Action
(MGNIRSA)
HYDERABAD
TELANGANA STATE - 500034**

Dr. D. Suresh
Registrar

MESSAGE

MGNIRSA is organizing a three-day national conference on **RECENT ADVANCEMENTS IN COMPUTER SCIENCE (CON-RACS 2019)** during 26 – 28 July 2019. In this connection, the MGNIRSA is publishing conference proceedings showcasing the outstanding articles as the topic is confined to the cutting-edge research and scientific advancements occurred in the field of computer science.

The participation of a large number of eminent thinkers, scientists, educationists, and students on this joyful occasion would have a meaningful impact on the partakers and would also be an important contribution in the direction of societal development.

I take this occasion to accolade the organising department and all the faculty members of the college for their contribution towards the smooth running of this national conference and express my best wishes to them for all their future endeavours.

[Dr. D. Suresh]



**Mahatma Gandhi National Institute of
Research and Social Action (MGNIRSA)**
HYDERABAD
TELANGANA STATE - 500034

Dr. D. Sirisha
Convenor, CONRACS 2019

MESSAGE

I wish to congratulate all the faculty members of the MGNIRSA, Hyderabad, for their valuable support in organizing a three-day national conference on **RECENT ADVANCEMENTS IN COMPUTER SCIENCE (CON-RACS 2019)** during 26 – 28 July 2019.

The proposed conference would create a platform to share the latest research findings in the computational technologies being carried in the country and throw light on the finding various solutions of the problems unsolved. The institute is highly indebted to the DST, ISRO, and INCOIS for their association in organizing this conference. I convey my best wishes to all the participants and I feel proud to be associated as Convenor of CON-RACS 2019 with conglomeration of experts.

[Dr. D. Sirisha]

CON-RACS (2019) PROGRAMME SCHEDULE

26.07.2019 (Day-1)

Time	Session
8:00AM – 10:00AM	Registration
10:00AM – 11:00AM	Inauguration Session
11:00AM – 11:15AM	Tea Break
11:15AM – 01:00PM	Technical Session1
01:00PM – 02:00PM	Lunch Break
02:00PM – 03:15PM	Technical Session2
03:15PM – 03:30PM	Tea Break
03:30PM – 05:00PM	Technical Session3

27.06.2019 (Day-2)

Time	Session
10:00AM – 11:15AM	Technical Session4
11:15AM – 11:30AM	Tea Break
11:30AM – 01:00PM	Technical Session5
01:00PM – 02:00PM	Lunch Break
02:00PM – 03:15PM	Technical Session6
03:15PM – 03:30PM	Tea Break
03:30PM – 05:00PM	Technical Session 7

28.07.2019 (Day-3)

Time	Session
10:00AM – 11:15AM	Workshop on Cyber Security
11:15AM – 11:30AM	Tea Break
11:30AM – 01:00PM	Workshop on Data Science
01:00PM – 02:00PM	Lunch Break
02:00PM – 03:15PM	Workshop on Machine learning
03:15PM – 03:30PM	Tea Break
03:30PM – 05:00PM	Valedictory Session

GUESTS OF HONOURS & SESSION CHAIRS

❖ **Dr. N. Srinivasa Rao**

Scientist-D,
Indian National Centre for Ocean Information Services
(INCOIS), Ministry of Earth Sciences (MoES)
Hyderabad.

❖ **Dr. D. Giri Babu**

Scientist-SF, Regional Remote Sensing Centre – West,
Indian Space Research Organization (ISRO), Jodhpur.

❖ **Dr. K. L. Raju**

Director, Logic Designer Pvt. Ltd.

❖ **Dr. Sita Rambabu**

Chairman, Computer Society of India (CSI Hyderabad
Chapter)

❖ **Prof. O.B.V. Ramaniah**

Dept. of Computer Science & Engineering
Jawaharlal Nehru Technological University (JNTUH)
Hyderabad.

❖ **Sri. S.P Vighneshwar**

Scientist-C,
Indian National Centre for Ocean Information Services
(INCOIS), Ministry of Earth Sciences (MoES)
Hyderabad.

❖ **Sri. N. Kiran Kumar**

Scientist-D
Indian National Centre for Ocean Information Services
(INCOIS), Ministry of Earth Sciences (MoES)
Hyderabad.

- ❖ **Sri. Ch. E. Sai Prasad**
Scientist, Computer Forensics Division
Central Forensic Science Laboratory
Hyderabad.
- ❖ **Sri P. Srinivas Murthy**
Scientist, Atomic Minerals Directorate for Exploration
& Research, Dept. of Atomic Energy,
Hyderabad.
- ❖ **Prof. A.V. Krishna Prasad,**
Dept. of Computer Science & Engineering,
MVSR Engineering College,
Hyderabad.
- ❖ **Dr. N. Uday Bhaskar**
Associate Professor and Head,
Dept. of Computer Science,
Government Arts and Science College, Anantapur, Andhra
Pradesh.
- ❖ **Dr. HR Chennamma**
Assistant Professor, Dept. of Computer Applications
JSS Science and Technology University
Mysuru.
- ❖ **Prof. P. Chitti Babu**
Dept. of Computer Science & Engineering
Annamcharya Institute of Science and Technology
Hyderabad
- ❖ **Dr. M. Uma Devi**
Assistant Professor, Dept. of Computer Science & Engineering
Universal College of Engineering,

Guntur, Andhra Prasad.

❖ **Smt. G. Geetha**

Scientist-B,

Indian National Centre for Ocean Information Services
(INCOIS), Ministry of Earth Sciences (MoES)

Hyderabad.

List of Original Papers Submitted to CONRACS 2019

SNO	TITLE OF THE PAPER	AUTHOR(S)	PAGE NO.
TECHNICAL SESSION-1			
1.	BLOCKCHAIN IMPLEMENTATION FOR THE MUTUAL FUND RTA	VIJAYA KILLU MANDA, VEDAVATI KATNENI, SATYA PRAKASH YAMIJALA	1
2	AN ANALYSIS OF PREVENTION OF FAKE NEWS_PAPER_CONFERENCE	R KAMALAKAR, KISHORE KUMAR GAJULA	9
3	BLOCKCHAIN BASED TRANSACTION SYSTEMS	ANNAPAREDDY VN REDDY, CHITTA VENKATA PHANI KRISHNA	21
4	BLOCKCHAIN FUNDAMENTALS PRESENTATION	M DIVYA SREE	33
5	BLOCKCHAIN TECHNOLOGY IN SUPPLY CHAIN MANAGEMENT	VARADA PALLY VINAY REDDY	37
TECHNICAL SESSION-2			
6	A NOVEL ARCHITECTURE TO CLOUD BASED SCADA SYSTEMS	RAMANA DUGYALA, N HANUMAN REDDY	51
7	CERTIFICATELESS DATA INTEGRITY CHECKING AND DATA SHARING WITH SENSITIVE INFORMATION HIDING IN CLOUD STORAGE-CONVERTED	K NAVEEN, C SHOBHA BINDU	58
8	SECURE AND EFFICIENT ACCESS CONTROL FOR MULTI-AUTHORITY CLOUD STORAGE	P VENKATESWARA RAO, V SUCHITRA, P MANEPPA	69
9	REVIEW ON IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE FOR OPTIMAL TASK SCHEDULING IN CLOUD ENVIRONMENTS	RAMAKRISHNA GODDU, KIRAN KUMAR REDDI	79
10	CLUSTERING OF ERROR DATA_DATA MINING	KARUNALA ARUN RAJ	95

		BAPUJI, B ANAND KUMAR, VORSU MALLAYA, A VINAYA BABU	
11	PROVIDING AUTHENTICATION FOR BIG DATA IN CLOUD COMPUTING	M VIJAYA LAKSHMI, M SRI LAKSHMI	105
12	A COMPARATIVE ANALYSIS OF HIGH-PERFORMANCE COMPUTING ON THE CLOUD FOR SCIENTIFIC APPLICATIONS	CHINTHI REDDY PRAKASH, K RAJESHWARA RAO	110
13	FUZZY KEYWORD SEARCH OVER ENCRYPTED DATA IN CLOUD COMPUTING	BOGA JAYARAM MYSA KALYAN CHAKRAVARTHY	121
14	WHY PREDICTIVE ANALYTICS IS KEY FOR EACH ONE? AND THIS TOOL IS THE PERFECT USE CASE FOR CLOUD COMPUTING	K MADAN MOHAN, PROF. P. PREMCHAND	136
TECHNICAL SESSION-3			
15	A SURVEY-ON-MULTIMEDIA-INFORMATION-RETRIEVAL-BASED-ON-ANNOTATION	H R CHENNAMMA	147
15	MALWARE DETECTION AND CONTROL IN DECENTRALIZED PEER TO PEER NETWORK	S GANESH MOORTHY, ABHIJIT P, ASHAR HENSON H	159
16	SELECTIVE JAMMING DROPPING INSIDER ATTACKS IN WIRELESS MESH NETWORKS	S GANESH MOORTHY, UJWAL U, SUJITH V	168
TECHNICAL SESSION-4			
17	ADVANCED TECHNIQUES FOR OUTLIER DETECTION IN HIGH DIMENSIONAL DATA	N JAYANTHI, B VIJAY BABU, N SAMBASIVARAO	179
18	MARKET DYNAMICS OF BITCOIN CURRENCY JULY 2019	DINESH K, JANET	186
19	A SURVEY ON DIAGNOSIS OF DENTAL CAVITIES	P T S S ROOPESH,	198

	DATA USING BIG DATA TECHNOLOGIES	ASADI SRINIVASULU, P V VENUGOPAL PRUDHVI	
20	OBSERVATIONAL ANALYSIS FOR QUALITY OF SOIL ORGANIC MATTER USING AGRICULTURAL DATA SCIENCE	GULLEDMATH SANGAYYA, ARUL KUMAR V	209
21	THE STUDY OF DATA MINING IN HEALTH CARE SECTOR	K RAJESWARA RAO , B MAHENDAR REDDY, K VENKATESWARA RAO, S BALAKRISHNA REDDY	226
22	TRENDS ENABLED IN DATA SCIENCE	KADARI SRINIVASA RAO	236
23	AN INSIGHT ON DATA SCIENCE CVK	C V KRISHNA VENI	242
24	LOADING, SEARCHING AND RETRIEVING DATA FROM CLUSTERED DATA NODES ON HDFS	M SREE RAMA MURTHY, N NAGAMALLESHWARA RAO	257
25	SURVEY ON VARIOUS PROCESS MODELS IN SOFTWARE DEVELOPMENT	LALITHA KUMARI, M MANASA	271
26	A COMPLETE SURVEY ON ITERATIVE CLUSTERING METHODS	M VIAYA LAKSHMI, A MANASA	278
27	BIG DATA ANALYSIS	P SOWMYA SREE	292
28	CHALLENGES IN PRODUCT DEVELOPMENT USING SCRUM MODEL - SUNIL REDDY	K SUNIL MANOHAR REDDY, G PRATHIBHA	297
TECHNICAL SESSION-5			
29	A WEB BASED APPROACH FOR HOUSE CONSTRUCTION	B. SANJANA G. YOGITHA VIJAY KUMAR	302
30	DIABETES DISEASE PREDICTION USING DATA MINING	K .SANKEERTHANA , K .LASYA, B.ASHRITHA, V.VIDHU	309
31	DEEP LEARNING APPLICATIONS IN MRI-BASED IMAGE ANALYSIS	VIJAY KUMAR.B B DEEPTHI REDDY	315
32	AN ENHANCED DEEP LEARNING METHOD FOR VIDEO RETRIEVAL	SADDAM HUSSAIN, C SHOBHA BINDU	326
33	FEATURE BASED MULTI LEVEL KEYFRAME SELECTION FOR VIDEO SUMMARIZATION	B SIRISHA, B SANDHYA	337
34	INTENSIFYING THE ACCURACY OF FINGER-VEIN IMAGE IDENTIFICATION USING CONVOLUTIONAL NEURAL NETWORKS	VINEET REDDY K, C SHOBHA BINDU	351

35	IMAGE SIMILARITY WITH COSINE DISTANCE	C SIVA JYOTHI, B SANDHYA	361
36	"INTERNET OF EVERY THING" APPLICATIONS, ITS FUTURE AND CHALLENGES	RAYEES FATHIMA	379
37	INNOVATIVE TECHNOLOGY TO PROTECT FARMERS FROM SNAKE BITE USING IOT	CHENNUR KEERTHIKA REDDY, VADLAPUDI POOJITHA, ARURU GURU GAYATHRI	385
38	INTERNET OF THINGS" APPLICATIONS, ITS FUTURE AND CHALLENGES	RAYEES FATHIMA	390
39	IOT AND SMART HOME BASED ON LI-FI TECHNOLOGY	R SUBRAMANYAM	400
40	A QUICK REVIEW OF INTERNET OF THINGS AND ITS LATEST APPLICATIONS	SELAM BHANUPRAKASH SANE DINESH KUMAR REDDY	406
41	CYBER – PHYSICAL SYSTEMS AND APPLICATIONS	D SUHASINI	409
42	DRUNK DETECTION FOR LOCKING IGNITION	K RAVALI	415
43	E-COMMERCE AN ANALYSIS OF PRESENT TRENDS, CHALLENGES & OPPORTUNITIES	J MALLAREDDY, T R SRINIVAS	425
44	INTERNET OF THINGS – APPLICATIONS AND HURDLES	T R SRINIVAS	436
45	ROBOTICS AND ADVANCED MANUFACTURING	K VINEELA	449
46	SECURE IOT INFRASTRUCTURE PAPER BY K NAGA MAHA LAKSHMI	K NAGAMAHA LAKSHMI	453
TECHNICAL SESSION-6			
47	'MACHINE LEARNING' THE FUTURE TECHNOLOGY	G SHANMUKHA SESA SAI, A S RAMCHARAN	460
48	APPLICATIONS OF BIGDATA IN MACHINE LEARNING	R V GANDHI	467
49	MACHINE LEARNING DATABASE FOR NATURAL LANGUAGE PROCESSING	N JAYANTHI, A RAVI PRASAD	476
50	MACHINE LEARNING APPROACH FOR PREDICTING MEDICAL DIAGNOSIS	S PRABHAKAR, M SUJATHA	480
51	ARTIFICIAL INTELLIGENCE AND ROBOTICS VIA GAMES	K VISHNUVARDHAN REDDY, M ROHITH	492
52	ML INTEGRATION WITH EVERYDAY DATA	FIZA TARANNUM, NEHA ANSARI	498
53	CROANN - CHEMICAL REACTION OPTIMIZATION OF ARTIFICIAL NEURAL NETWORKS FOR EFFECTIVE FORECASTING OF CRUDE OIL PRICES TIME SERIES	SARATH CHANDRA NAIK	502
54	AIRLINE BOT – THE TRANSFORMED EXPERIENCE	RINCY MARIAM THOMAS, SUPRIYA TUNNA, C KISHORE KUMAR REDDY, B V RAMANA MURTHY	521
TECHNICAL SESSION-7			
55	DIGITAL BASED HUMANOID ROBOT(DBHR)	RAVI VINEET SHARMA	532

56	ANDROID BASED SIGN LANGUAGE TRANSLATOR	G. PRANAY GOUD, M. RAVI	534
57	WIRELESS SENSOR NETWORKS APPLICATIONS AND CHALLENGES	G RAJITHA DEVI	542
58	SOCIAL APPLICATIONS IN THE HOME NETWORKS - JAIRAM BOGA	BOGA JAYARAM, MYSA KALYAN CHAKRAVARTHY	551
59	CRITICALITY OF ROLE OF OPTIMIZERS IN MACHINE LEARNING TECHNIQUES	AJEET K JAIN, M SHWETHA	558

Blockchain Implementation for the Mutual Fund RTA

Vijaya Killu Manda¹, Vedavathi Katneni², Satya Prakash Yamijala³

^{1,2}GITAM Deemed to be University, Visakhapatnam, INDIA

³DXC Technologies, Madhapur, INDIA

¹mvkillu@gmail.com, ²vedavathi.katneni@gitam.edu, ³satya.yamijala@gmail.com

Abstract. Registrar and Transfer Agents (RTA or R&T Agents) are important financial intermediaries in the investment asset management industry who can use blockchain technology in their financial transaction record keeping and investor (customer) service activities. This paper examines the operational, banking, financial and non-financial transactions of an RTA for possible cost and process optimization with a blockchain implementation. Enterprise Ethereum Alliance (EEA) standards are considered so that the system is enterprise-grade, scalable and plugs with other vendor systems. It is found that the on-chain and off-chain features of the blockchain can help in document management as well as optimal resources of the RTA.

Keywords: asset management company, financial intermediaries, blockchain architecture, blockchain standard, blockchain layers, offchain computation

1 Introduction

Registrar and Transfer Agents (RTA or R&T Agents) provide vital financial transaction record keeping services and are a one-stop contact point for information on mutual funds (mf) (such as scheme details, statutory information, NAV updates etc.) for the investor. The Asset Management Company (AMC) is already burdened with the scheme design, portfolio management and related activities and hence outsourcing time-critical operational activities such as transaction processing, statement generation and customer support to an RTA makes sense. RTAs have branches across the country and hence AMCs need not open branch offices in the same locations again. Further, an investor visiting an RTA branch can get support of multiple AMCs from a single location. Finally, because RTAs have inter-connectivity with other RTAs, sending consolidated account statement will be a breeze.

Blockchain is spearheading the fintech revolution, the beginning of which we are witnessing right now. While blockchain as a disruption technology eliminates non-value adding “intermediaries”, the future of RTAs is in doubt and is a huge topic for a separate debate. This paper presumes that the role of RTAs is too huge to be eliminated from the system, at least for now, considering that the industry is in early stages of realizing the technological importance. The next steps will be to slowly migrate towards practical implementations before disrupting the “middlemen”. In fact, this is not the first time that the need for RTAs is questioned, or rather, challenged. Back in 2009, financial market

regulator, the Securities Exchange Board of India (SEBI) began allowing mf orders to be routed through the stock exchanges whereby investors can call their stock broker to place their mf transaction request. Funds get adjusted with the stock broker ledger balance and units will be held in demat format thereby reducing the need on an RTA. However, this initiative got lukewarm response back then [5] and the low and dull response remained the same even after a decade. This highlights the indispensable role of an RTA.

This paper explores the various RTA functions and device methods by which blockchain technology can fit in their existing processes and thereby bring in process optimize so as to derive cost cutting and time saving advantages

2. Research Design

2.1 Research Methodology

Being an evolving technology, Exploratory Research Methodology, as discussed by Robson [8], will be used to *find out what is happening* with the blockchain technology, to *seek new insights* (in terms of developments and evolution), to *ask questions* (in regard to applicability of the technology for the RTA industry) and to *access phenomenon* in a new light.

2.2 Data Sources

Being an evolutionary technology, only secondary data sources are used to collect information from journals, yellow papers and official websites.

2.3 Literature Study

Blockchain benefits for the banking and financial industry [1] in general and for sub-sectors such as the mf industry in particular [2] are already well documented. Trends show that it is getting increased acceptance beyond the typical use-case level. However, there is hardly any efforts put on studying RTAs. Clearly, there is a literature gap in regard to blockchain usage by the RTA and this paper attempts to fill it.

2.4 Need for study

While there are no industrial figures in regard to the quantum of mf transactions being processed every day, the industry size metrics would give a broad estimate. Globally, there are 114,000 open-ended regulated mf schemes managing US \$49 trillion as of end of 2017[4]. Back in India, the number of mf folios have almost doubled from 4.03 crore (in Dec 2014) to 8.03 crore (as of Dec 2018). Assets being managed by Indian MFs are now worth Rs. 24.25 trillion (Feb 2019). Increased financial literacy and investment awareness is making investors to remain invested for the long term. As much as 29.7% of equity investors are staying invested for more than 2 years [3].

Financial Regulators (such as SEBI in India), recognize and regulate RTAs as “intermediaries” and grants certificate of registration, does inspection of books and records for compliance etc. SEBI Regulations of 1993, 1996 and 2008 are the primary regulations governing RTA operations in India.

This study is of significance considering the wide applicability of process optimization and time savings that come along with blockchain implementation for the RTAs.

3. Blockchain and the RTAs

Blockchain is a decentralized peer-to-peer network running a virtual machine (such as the Ethereum Virtual Machine (EVM)) with a clear separation between public and private layers wherein blocks containing information are chained (connected) to the previous block on the network. The Enterprise Ethereum has some unique features such as improved speed of processing (throughput), ability to perform transactions in private (off-chain) and membership enforcement (permissioning).

3.1 Operational aspects of RTA

RTA is responsible for safe and accurate maintenance of records, facilitate transaction processing and addressing investor complaints in a timely manner. They are expected to have mechanisms such as maker-checker for ensuring checks and balances in all transactions and to maintain compliance at all times. Being official investor service centres (ISC), they are responsible for investor-facing transactions of a fund.

RTAs perform key operational aspects in the mf transaction cycle. These include allowing customers to transact (purchase, redeem, switch) with the fund either off-line or on-line, issue and redeem mf units with the appropriate NAV, create and update folios reflecting investor transactions, updating unit capital account on a daily basis so that fund accounts / custodian can use this information for determining portfolio value, update the AMC and its fund manager about the inflow and outflow after banking the payment instruments (such as cheques and demand drafts), process dividends and redemptions and sending periodic statements and statutory information to investors [6].

All investor requests (financial or otherwise) are to be timestamped and serial numbered. This timestamping process will differ a little if a payment instrument (such as a

cheque) is accompanied. Blockchain implements timestamping as a part of its core service offering and in an automated method.

With Independent Financial Advisors (IFA) and investors allowed to participate using the special stock exchange platform window – such as the Mutual Fund Service System (MFSS) of the National Stock Exchange (NSE) and BSE StAr of the Bombay Stock Exchange (BSE), RTAs have to make special arrangements for data collection and dissemination in a consistent manner.

Apart from servicing AMCs and investors, RTAs also support and assist Mutual Fund Distributors (MFD) who in-turn can appoint sub-brokers under them. The MFDs are already part of a distribution network called FinNet (in India), an internet-based platform for allowing distributors to allow their investor clients to transact in 35-odd mutual fund schemes.

Clearly, the RTA blockchain need to handle both information and fund flow in an optimal manner.

2.1 Banking Operations of MF

Cash Management Services (CMS) are used in mfs for effective utilization of funds. The various types of accounts being maintained are: Collection Account, investment Account, Redemption Account and Expenses Account. A variety of financial instruments (such as cheques, demand drafts etc.) and banking services (ABSA, NACH of NPCI, ECS etc.) apart from online banking and their modern extensions (such as UPI, wallet payments etc.) are allowed as part of banking transitions of a mutual fund.

2.2 Financial Operations of MF

The mf application form is first contact point between the investor and the MF. RTAs collect, capture, store and maintain a lot of investor information as mandated by SEBI. Transactions are allowed using the transaction slips and a variety of payment instruments are to be accepted. Purchase transactions are to be processed as per time-line. Statement of Account (SoA) has to be sent within 5 business days. Of course, with transactions being processed online, SoAs are often sent within 1 to 2 working days. Blockchain can interfere and can cut down this time to few hours. When blockchain networks interconnect different RTAs, sending a Consolidated Account Statement (CAS) will be on-the-fly. Systematic Investment Plan (SIP) transactions involve processing to be done by the bank. Redemptions

are done keeping in mind on the applicability of exit load of the scheme. Systematic Transfer and Withdrawal (STP / SWP) are to be facilitated.

While cryptocurrencies cannot be right away allowed in mutual funds because of regulatory restrictions, bank transaction requirements can be addressed with blockchain.

2.3 Non-Financial Transactions

For an RTA, non-financial transactions are equally important as financial and banking transactions owing to regulatory requirements of providing timely response to investor requests. The bulk of the non-financial transactions include change of investor details (such as Address, Bank details, Option of a scheme, Name of an individual investor, Corporate name or status, Authorised signatories etc.). All these non-financial requests are to be supported by a request letter and necessary proofs which are scanned and stored permanently in digital format. Blockchain has the ability to store documents on-chain in the form of the document itself or by storing a hash on-chain but keeping the actual document off-chain. Blockchain provides tamper resistance, visibility and decentralized storage abilities in this regard.

2.4 EEA Layered Architecture

In the process of defining and implementing standards, the EEA prepared a five-layered Enterprise Ethereum Architecture Stack. Confirming to the standards have many long-term benefits for the blockchain implementors because it allows scalability, easy communication and plugs-in to systems of other vendors, helps in inter-operability with other disruption technologies (such as IoT, Machine Learning, Artificial Intelligence) etc. These things are important considering clients (such as RTAs and AMCs) indeed have plans to use a mix of various technologies so as to have an edge over their competitors.

At the bottom of the EEA stack is the Network layer where the communication between blockchain nodes are facilitated by two protocols – DEVp2p wire protocol (that is used for establishing and maintaining communication for higher layer protocols) and the Ethereum wire protocol (that is used for exchanging block and transaction information between Ethereum client nodes).

The Core Blockchain layer takes care of storage / ledgering (to store the blockchain state), execution (implements the Ethereum Virtual Machine (EVM) or the Ethereum-flavored WebAssembly [eWASM]) and consensus (by allowing consensus algorithms such as Proof of Stake to run).

The Privacy / Scaling layer carries necessary extensions to allow enterprise-grade deployments at two levels – Level 1 (on-chain scaling implemented at protocol layer level) and Level 2 (off-chain scaling implemented using smart contracts at application protocol layer).

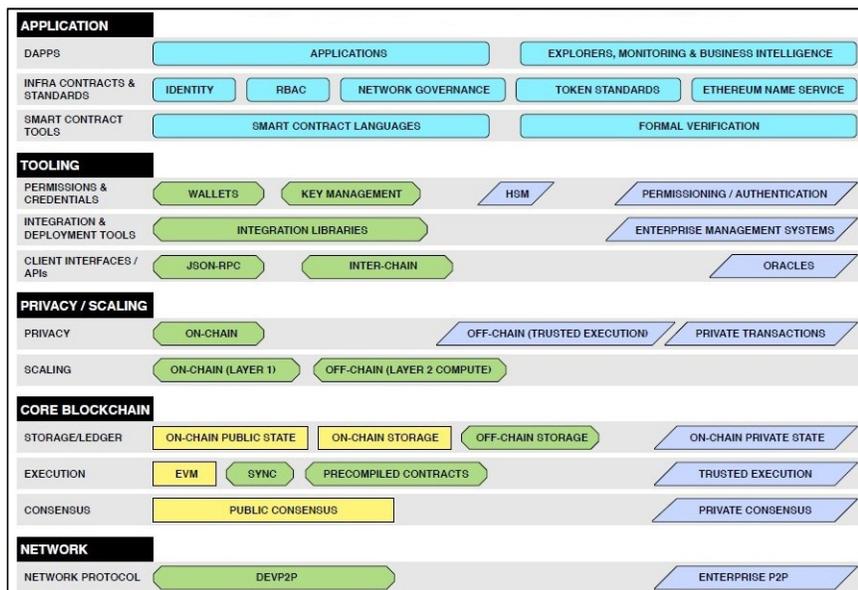


Fig. 1. The Enterprise Ethereum Architecture Stack provides a five-layer standardized implementation of enterprise grade apps. Source: Enterprise Ethereum Alliance

Communication with clients is handled by the API's (such as JSON-RPC API) provided by the Tooling layer for activities such as submitting transactions for execution etc. Compilation and verification of smart contracts will be provided by support to languages such as Solidity, LLL etc.

The Application layer is the top most layer providing highest level services by residing partially or fully outside of the client. Wallet services, Ethereum Naming Services (ENS), node monitoring etc. This layer provides the ground for the running Decentralized Applications (DApps), certain enablers that allow extensions (Infrastructure Contracts and Standards) and support to language parsers, compilers and debuggers for smart contract tools.

While the bottom three layers are common for any enterprise, the top two layers (Tooling and Application) would require special and specific changes for use by the RTAs.

2.3 Off-chain Computation

The EEA Off-chain trusted compute specification [7] resembles a distributed computer wherein the blockchain implementation will have n Member Enterprises (ME). Each ME will have Requestors (that sends a Work Order), EE Client and a number of Workers (that execute a Work Order) running under it that are managed by a worker service (WS), all of which are registered with the main blockchain. The workers are computational resources that take up the responsibility of executing a Work Order by doing the actual computation necessary and returning back the computational result. Imagine the RTA blockchain with ME's being RTA branch office nodes. Each ME will have multiple workers who do transaction processing (such as processing a SIP, a redemption request etc.). When one branch ME is done processing all its work orders and is in idle state, another branch ME can send a work request to the Requestor of the idle ME to get the transaction processed. This way, optimal resource utilization of the blockchain can be done. The work orders are issued either by a DApp or an application smart contract. Either the direct mode or the proxy mode can be used by the Requestor to submit a work order. A Work Order Receipt will be created by the Requestor and are updated by Workers to tell if the transaction processing is a success or a failure. Since all these transactions happen off-chain, higher throughput can be achieved.

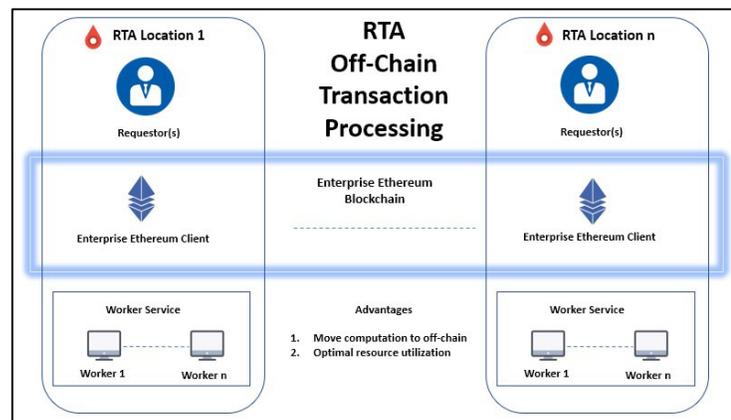


Fig. 2. The Off-Chain Transaction Processing for the RTA is based on EEA Off-chain Trusted Compute Specification V0.5

3. Research Findings

1. Blockchain can help in various operational aspects of an RTA, particularly in the banking, financial and non-financial transactions.
2. Using the various architectures and the five-layered system proposed by the Enterprise Ethereum Alliance (EEA) can help in building scalable, inter-operable and pluggable systems.

3. Blockchain can accommodate direct on-chain and hash-based off-chain storage of documents which are important in non-financial transaction processing activity.

4. Resource intensive computational operations can be moved off-chain and so, high industry-grade performance throughput can be obtained.

4. Scope for further development

1. This paper touches only the mutual fund aspects of an RTA and clearly this can be extended to other investment instruments (such as units issuing insurance policies, PMS, venture capital and hedge fund activities etc.).

2. The exchange routed MFs (including ETFs) require a separate through discussion as it includes several other entities that are not of the core traditional MF route.

3. Some RTAs and AMCs provide auxiliary services such as capital gains statement etc. which can also be considered for implementation over blockchain.

5. Conclusions

RTAs provide record keeping and investor support services and are vital financial intermediaries in the mf industry. Various features of the Blockchain technology, particularly the Enterprise Ethereum, can help RTAs to migrate to the Blockchain so that process optimization and cost cutting can be done. The services of EEA (such as the architectural framework and the layered approach) can help in bringing about standards in the blockchain industry and help derive long term benefits because of easy inter-operability with multiple vendors. Moving transactions and storage off-chain can reduce some burden and improve throughput of the blockchain network as well as in optimal system resource utilization.

6. References

1. Vijaya Kittu, M., & Aruna, P. (2018). Status Check on Blockchain Implementations in India. International Conference on Technological Innovations in Management Ecosystem. doi:10.2139/ssrn.3265654
2. Vijaya Kittu, M., & Prasada Rao, S.S. (2018). Blockchain Technology for the Mutual Fund Industry. National Seminar on Paradigm Shifts in Commerce and Management, 12-17. doi:10.2139/ssrn.3276492
3. AMFI. (Feb 2019). *Industry Trends*. AMFI
4. Investment Company Institute. (2019). *Trends in Mutual Fund Investing - February 2019*. ICI Global.
5. Sriram, V. (2016). *CAMS: An Eye for Detail*. Retrieved from http://www.acsysindia.com/links/CAMS_An_Eye_for_Detail.pdf
6. NISM. (2017). *Workbook for NISM Series IIB: Registrars to an issue and Share Transfer Agents (Mutual Funds) Certification Examination*. National Institute of Securities Markets.
7. Enterprise Ethereum Alliance. (15 October 2018). *Off-Chain Trusted Compute Specification V 0.5*.
8. Robson, C. (2002). *Real World Research* (2nd ed.). Oxford: Blackwell.

AN ANALYSIS FOR PREVENTION OF FAKE NEWS USING BLOCKCHAIN TECHNOLOGY

^{#1}Dr. R.KAMALAKAR, *I/C, Head, Dept of Computer Science, Satavahana
University, Karimnagar.*

^{#2}KISHOR KUMAR GAJULA, *Ph.D Scholar, Dept of CSE, Shri JJT University,
Rajasthan.*

Abstract: Nowadays, fake news is hovering far and wide vaingloriously, despite making fear its objective gatherings. Dramatist revealing is an exceptional instance of "Fake News" that fuses intentionally contaminated or fabricated information which is circulated in a couple of media, for instance, in TVs, traditional papers and online by means of electronic systems administration media locales. The web based life stages join Facebook, Twitter, Instagram, Reddit and various destinations. Moreover, individuals who make fake news are conventionally masters who ability to control and get the thought of people towards their articles or chronicles they convey. Nevertheless, not all news is fake by means of electronic systems administration media locales. The evident system is spreading a wide scope of fake information all through different electronic life stages since they are before long the most dynamic correspondence organizes that exist in this age. Conversely with standard media, for instance, magazines, papers, TVs, and radios; web based life stages have accomplished the top spot similar to the sign of fake news. Also, the inspiration driving why fake news is made should be conceivable as a sort of activism, confronting or supporting political issues, for adjustment, or despite for fame and reputation. In this paper we propose the examination of foreseeing fake news using square chain development using the timestamps and land regions and marks.

Keywords: Blockchain, Fake news, big data, Ethereum, hash functions, digital signatures.

I.INTRODUCTION

Will fake news cause fiendishness to individuals all in all? The fitting reaction is most likely yes if you are one of the concentrated on observers, who examine on some phony articles by means of electronic systems administration media. Some no matter how you look at it counterfeit news may cause revolts, trailed by making hurt the all inclusive community, especially if it has been associated with political issues. The most notable, are disrupting impacts among ideological gatherings or even among democrats and republicans. Protestation and crowds may occur, which would then have the option to provoke a couple of veritable and dangerous conditions. In the earlier months, during the choice days, the articulation "Fake news" has ended up being notable and is seen as a hazard to the council. On November 27th, President Trump proclaims the "Fake News Trophy," that was dispersed in the New York Post article. The President responded to the post through Twitter, his most utilized online life webpage. He communicated, "We should have a test as for which of the Networks, notwithstanding CNN and barring Fox, is the most exploitative, savage or possibly bent in its

political incorporation of your most adored President(me)" according to maker Fredericks. This declaration was posted with thoroughgoing joke by the president to CNN as a method for retaliation towards what they posted on President Trump during the choice. It was the moment that "Fake News" transformed into the element in numerous news frameworks' articles, especially through online systems administration media destinations.

Fake News

Fake News is, simply, envisioned information. Incredibly, it is routinely difficult to spot envisioned from certified. For instance, in a progressing audit, when the UK's Channel 4 News demonstrated three certified and three fake stories to 1,684 adults, only 4% of the respondents had the alternative to separate all of the records successfully, and about half thought that at any rate one of the fakes was authentic

II. LITEARTURE REVIEW

During a progressing TED talk, Yuval Noah Harari expressed: "I think fake news has been with us a long time; essentially consider the Bible!". In actuality, the most dependable instance of declaration is seen as the Behistun Inscription, made around 515 BC, which is an etching in three unmistakable cuniform languages on a slope at Mount Behistun in Kermanshah Province, Western Iran. It nuances the rising to the situation of power of the Persian Empire of Darius I and his achievement in stifling various uprisings. In any case, Pope Gregory XV was the first to use the term 'exposure', when in 1622, he molded the 'Congregatio de Propaganda Fide', or "social event for spreading the certainty." The word itself begins from the Latin word 'propagare', which means multiplication. In this way, attention is fathomed to mean the multiplication of a conviction framework. A dynamically current instance of deliberate attention, yet still one-hundred years old, was delineated by Dr David Clarke in a progressing piece for the BBC18.

Dr Clarke tells how, in 1917, the British Government, in an in the long run powerful undertaking to bring China onto the Allied side in The Great War, made a repulsive tale about the German military, whom they (untrustworthily) accused for removing glycerine from human bodies. Plainly, Conservative MP John Charteris, Head of Intelligence at the period of the story's production, transposed captions from a photograph that showed a train of dead steeds that ought to have been rendered onto another exhibiting a train taking dead warriors for internment. Grievously, the story was later used by the Nazi Party as affirmation of British lies during the Great War, and it may have provoked inquiries concerning reports on Nazi shock during the Second World War; as Dr Clarke comments: "lies have results".

The Nazi Party, understanding the noteworthiness of war proclamation, molded the Reich Ministry of Public Enlightenment and Propaganda. The Ministry's head, Joseph Goebbels, used his control of the press to help brace Nazi conviction framework through fake news: "If you tell a comparative lie enough events, people will confide in it; and the more prominent the deception, the better"¹⁹. Much like Nazi Germany, Stalinist Russia, attempting to induce its kinfolk that the Soviet Union savored the experience of significantly higher desires for ordinary solaces than those in the Capitalist West, used deliberate attention extensively²⁰. During the lead-up to the Second World War, the Soviet media smothered strange end through the limitation of boisterous voices. Paper highlights took a standard structure: "all experts respected the methodology (of the Russian

Government) with satisfaction." They reiterated the message routinely, offering certainty to Goebbels' mantra that if you lie as often as possible enough, people will confide in it. Soviet deliberate attention continued after the war also, with books overwhelmingly blue-penciled and papers multiplying ideolised reality. Television and radio gave that reality a degree of show. In the meantime, cinematography took a triumphalist tone, depicting happy lives and the fulfillment of the 'Soviet dream'.

The Russian State has not been the sole purveyors of fake news in the front line world. In 1928, Cornell Graduate Edward Bernays circulated a book called Propaganda, which has advanced toward ending up, essentially, a manual of mass manipulation²⁴. The book opens with the going with entry: "The insightful and smart control of the dealt with affinities and evaluations of the larger part is a noteworthy segment in a notoriety based society. The people who control this disguised arrangement of society include a subtle government which is the certified choice force of our country." really, before the First World War, the term intentional exposure was not used antagonistically, anyway the open began to question the term once they comprehended how much the Anglo-American political mechanical assembly had sent declaration attempting to criticize "The Hun". Its use by the Nazi Party in the Second World War²⁵, and later by Communist Russia, appears to have fixed the term's predetermination; by and by proclamation has incredibly negative ramifications. In any case, that does not suggest that its usage in the West has diminished. Following the war, U.S. President Truman incited NSC, a procedure to contain the Soviet state using wide-going undercover assignments, including deliberate exposure. During the 1960s and 1970s, the media organizations of Western nations were instrumental in progressing neo-dominion (the demonstration of applying effect or order over less made countries by using trade courses of action, financial or cash related strategies) and weakening undertakings at selfdetermination by third world countries²⁵. There are late occasions of Western declaration also; in 2005, the U.S. Government endeavored to impact general supposition with respect to the upsides of the Iraq War by consuming US\$300 million on a movement to multiply 'positive news'.

III. BLOCKCHAIN MARKET

To by a wide edge most of individuals if all else fails, blockchain is related with cryptographic sorts of cash, for instance, Bitcoin, Ether, or diverse other virtual "coins." The clarification for the strong association between cutting edge money and blockchain is that PC based financial principles are the distinguish the arrangement guideline of blockchain is still most regularly found today. It's what made blockchains certainly found regardless.

Unequivocally when Bitcoin progressed blockchain in 2009, after the arrangement of a white paper made by the pseudonymous Satoshi Nakamoto,⁸ computer specialists and cryptographers had only thought concerning a great deal of blockchain's fundamental improvement, generally in educational circles.⁹ For instance, with an outrageous objective to make instruments to fight email spam and securely send partitions on the web, PC analysts had explored cryptography and secure transmission shows up. The usage of cryptography to pass on open and private keys, or the specific figurings that make long numerical codes (hashes), moved

during this period of research. In blockchain propels today, these for all intents and purposes indistinguishable structure measures have now gotten a certainly wide field of utilization.

Nakamoto's flimsy paper on the proposition for Bitcoin can be seen as the repurposing of a wide degree of plans and developments streamed at the fortunate time. The likelihood of a sidekick reviewed and deductively strong piece system hit a nerve for a couple, obviously in light of the general outcome from the Great Recession. As certified cash related foundations fizzled and caused a general bank run, the silver shot of a secured progress managing most likely a segment of the world's budgetary issues sounded curious and inviting.

Blockchain was first uncommon as a bank-less, pitiful, and secure way to deal with oversee pay for things, making a main impetus for electronic money related structures by record trades as secure and lasting. Inside fundamental estimation of modernized sorts of cash is the trust of its customers that it will be seen as a sort of part—basically like by uprightness of national money related models far and wide—and that, by ethicalness of blockchain's mechanical properties, it can't be faked or hacked.

Much all the additionally starting late, notwithstanding, various endeavors have begun inspecting various roads concerning blockchain's structure rule for applications past mechanized money. The more sweeping media industry, which joins revealing, progressed driving, swarm examination, music, et al., is just one of many.

The purpose of blockchain

Generally speaking, blockchains are meant to do three things:

- **store** (small amounts of) data (in containers called “blocks”);
- **keep track** of all modifications made to the data (by threading them into a “chain” that cannot be altered); and
- **secure** the data and its many edited versions in a way that multiple users can agree on, including how data is stored, protected, and remains unchanged. This is where cryptography, proof-of-work, and community consensus come in.

In most blockchain models, a fourth section is in like manner included: the helper to look into the arrangement of people who store, screen, and secure data. Catalysts motivate excavators to make and support deters in the blockchain, or knock accomplices to police and control a blockchain. (More on excavators and accomplices later.)

The most unmistakable terms used by blockchain explanations are decentralized, immutable, clear, dispersed (record), and trustless frameworks. Most of these portrayals suggest a comparable basic idea: let there be a record of stuff (data) that people can agree on and not change later on, and secure that stuff with the objective that all its history is consistently undeniable to everyone. In the domain of blockchain, trust and memory are manufactured square by square.

The hinders that contain data have a uniform structure across over for all intents and purposes all blockchain applications.

The block

Each block consists of a:

- **version number** (to mark the position it occupies in the chain)
- **header hash** (a number code that links it to the previous block's output hash)
- **timestamp** (the time when the block was created)
- **Merkle root** (the block's content encoded into a hash, which is a number code, as mentioned above)
- **nonce** (a random number used to randomize and create the output hash of the block, which will then link it to the next block and thus lock the block into its place within the chain)
- **output hash** (the version number, the header hash, the time stamp, the Merkle root, and the nonce ALL encoded into yet another single code number).

There are also some other general properties of Blocks. Blocks are commonly a couple of megabytes in size since they just contain arrangement of numbers (hash yields). In numerous blockchains, each square is accessible to anyone with a PC and a web affiliation. Entire blockchains can be downloaded to a PC (e.g., the entire history of Bitcoin trades is by and by accomplishing 200 gigabytes, notwithstanding the way that it used to be only two or three gigabytes just a few years earlier).

Blocks are clear, yet garbage to the human eye with their long lines of numbers and letters. Whenever opened, all hashes (the gigantic, by and large 32-to 64-digit numbers and letters) can be seen. The reason for blockchain is that paying little mind to whether somebody needs to change something in a square by revamping a couple of characters in the substance of any square, the entire chain following that square will be balanced. This is basically a domino sway, in light of the way that the going with Blocks are encoded reliant on the data found in the past Blocks.

Since the system embraced chain of Blocks is continually secured on various PCs in the framework, it might be restored adequately. The past the square is in a chain, the more inconvenient it is to change since a consistently expanding number of Blocks after it would ought to be adjusted. Eventually, changing a lone square only one position removed from the "freshest" square is incomprehensible without being recognized by a huge amount of centers in the framework.

IV. HASHING TECHNIQUES

Hashing is the PC's approach to manage get information. There are different ways information that is comprehensible for people can be changed into a PC huge code, at any rate the structure standard is clear: input information is facilitated with a numerical code, similar to an individual name is united with a phone number. The way where a PC relates the two is through hashing, so it acknowledges this information pair and can recover it effectively if a client needs to get to it.

In blockchains, the information contained in each square is hashed, which recommends that sentences, words, dates, or numbers that look great to people are encountered the hashing estimation. Hashing estimations complete a development of ordained coherent checks to make a code for any sort of information. Undoubtedly, hashing still looks similarly as numbers or messages were mixed to pass on an unlimited code.

Notwithstanding, PCs can rapidly switch their figurings and produce the information, which people can examine or see once more, if the PC is settled what the data information was. A little while later, these estimations perceive any sort of data, and produce diverse fixed length, for example, 32 or 64 digits. Hash yields as frequently as conceivable resemble a long queue of the two numbers and letters since most standard hashing calculations encode hexadecimally, deriving that every "digit" can contain 16 characters: the numbers 0 to 9 and the letters A to F—10 numbers and six letters outright.

For cryptographic purposes, express hashing estimations are helpful in light of the way that the way wherein they scramble any sort of duty to a limited strategy of numbers is sporadic. So flighty that if the information is dim, somebody would need to try all conceivable hash mixes to consider the information string. Undoubtedly, even on the snappiest PCs, testing all mixes would require a gigantic number of years.

The yield (or, in the language of cryptography, the development or the hash) is free of the information's length. A solitary holler etching will pass on a hash of a similar length as the hash produced using a whole novel. Significantly more curiously, if a solitary complaint etching is erased from the substance of said novel, the hashing estimation will make a very amazing number for its review.

Be that as it may, randomization just works one way: while the review that is being passed on is flighty, a near information will dependably make an equivalent layout. PC experts call this single bearing nature of hashing deterministic. There is no certified technique to manage hash yields, for instance, finding what the fundamental hash yield for the number 9 is and trying to find it in the long numerical hash that is made for the number 19: the two yields will be thoroughly alarming, and the yield code of 19 may genuinely take after the hash yield of our novel in the past model.

In blockchains, Blocks don't for the most part contain humanly rational numbers or messages, at any rate rather pass on sets of hash yields. Everything in a rundown, each exchange, or each lump of information in the square is made into a hash, and the subsequent hashes are facilitated thoughtlessly to change into the information for a solitary new hash. Exactly when each bit of information in the square is hashed, and the majority of the hashes are hashed with one another over and over, a particular last hash is made. This last hash is known as the Merkle root, named after the mathematician Ralph Merkle who approved along these lines of multi-layered and secure strategy for verifying information.

The "root" starts from the tree-structure of the manner by which hashes are related with one another: hashes look like leaves that are hardened two by two on a branch (which is the going with hash), and branches are participated in the base of the tree (the last hash). Verifying information in such an alternate leveled course inside each square is valuable in light of the

manner in which that all of information ends up deterministic. Each square in the one-square per-level Jenga is itself contained a progressively small Blocks supporting one another. When one square is cleared or changed, everything above it—inside the square and following the square—breakdown.

Adding another square to the chain

So far we have talked about what is verified in a solitary square, and how things are verified in it. The last computational part of the blockchain is the affiliation that strings avoids into a chain. As referenced starting at now, the basic bit of that string is the square hash, which is the hash yield of the alteration number, the Merkle root, the timestamp, the header hash (the past square's yield hash), and the nonce. These numbers are consolidated by the hashing calculation to make the square hash, which the going with square will contain as its header hash.

Notwithstanding, which square hash will be respected? Focuses in the structure are up to this point doing combating to have their new store of information (the square) added to the chain. So who picks which square gets included? This is the recognize the confusing number, the "nonce," transforms into an irreplaceable factor. The "n(umber just used)once" is an abstract number, and is routinely shorter than the hash number. Since it is blended in with the Merkle root, the modification number, and the various numbers contained in the square, it picks the last square hash.

On the off chance that the nonce changes, so does the square hash. Right when diggers in the blockchain system attempt to locate the last hash for a square and get it grasped by their colleagues in the structure, they try a huge number of groupings of the nonce to locate that last hash. Since the nonce is optional, it can't be gotten from some other information encoded in the square, and that is the point. Blending this erratic fragment into Blocks requires the battling PCs to dependably recalculate the yield hash when making another square.

There is a last curve to making another square, in any case. So as to abstain from figuring trillions of kinds of the yield hash (which would take a gigantic number of years) to locate a particular one, the blockchain show is proposed to make diggers locate any sporadic hash that is more little than an objective number. This clearly basic specific course of action was Bitcoin's innovative thought, in order to make square support less tedious, progressively strong, and monetarily fulfilling.

Note, regardless, that by chopping down the estimation of the objective number, the blockchain system doesn't diminish the measure of potential blueprints a PC must undertaking. It just diminishes the measure of palatable game-plans from trillions and trillions (this is the reason the burden ascends by chopping down the objective number). Since focuses are not required to locate a particular hash (they don't need to make an understanding of a hash to discover the information string), they simply need to discover a hash that is inside the cutoff reasons for the objective number by embeddings another nonce each and every time they attempt to locate the new square hash.

This is still time and centrality debilitating, at any rate reasonable for profound PCs that can try different groupings extremely quick. Blockchain demonstrates by and large change the objective level subject to the degree of conflict among focus focuses to decently pace out hash-

approaches in the structure. In the Bitcoin arrange today, another square is made typically all around.

V. PREVENTION METHODS

Blocks require a great deal of computational starters and messes up to accomplish a fittingly stunning square hash. The time and criticalness use of PCs to mine each new square is the evidence that people in the structure spot work into keeping up the framework. The first Bitcoin declaration by Satoshi Nakamoto considered this the "one-CPU-one-vote" model. In cryptographic money applications, made by excavators, which is completely settled on PCs' CPU yield, is repaid by getting cash as robotized tokens, for example, Bitcoin or Ether. Despite whether two focus focuses locate a reasonable nonce for the square hash at about an equivalent time, the triumphant square will be the one that required even somewhat more CPU output.² sorts of conditions for evading of phony news methodology.

i). Proof-of-Work model ii). Proof-of-Stake model

i). Proof-of-Work Model,

This model is the **Proof-of-Work** scenario. The security and decency of the blockchain is undergirded by a tremendous equipped power of PCs that encode gigantic proportions of data, and screen each new piece of data added to the chain. Steady and customized perception of the framework is required since software engineers may need to change bits of the blockchain in order to record fake trades that would add mechanized money to their propelled wallets. As we have seen, regardless, despite altering a lone character in a lone square's single hash will trigger the Merkle root to change, and nearby it the entire square's yield hash, and along these lines coming up next square's header hash, similarly as coming up next one's, and so forth.

To revalidate each following square in order to cover someone's modifying would require the recalculation of each square's nonce to find new yield hashes to relink them to coming about squares. This would take significant lots of work during which by far most of the framework would recognize what's happening, and brief them to intervene. These center points have the privilege and avowed blockchain locally on their PCs, and can restore it and twofold check it with various center points.

The estimation of the tokens allowed to excavators is, clearly, dependant on the swapping scale among tokens and "authentic" fiscal models. The economy of blockchain-set up together tokens depends regarding a variety of human factors, for instance, trust in the framework, media thought, the risk of rule, publicize alert, and so on. Like stocks and money related norms, the monetary estimation of tokens waver reliant on token holders' trust.

Additionally, the infrastructural state of mining new tokens have changed essentially. Today diggers much of the time blend into pools to perform computational assignments together and a short time later offer token prizes among themselves. PC getting ready units unequivocally proposed to beneficially run hashing counts are in like manner open accessible. These are called ASICs, or application-unequivocal joined circuits. This is to express that while mining computerized types of cash may have been practical as a diversion 10 years back, with less PCs in

the framework and a lower inconvenience rate, today nature of mining pools infers that single a lot of pool managers control all miners.

These chiefs oftentimes go into phony concurrences with each other to avoid pool centralization where one pool stores in excess of 50 percent of the CPU power of the entire framework, thusly haggling the genuineness of the blockchain. In case most of the framework center points are utilized to support a comparative square, they become fit for what is known as the "51-percent attack," when there are more PCs to retroactively adjust the blockchain and favor the remedy than there are companions to transform those modifies. Essentially, while centralization is really what blockchain frameworks expected to keep up a vital good ways from, the current physical structure of pervasive blockchains, as Bitcoin, is consolidating at an exasperating rate.

ii). Proof-of-Stake model:

Peercoin was the essential standard computerized money to use PS as its key controlling rule, and Ethereum is going with a similar example. As opposed to anticipating that PCs should race against each other and induce the nonce to make a square hash, a singular PC inside the framework is picked to endorse another square. The center point is picked reliant on its stake in the framework, which can be the proportion of cryptographic cash it holds. It is normal that in case someone holds more money in a given advanced cash, the more stake they will have—and wager—to carefully endorse another square. The validator fundamentally bets their own money to endorse another square, and gets all the trade charges from the square they had the alternative to favor.

Various center points in the framework trust the validator center subject to how much stake this center has in contributed advanced cash. In PS, no new coins are mined, and rather validators get cash related reward as trade charges (this is the inspiration to transform into a validator). Validators are looked over a pool of competitors after they are checked on and enrolled into the applicant pool.

Instead of the PS structure, the benefits of PW is the framework effect of shared trustlessness, which keeps everyone on their toes to screen in case someone needs to hack the blockchain. Its disservices are: wasteful imperativeness use and the mystery of individuals. Uncertainty is imparted in the structure standard of blockchain with the objective that part center points remain caution and hold each other inside appropriate cutoff points.

Then again, the benefit of PS is the trust and straightforwardness of the checked system of friends, yet the drawback is that carefulness is brought into the structure. Someone needs to make a show by which validator competitors are picked, and dynamically human organization is required to deal with false on-screen characters. To get stake in the framework, a potential new performer must contribute capital through the purchase of coins, which in this way may make irregularity in the structure straightforwardly from the most punctual beginning stage.

Notwithstanding whether the proportion of coins that can be picked up is topped, and paying little mind to whether token trades are vivaciously managed to keep up a key good ways from

liquidity, settling inquiries by demonstrating the proportion of arranging concessions a center point holds may continue conflicting force components. Associations and overwhelming parts may be encircled, and dealing with staggering on-screen characters in the system who may end up being successfully baffled can put weight on the framework.

VI.CONCLUSION

As a conclusion, this examination has given starter bits of data in the party of blockchain and online life to make trust and predict counterfeit news. There are different difficulties in checking news substance and equalization counterfeit news by methods for electronic frameworks organization media. Broad research is required around there to get the potential prizes. This examination locale has goliath certifiable political, economy and social effects. With blockchain improvement that are advancing and redesigning determinedly, there are different new open gateways for awkward progress in present and future natural course of action of electronic life so as to fix up a persistently trustable, solid, direct, and secure central surfaces for our bleeding edge society.

REFERENCES:

- [1] S. Haber and W. S. Stornetta, "How to time-stamp a digital document," *Journal of Cryptology*, vol. 3, no. 2, pp. 99-111, 1991.
- [2] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," 2008 [Online]. Available: <https://bitcoin.org/bitcoin.pdf>.
- [3] The Economist, "The great chain of being sure about things," 2015 [Online]. Available: <https://www.economist.com/news/briefing/21677228-technology-behind-bitcoin-lets-people-who-do-not-know-or-trusteach-other-build-dependable>.
- [4] Bitcoinwiki, "Genesis block," 2017 [Online]. Available: https://en.bitcoin.it/wiki/Genesis_block.
- [5] E. Robert, "Digital signatures," 2017 [Online]. Available: http://cs.stanford.edu/people/eroberts/courses/soco/projects/public-key-cryptography/dig_sig.html.
- [6] Ethereum [Online]. Available: <https://www.ethereum.org>.
- [7] S. Popov, "A probabilistic analysis of the Nxt forging algorithm," *Ledger*, vol. 1, pp. 69-83, 2016.
- [8] Nxt wiki, "Whitepaper: Nxt," 2016 [Online]. Available: <https://nxtwiki.org/wiki/Whitepaper:Nxt>.
- [9] The Public Disputes Program, "A short guide to consensus building," [Online]. Available: http://web.mit.edu/publicdisputes/practice/cbh_ch1.html.
- [10] A. Back, "Hashcash: a denial of service counter-measure," 2002 [Online]. Available: <ftp://sunsite.icm.edu.pl/site/replay.old/programs/hashcash/hashcash.pdf>.
- [11] Bitcoinwiki, "Proof of Stake," 2014 [Online]. Available: https://en.bitcoin.it/wiki/Proof_of_Stake.

- [12] Intel, "Sawtooth v1.0.1," 2017 [Online]. Available: <https://sawtooth.hyperledger.org/docs/core/releases/latest/introduction.html>.
- [13] M. Milutinovic, W. He, H. Wu, and M. Kanwa, "Proof of luck: an efficient Blockchain consensus protocol," in *Proceedings of the 1st Workshop on System Software for Trusted Execution*, New York, NY, 2016.
- [14] S. Park, K. Pietrzak, A. Kwon, J. Alwen, G. Fuchsbauer, and P. Gazi, "Spacecoin: a cryptocurrency based on proofs of space," 2017 [Online]. Available: <https://eprint.iacr.org/2015/528>
- [15] C. Cachin, "Architecture of the hyperledger blockchain fabric," in *Proceedings of ACM Workshop on Distributed Cryptocurrencies and Consensus Ledgers*, Chicago, IL, 2016.
- [16] QuorumChain Consensus [Online]. Available: <https://github.com/jpmorganchase/quorum/wiki/Quorum-Chain-Consensus>.
- [17] L. Lamport, "Paxos made simple," *ACM Sigact News*, vol. 32, no. 4, pp. 51-58, 2001.
- [18] M. Castro and B. Liskov, "Practical Byzantine fault tolerance," in *Proceedings of the Third Symposium on Operating Systems Design and Implementation*, New Orleans, LA, 1999, pp. 173-186.
- [19] M. Castro and B. Liskov, "Byzantine fault tolerance," U.S. Patent 6671821 B1, Dec 30, 2003.
- [20] C. Pommier, "How the private and public key pair works," 2017 [Online]. Available: <https://www.symantec.com/connect/blogs/how-private-and-public-key-pair-works>.
- [21] Elliptic curve digital signature algorithm, 2017 [Online]. Available: https://en.bitcoin.it/wiki/Elliptic_Curve_Digital_Signature_Algorithm.
- [22] Bitcoin Stack Exchange, "Can someone explain how the Bitcoin Blockchain works?," 2017 [Online]. Available: <https://bitcoin.stackexchange.com/questions/12427/can-someone-explain-how-the-bitcoinblockchain-works>.
- [23] Bitcoinwiki, "Block hashing algorithm," 2015 [Online]. Available: https://en.bitcoin.it/wiki/Block_hashing_algorithm.
- [24] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Advances in Cryptology– CRYPTO'87*. Heidelberg: Springer, 1987, pp. 369-378.
- [25] Bitcoinwiki, "SHA-256," 2016 [Online]. Available: <https://en.bitcoin.it/wiki/SHA-256>.
- [26] J. Tromp, "Cuckoo cycle: a memory-hard proof-of-work system," 2015 [Online]. Available: <https://eprint.iacr.org/2014/059.pdf>.
- [27] K. Schwarz, "Cuckoo Hashing," [Online]. Available: <http://web.stanford.edu/class/cs166/lectures/13/Small13.pdf>.
- [28] S. King, "Primecoin: cryptocurrency with prime number proof-of-work," 2013 [Online]. Available: <http://primecoin.io/bin/primecoin-paper.pdf>.

[29] C. K. Caldwell, "Cunningham chain," 2017 [Online]. Available: <http://primes.utm.edu/glossary/xpage/CunninghamChain.html>.

[30] M. Liskov, "Fermat primality test," in *Encyclopedia of Cryptography and Security*. Boston, MA: Springer, 2005, pp. 221-221.

Blockchain Based Transaction Systems”

Annapareddy V N Reddy¹, Chitta Venkata Phani Krishna²

¹Assistant Professor, Vikas College of Engineering and Technology,
Vijayawada, nrnagarjunareddy@gmail.com. ²Professor, Teegala Krishna Reddy Engineering College,
Hyderabad, phanik16@gmail.com.

Abstract:

A blockchain is basically distributed information of records or public ledger of all transactions or digital events that are dead and shared among taking part parties. Every dealing in the public ledger is verified by agreement of a majority of the participants within the system. And, once entered, info will ne'er be erased. The blockchain contains an exact and verifiable record of every single dealing ever created. Bit coin, the suburbanized peer-to-peer digital currency, is the most popular example that uses blockchain technology. The digital currency bit coin itself is extremely controversial however the underlying blockchain technology has worked cleanly and located wide selection of applications in each monetary and non-financial world. The main hypothesis is that the blockchain establishes a system of making a distributed consensus within the digital on-line world. This permits taking part entities to understand for sure that a digital event happened by making A positive record in an exceedingly public ledger. It opens the door for developing a democratic open and ascendible digital economy from a centralized one. There are tremendous opportunities during this unquiet technology and revolution during this area has simply begun. This written report describes blockchain technology and a few compelling specific applications in each financial and non-financial sector. We have a tendency to then inspect the challenges ahead and business opportunities in this elementary technology that's prepared to revolutionize our digital world.

1. Introduction:

Blockchains square measure tamper evident and tamper resistant digital ledgers enforced in an exceedingly distributed fashion (i.e., while not a central repository) and frequently while not a central authority (i.e., a bank, company or government). At their basic level, they allow a community of users to record transactions in an exceedingly shared ledger inside that community, such below traditional operation of the blockchain network no dealings may be modified once printed. In 2008, the blockchain plan was combined with many different technologies and computing ideas to form trendy crypto currencies: electronic money protected through science mechanisms rather than a central repository or authority. This technology became wide better-known in 2009 with the launch of the Bit coin network, the first of many trendy crypto currencies. In Bit coin, and similar systems, the transfer of digital information that represents electronic money takes place in an exceedingly distributed system. Bit coin users will digitally sign and transfer their rights thereto data to a different user and also the Bit coin blockchain records this transfer publically, permitting all participants of the network to independently verify the validity of the transactions. The Bit coin blockchain is severally maintained and managed by a distributed cluster of participants. This, together with science

mechanisms, makes the blockchain resilient to makes an attempt to change the ledger later (modifying blocks or formation transactions). Blockchain technology has enabled the event of the many crypto currency systems like Bit coin and Ethereum1. As a result of this, blockchain technology is often viewed as guaranteed to Bit coin or presumably crypto currency solutions normally. However, the technology is on the market for a broader style of applications and is being investigated for a variety of sectors. The numerous parts of blockchain technology together with its reliance on science primitives and distributed systems will create it difficult to know. However, each component may be represented merely and used as a building block to know the larger complex system. Blockchains may be informally outlined as: Blockchains are distributed digital ledgers of cryptographically signed transactions that are grouped into blocks. Every block is cryptographically connected to the previous one (making it tamper evident) when validation and undergoing a accord call. As new blocks are added, older blocks become harder to switch (creating tamper resistance). New blocks are replicated across copies of the ledger at intervals the network, and any conflicts are resolved mechanically victimization established rules.

1.1 Background and History:

The core ideas behind blockchain technology emerged within the late Nineteen Eighties and early Nineties. In 1989, Leslie Lamport developed the Paxos protocol, and in 1990 submitted the paper The Part Time Parliament [2] to ACM Transactions on laptop Systems; the paper was finally published in an exceedingly 1998 issue. The paper describes a accord model for reaching agreement on a result in a network of computers wherever the computers or network itself could also be unreliable. In 1991, a signed chain of data was used as associate degree electronic ledger for digitally linguistic communication documents in an exceedingly means that would simply show none of the signed documents within the assortment had been modified. These ideas were combined and applied to electronic benefit 2008 and described within the paper, Bit coin: A Peer to see Electronic money System that was revealed pseudonymously by Satoshi Nakamoto, and so later in 2009 with the institution of the Bitcoin crypto currency blockchain network. Nakamoto's paper contained the blueprint that the majority modern crypto currency schemes follow (although with variations and modifications). Bitcoin was simply the primary of the many blockchain applications. Many electronic money schemes existed before Bitcoin (e.g., ecash and NetCash), however none of them achieved widespread use. The utilization of a blockchain enabled Bitcoin to be enforced in an exceedingly distributed fashion specified no single user controlled the electronic money and no single purpose of failure existed; this promoted its use. Its primary profit was to alter direct transactions between users while not the necessity for a trusty third party. It additionally enabled the provision of recent crypto currency in an exceedingly outlined manner to those users World Health Organization manage to publish new blocks and maintain copies of the ledger; such users area unit referred to as miners in Bitcoin. The machine-controlled payment of the miners enabled distributed administration of the system while not the necessity to arrange. By using a blockchain and consensus-based maintenance, a self-policing mechanism was created that ensured that solely valid transactions and blocks were else to the blockchain. In Bitcoin, the blockchain enabled users to be ominous. This suggests that users area unit anonymous, however their account identifiers area unit not; to boot, all transactions area unit in public visible. This has effectively enabled Bitcoin to supply pseudo-anonymity as a result of

accounts is often created with none identification or authorization method (such processes area unit generally needed by Know-Your-Customer (KYC) laws). Since Bitcoin was onymous, it had been essential to possess mechanisms to make trust in associate degree environment wherever users couldn't be simply known. Before the utilization of blockchain technology, this trust was generally delivered through intermediaries trusty by each parties. Without trusty intermediaries, the required trust among a blockchain network is enabled by four key characteristics of blockchain technology, represented below:

- **Ledger** – the technology uses associate degree append solely ledger to produce full transactional history. Unlike ancient databases, transactions and values in an exceedingly blockchain don't seem to be overridden.
- **Secure** – blockchains area unit cryptographically secure, guaranteeing that the information contained within the ledger has not been tampered with, which the information among the ledger is attestable.
- **Shared** – the ledger is shared amongst multiple participants. This provides transparency across the node participants within the blockchain network.
- **Distributed** – the blockchain may be distributed. This enables for scaling the amount of nodes of a blockchain network to form it a lot of resilient to attacks by unhealthy actors. By increasing the amount of nodes, the flexibility for a foul actor to impact the accord protocol utilized by the blockchain is reduced.

2. Blockchain Categorization:

Blockchain networks are often classified supported their permission model, that determines United Nations agency can maintain them (e.g., publish blocks). If anyone will publish a replacement block, it's permission less. If solely explicit users will publish blocks, it's permission. In easy terms, a permission blockchain network is sort of a company computer network that's controlled, whereas permission less blockchain network is just like the public web, wherever anyone will participate. Permission blockchain networks area unit typically deployed for a bunch of organizations and people, typically referred to as an association. This distinction is important to know because it impacts a number of the blockchain elements mentioned later during this document.

2.1Permissionless:

Permission less blockchain networks area unit localised ledger platforms hospitable anyone publishing blocks, while not having permission from any authority. Permission less blockchain platforms area unit usually open supply software package, freely out there to anyone UN agency needs to transfer them. Since anyone has the correct to publish blocks, this leads to the property that anyone will read the blockchain likewise as issue transactions on the blockchain (through together with those transactions among printed blocks). Any blockchain network user among permission less blockchain network will browse and write to the ledger. Since permission less blockchain networks

are hospitable all to participate, malicious users might decide to publish blocks in a very method that subverts the system discussed intimately to stop this, permission less blockchain networks usually utilize a multiparty agreement or 'consensus' system users to expend or maintain resources once trying to publish blocks. This prevents malicious users from simply subverting the system. Samples of such agreement models embrace proof of labor (see Section four.1) and proof of stake (see Section four.2) methods. The agreement systems in permission less blockchain networks sometimes promote non-malicious behavior through rewarding the publishers of protocol-conforming blocks with a native crypto currency.

2.2 Permissioned:

Permissioned blockchain networks area unit ones wherever users publication blocks should be approved by some authority (be it centralized or decentralized). Since solely approved user's area unit maintaining the blockchain, it's doable to limit browse access and to limit United Nations agency will issue transactions. Permissioned blockchain networks could therefore permit anyone to browse the blockchain or they will restrict browse access to approved people. They conjointly could permit anyone to submit transactions to be enclosed within the blockchain or, again, they will limit this access solely to approved individuals. Permissioned blockchain networks are also instantiated and maintained exploitation open source or closed supply package. Permissioned blockchain networks will have a similar traceability of digital assets as they pass through the blockchain, yet because the same distributed, resilient, and redundant knowledge storage system as permission less blockchain networks. They conjointly use agreement models for publication blocks; however these strategies usually don't need the expense or maintenance of resources (as is that the case with current permission less blockchain networks). This is often as a result of the institution of one's identity is needed to participate as a member of the permission blockchain network; those maintaining the blockchain have grade of trust with one another, since they were all authorized to publish blocks and since their authorization is revoked if they move. Consensus models in permissioned blockchain networks are then typically quicker and fewer computationally big-ticket. Permissioned blockchain networks may additionally be utilized by organizations that require to a lot of tightly control and shield their blockchain. However, if one entity controls WHO will publish blocks, the users of the blockchain can get to have trust in this entity. Permissioned blockchain networks may additionally be utilized by organizations that want to figure along however might not totally trust one another. They will establish a permissioned blockchain network and invite business partners to record their transactions on a shared distributed ledger. These organizations will verify the consensus model to be used, supported what quantity they trust each other. On the far side trust, permissioned blockchain networks offer transparency and insight which will facilitate higher inform business selections and hold misbehaving parties responsible. This will expressly embrace auditing and oversight entities creating audits a relentless prevalence versus a natural event. Some permissioned blockchain networks support the power to by selection reveals dealings information supported a blockchain network user's identity or credentials. With this feature, some degree of privacy in transactions could also be obtained. For instance, it can be that the blockchain records that a dealings between 2 blockchain network users materialized, however the particular contents of transactions is merely accessible to the concerned parties. Some permissioned blockchain networks need all users to be approved to send and receive transactions (they don't seem to be anonymous, or maybe pseudo-anonymous). In such systems parties work together to attain a shared business method with natural

disincentives to commit fraud or otherwise behave as a foul actor (since they will be identified). If unhealthy behavior were to occur, it is well known wherever the organizations are incorporated, what legal remedies are on the market and the way to pursue those remedies within the relevant scheme.

3. Blockchain Limitations and Misconceptions:

There is an inclination to overhype and overuse most emerging technology. Several comes can attempt to incorporate the technology, even though it's superfluous. This stems from the technology being comparatively new and not well understood, the technology being encircled by misconceptions, and also the worry of missing out. Blockchain technology has not been immune. This section highlights a number of the restrictions and misconceptions of blockchain technology.

3.1 Immutability:

Most publications on blockchain technology describe blockchain ledgers as being immutable. However, this can be not strictly true. They're tamper evident and tamper resistant that may be a reason they are trusty for money transactions. They cannot be thought of fully immutable, because there are things during which the blockchain is changed. During this section we'll look at other ways during which the idea of fixity for blockchain ledgers is profaned. The chain of blocks itself can't be thought of fully immutable. for a few blockchain implementations, the foremost recently revealed, or 'tail' blocks are subject to being replaced (by a longer, various chain with totally different 'tail' blocks). As noted earlier, most blockchain networks use the strategy of adopting the longest chain (the one with the foremost quantity of labor place into it) as truth once there are multiple competitor chains. If 2 chains are competitor, however every embrace their own distinctive sequence of tail blocks, whichever is longer are adopted. However, this does not mean that the transactions at intervals the replaced blocks are lost – rather they'll have been enclosed during a totally different block or came back to the unfinished dealings pool. This degree of weak fixity for tail blocks is why most blockchain network users wait many block creations before considering dealings to be valid. For permission less blockchain networks, the adoption of an extended, alternate chain of blocks may be the results of a type of attack referred to as a fifty one foray. For this, the aggressor merely garners enough resources to outgo the block creation rate of remainder of the blockchain network (holding quite fifty one you look after the resources applied towards manufacturing new blocks). Betting on the size of the blockchain network, this might be an awfully price prohibitory attack administrated by state level actors. The price to perform this kind of attack will increase the more back within the blockchain the aggressor needs to create an amendment. This attack isn't technically tough (e.g., it is just continuance the traditional method of the blockchain implementation, however with hand-picked transactions either enclosed or omitted, and at a quicker pace), it's simply high-priced. For permission blockchain networks, this attack is mitigated. There's usually associate owner or pool of blockchain network users WHO permit commercial enterprise nodes to affix the blockchain network and take away commercial enterprise nodes from the blockchain network, which supplies them an excellent amount of management. There's less probably to be competitor chains since the owner or pool will force commercial enterprise nodes to collaborate fairly since non-cooperating commercial enterprise nodes will merely have their privileges removed. There are probably extra legal contracts in situ for the blockchain network users which can embrace clauses for misconduct and also the ability to

require legal action. Whereas this management is helpful to stop misconduct, it implies that any variety of blocks can be replaced through legitimate strategies if desired by the owner or pool.

3.2 Users Involved in Blockchain Governance:

The governance of blockchain networks deals with the foundations, practices and processes by that the blockchain network is directed and controlled. a standard idea is that blockchain networks are systems while not management and possession. The phrase “no one controls a blockchain!” is often exclaimed. This is often not strictly true. Permission blockchain networks are typically setup associated pass by an owner or pool that governs the blockchain network. Permission less blockchain networks are typically ruled by blockchain network users, commercial enterprise nodes, and software developers. Every cluster encompasses a level of management that affects the direction of the blockchain network’s advancement. Software developers produce the blockchain software system that’s utilized by a blockchain network. Since most blockchain technologies are open supply, it’s attainable to examine the ASCII text file, and compile it independently; it’s even attainable to make separate however compatible software system as a means of bypassing pre-compiled software system free by developers. However, not each user can have the flexibility to try and do this, which implies that the developer of the blockchain software system can play a large role within the blockchain network’s governance. These developers might act within the interest of the community at massive and are control responsible. For instance, in 2013 Bit coin developers released a brand new version of the foremost well-liked Bit coin shopper that introduced a flaw and began two competitive chains of blocks. The developers had to make a decision to either keep the restructure (which had not however been adopted by everyone) or revert to the recent version. Either selection would end in one chain being discarded—and some blockchain network user’s transactions becoming invalid. The developers created a selection, reverted to the recent version, and with success controlled the progress of the Bit coin blockchain. This example was associate unintentional fork; but, developers will by {design} design updates to blockchain software system to alter the blockchain protocol or format. With enough user adoption, a successful fork is created. Such forks of blockchain software system updates are typically mentioned at length and coordinated with the concerned users. For permission less blockchain networks, this is usually the commercial enterprise nodes. There’s typically an extended discussion associated adoption amount before an event happens wherever all users should switch to the fresh updated blockchain software system at some chosen block to continue recording transactions on the new “main” fork. For permission less blockchain networks, though the developers maintain an oversized degree of influence, users will reject a amendment by the developers by refusing to put in updated software system. Of the blockchain network users, the commercial enterprise nodes have important management since they produce and publish new blocks. The user base sometimes adopts the blocks created by the commercial enterprise nodes but isn’t needed to try and do therefore. A motivating aspect result of this is often that permission less blockchain networks are basically dominated by the commercial enterprise nodes and will interact a section of users by forcing them to adopt changes they will afflict to remain with the most forks. For permission blockchain networks, management and governance is driven by members of the associated owner or pool. The pool will govern WHO will be a part of the network, when members are off from the network, cryptography pointers for good contracts, etc. In summary, the software system developers, commercial enterprise nodes, and blockchain network users all play a part within the blockchain network governance.

3.3 Beyond the Digital:

Blockchain networks work extraordinarily well with the information inside their own digital systems. However, once they have to be compelled to act with the important world, there square measure some problems (often known as the Oracle downside). A blockchain network is an area to record each human computer file as well as device computer file from the important world; however there could also be no technique to work out if the input data reflects universe events. A device may be out of whack and recording knowledge that's inaccurate. Humans may record false info (intentionally or unintentionally). These issues don't seem to be specific to blockchain networks, however to digital systems overall. However, for blockchain networks that square measure onymous, managing knowledge deception outside of the digital network is particularly problematic. For example, if a crypto currency group action occurred to buy a real-world item there's no way to confirm inside the blockchain network whether or not the cargo occurred, without relying on outside device or human input. Many comes have tried to handle the 'Oracle problem' and make reliable mechanisms to ingest external knowledge in a very means that's each trustworthy and correct. For instance, comes like 'Oraclize' offer mechanisms to require net API knowledge and convert it into blockchain legible byte/opcode. Inside the context of localized applications, these comes could also be thought-about centralized as they supply single points of failure for attackers to compromise. As a result, projects like 'Mineable Oracle Contract' have recently arisen to modify oracle consumption in a very way that's impressed by blockchain technology and designed atop established agreement models and economic incentives.

3.4 Blockchain Death:

Traditional centralized systems square measure created and brought down perpetually, and blockchain networks will seemingly not diverge. However, as a result of their suburbanized, there's an opportunity that once a blockchain network "shuts down" it'll ne'er be absolutely clean up, which there might continuously be some lingering blockchain nodes running. A defunct blockchain wouldn't be appropriate for a account, since while not several publishing nodes, a malicious user might simply overpower the few publication nodes left and redo and replace any variety of blocks.

4. Application Considerations:

Since blockchain technology continues to be new, a great deal of organizations area unit staring at ways that to include it into their businesses. The concern of missing out on this technology is sort of high, and most organizations approach the matter as "we wish to use blockchain somewhere, wherever will we tend to do that?" that ends up in frustrations with the technology because it can't be applied universally. A better approach would be to 1st perceive blockchain technology, wherever it fits, and then identify systems (new and old) that will match the blockchain paradigm. Blockchain technology solutions could also be appropriate if the activities or systems need options such as:

- Several participants
- Distributed participants

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- wish or want for lack of sure third party
- Work flow is transactional in nature (e.g., transfer of digital assets/information between parties)
- A desire for a globally scarce digital symbol (i.e., digital art, digital land, digital property)
- A desire for a localized naming service or ordered written account
- A desire for a cryptographically secure system of possession
- A desire to cut back or eliminate manual efforts of reconciliation and dispute resolutions
- A desire to modify real time observance of activity between regulators and controlled entities
- A desire for full source of digital assets and a full transactional history to be shared amongst participants

Several agencies and organizations have developed guides to assist verify if a blockchain is suitable for a specific system or activity, and which sort of blockchain technology would be of most profit. During this section, some articles and recommendation area unit highlighted from many completely different sectors – federal, academia, technical publications, technology websites, and software developers. The U.S Department of Office of Homeland Security (DHS) Science & Technology board has been investigation blockchain technology and has created a multidimensional language to assist one verify whether a blockchain could also be required for a development initiative. The multidimensional language is reproduced here, with permission.

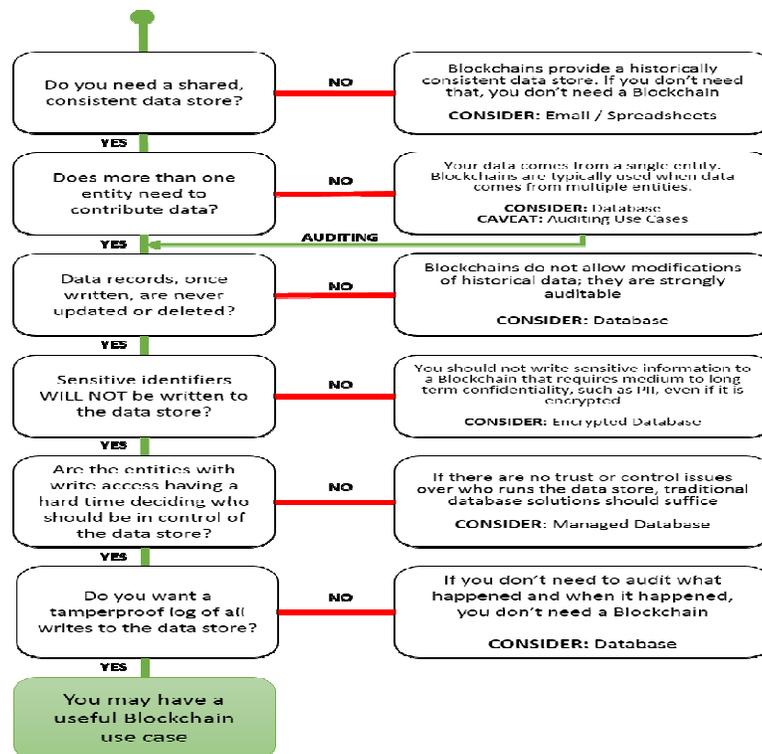


Figure 1 - DHS Science & Technology Directorate Flowchart

5. Additional Blockchain Considerations:

When deciding whether or not to utilize a blockchain, one should take into thought extra factors and confirm if these factors limit one's ability to use a blockchain or a selected sort of blockchain:

- **Information Visibility**

- Permission blockchain networks might or might not reveal blockchain information publicly. The information might solely be offered to those among the blockchain network. Consider eventualities wherever information is also ruled by policy or rules (such as personally distinctive data (PII) or General information Protection Regulation (GDPR) regulations). Information like this might or might not be applicable to store even among a permission blockchain network.
- Permission less blockchain networks will enable anyone to examine and contribute to the blockchain the information is mostly public. This results in many queries that must be thought-about. Will thin formation for the applying ought to be offered to everyone? Is there any hurt to having public data?

- **Full transactional history:**

Some blockchain networks give a full public history of a digital plus – from creation, to each dealing it's enclosed in. This feature is also beneficial for a few solutions, and not helpful for others.

- **Faux information Input** – Since multiple users' square measure contributive to a blockchain, some could submit false information, mimicking information from valid sources (such as detector data). It's tough to automatize the verification of information that enters a blockchain network. Good contract implementations might give extra checks to assist validate information wherever doable.

- **Tamper evident and tamper resistant information** – several applications follow the "CRUD" (create, read, update, delete) functions for information. With a blockchain, there's solely "CR" (create, read). There square measure strategies that may use to "deprecate" older information if a newer version is found, however there's no removal method for the initial information. By using new transactions to amend and update previous transactions, information will be updated where a providing a full history. However, though replacement dealing marked associate degree older dealing as "deleted" – {the information the info the information} would still be gift within the blockchain data, though it's not shown among associate degree application process the information.

- **Dealings Per Second** – Transaction process speed is extremely smitten by the consensus model used. Presently transactions on several permission less blockchain networks don't seem to be dead at constant pace as different data technology solutions due to a slow publication time for blocks (usually in terms of seconds, however generally minutes). Thus, some retardation in blockchain dependent applications might occur whereas waiting for knowledge to be announced. One should rise if their application will handle comparatively slow transaction processing?

- **Compliance** – the utilization of blockchain technology doesn't exclude a system from following any applicable laws and laws. As an example, there square measure several compliance

considerations with regards to legislation and policies tied to PII or GDPR that establish that sure info mustn't be placed on the blockchain. Additionally, certain countries might limit the sort of information which will be transferred across its geographic boundary. In alternative instances, sure legislation might dictate that the "first write" of financial transactions should be written to a node that is gift inside their borders. In any of those cases, a public, permission less chain is also less acceptable, with a permission or hybrid approach needed to satisfy restrictive wants. An additional example of laws and laws square measure for any blockchain network that manages federal records. Federal records square measure subject to several laws and laws. Federal agencies themselves should follow specific federal pointers once utilizing blockchain technology.

• **Permissions** – For permission blockchain networks, there square measure issues around the permissions themselves

- Roughness – do the permissions inside the system provide enough roughness for specific roles that users may have (in a way like Role-Based Access Control methods) to perform actions inside the system
- Permission blockchain networks provide additional ancient roles like administrator, user, validate, auditor, etc.
- Administration – World Health Organization will administer permissions? Once permissions square measure administered to a user, will they simply be revoked?

• **Node Diversity** – A blockchain network is barely as sturdy because the combination of all the existing nodes collaborating within the network. If all the nodes share similar hardware software, geographic location, and electronic communication schema then there exists an explicit quantity of risk related to the chance of undiscovered security vulnerabilities. This risk is satisfied through the decentralization of the network of heterogeneous devices, which may be outlined as "the non-shared characteristics between anybody node and therefore the generalized set".

5. Conclusions:

Blockchain technology could be a new tool with potential applications for organizations, enabling secure transactions while not the requirement for a central authority. Beginning in 2009¹³, with Bit coin leveraging blockchain technology, there has been Associate in nursing increasing range of blockchain technology-based solutions. The first applications were electronic money systems with the distribution of a worldwide ledger containing all transactions. These transactions are secured with cryptographically hashes, and transactions are signed and verified exploitation asymmetric-key pairs. The group action history efficiently and firmly records a sequence of events in an exceedingly approach that any arrange to edit or amendment a past group action also will need a calculation of all later blocks of transactions. The use of blockchain technology remains in its early stages; however it's designed on wide understood and sound cryptographically principles. Currently, there's a great deal of promotion round the technology, and many projected uses for it. Moving forward, it's doubtless that the promotion can die down, and blockchain technology can become simply another tool which will be used.

As elaborated throughout this publication, a blockchain depends on existing network, cryptographically, and recordkeeping technologies however use them in an exceedingly new manner. It'll be necessary that organizations are ready to check up on the technologies and each the benefits and drawbacks of using them. Once a blockchain is enforced and wide adopted, it should become troublesome to change it. Once information is recorded in an exceedingly blockchain, that information is sometimes there forever, even when there is a slip. Applications that utilize the blockchain as a knowledge layer work round the reality that the particular blockchain information cannot be altered by creating later blocks and transactions act as updates or modifications to earlier blocks and transactions. This computer code abstraction permits for modifications to operating information, whereas providing a full history of changes. For a few organizations these are fascinating options. For others, these is also deal breakers preventing the adoption of blockchain technology. Blockchain technology remains new and organizations ought to treat blockchain technology like they would the other technological answer at their disposal--use it solely in applicable situations.

6. References:

- [1] Clarke, A.C., "Hazards of Prophecy: The Failure of Imagination," from Profiles of the Future: An Inquiry into the Limits of the Possible, 1962.
- [2] Lamport, Leslie. "The Part-Time Parliament." ACM Transactions on Computer Systems, vol. 16, no. 2, Jan. 1998, pp. 133–169., <https://dl.acm.org/citation.cfm?doid=279227.279229>.
- [3] Narayanan, A., Bonneau, J., Felten, E., Miller, A., and Goldfede, S., Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction, Princeton University Press, 2016.
- [4] Nakamoto, S., "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008. <https://bitcoin.org/bitcoin.pdf>
- [5] National Institute of Standards and Technology, Secure Hash Standard (SHS), Federal Information Processing Standards (FIPS) Publication 180-4, August 2015. <https://doi.org/10.6028/NIST.FIPS.180-4>
- [6] National Institute of Standards and Technology (NIST), Secure Hashing website, <https://csrc.nist.gov/projects/hash-functions>
- [7] "Hash per Second." Bitcoin Wiki, http://en.bitcoin.it/wiki/Hash_per_second.
- [8] National Institute of Standards and Technology, SHA-3 Standard: PermutationBased Hash and Extendable-Output Functions, Federal Information Processing Standards (FIPS) Publication 202, August 2015. <https://doi.org/10.6028/NIST.FIPS.202>
- [9] National Institute of Standards and Technology (NIST), Digital Signature Standard, Federal Information Processing Standards (FIPS) Publication 186-4, July 2013. <https://doi.org/10.6028/NIST.FIPS.186-4>
- [10] "LDAP.com." LDAP.com, <https://www.ldap.com>.
- [11] "How Is the Address of an Ethereum Contract Computed?" Ethereum Stack Exchange, 29 Jan. 2016, 22:14, <https://ethereum.stackexchange.com/questions/760/how-is-the-address-of-anethereum-contract-computed>.
- [12] Bahsoun, J.P., Guerraoui, R., and Shoker, A., "Making BFT Protocols Really Adaptive," 2015 IEEE International Parallel and Distributed Processing Symposium, Hyderabad, India, pp. 904-913, 2015. <https://doi.org/10.1109/IPDPS.2015.21>
- [13] Lamport, L. "Time, Clocks, and the Ordering of Events in a Distributed System." Communications of the ACM, vol. 21, no. 7, January 1978, pp. 558–565., doi:10.1145/359545.359563. <https://amturing.acm.org/p558-lamport.pdf>.
- [14] Todd, P. Bitcoin Improvement Protocol (BIP) 65, "OP_CHECKLOCKTIMEVERIFY," October 1, 2014. <https://github.com/bitcoin/bips/blob/master/bip-0065.mediawiki>
- [15] Wong, J. and Kar, I., "Everything you need to know about the Ethereum 'hard fork,'" Quartz Media, July 18, 2016. <https://qz.com/730004/everything-you-need-to-know-about-the-ethereum-hard-fork/>

- [16] Chen, L., Jordan, S., Liu, Y.-K., Moody, D., Peralta, R., Perlner, R., and SmithTone, D., Report on Post-Quantum Cryptography, National Institute of Standards and Technology Internal Report (NISTIR) 8105, April 2016. <https://doi.org/10.6028/NIST.IR.8105>
- [17] Szabo, N. "Smart Contracts," 1994.
<http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html>
- [18] Mell, P., Kelsey, J., and Shook, J., "Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness." October 7, 2017. https://doi.org/10.1007/978-3-319-69084-1_31
- [19] "Majority Attack." Bitcoin Wiki, https://en.bitcoin.it/wiki/Majority_attack.
- [20] Greenspan, G. "The Blockchain Immutability Myth." CoinDesk, May 9, 2017, <https://www.coindesk.com/blockchain-immutability-myth/>.
- [21] Narayanan, A., "Analyzing the 2013 Bitcoin fork: centralized decision-making saved the day," MultiChain, July 28, 2015. <https://freedom-totinker.com/2015/07/28/analyzing-the-2013-bitcoin-fork-centralized-decisionmaking-saved-the-day>
- [22] Buck, J. "Blockchain Oracles, Explained." Cointelegraph, October 18, 2017, <https://cointelegraph.com/explained/blockchain-oracles-explained>
- [23] <https://medium.com/@kleffew/truthpoint-angelhacks-dc-submission5569252d795a>
- [24] Greenspan, G., "The Blockchain Immutability Myth," MultiChain, May 4, 2017. <https://www.multichain.com/blog/2017/05/blockchain-immutability-myth/>
- [25] de Vries, A. "Bitcoin's Growing Energy Problem." Joule, vol. 2, no. 5, 16 May 2018, pp. 801–805., <https://doi.org/10.1016/j.joule.2018.04.016>.
- [26] Deetman, S., "Bitcoin Could Consume as Much Electricity as Denmark by 2020," Motherboard, March 29, 2016. https://motherboard.vice.com/en_us/article/bitcoin-could-consume-as-much-electricity-as-denmark-by-2020
- [27] Hern, A., "Bitcoin mining consumes more electricity a year than Ireland," The Guardian, November 27, 2017. <https://www.theguardian.com/technology/2017/nov/27/bitcoin-mining-consumeselectricity-ireland>
- [28] Power Compare, <https://powercompare.co.uk/bitcoin/>
- [29] Loh, T. "Bitcoin's Power Needs May Be Overblown." Bloomberg.com, Bloomberg, January 16, 2018, <https://www.bloomberg.com/news/articles/2018-01-16/bitcoin-s-power-needs-may-be-overblown-recalling-pot-growing>.

BLOCKCHAIN FUNDAMENTALS PRESENTATION

M.Divya Sree

Assistant Professor, Faculty Of Computer Science

A.V. College Of Arts, Science And Commerce

Mobile no: 9505096391

Email. divya.mende39@gmail.com

ABSTRACT

A block chain is a public ledger to which everyone has access but without a central authority having control. It is an enabling technology for individuals and companies to collaborate with trust and transparency. One of the best know applications of block chains are the cryptographic currencies such as Bitcoin and others, but many other applications are possible. Block chain technology is considered to be the driving force of the next fundamental revolution in information technology. Many implementations of block chain technology are widely available today, each having its particular strength for a specific application domain. The tutorial provides the participants with insights and practical experience on Block chain technology and applications in practice, as well as theory based exploration of possible business cases.

Keywords: Blockchain, Bitcoin, Cryptographic currency, Block chain applications

1. INTRODUCTION:

The block chain is defined as an open ledger that offers decentralization to the parties. In addition, it also offers transparency, immutability, and security. It has many features including being open, distributed, ledger, P2P and permanent. Block chain term was first introduced in the white paper of Bitcoin in 2009 by Satoshi Nakamoto. From there, it has come a long way as more and more organizations are interested in it. Right now, Bitcoin is on its way to implementing lightning network and other advanced features. A blockchain is a public ledger to which everyone has access but without a central authority having control. It is an enabling technology for individuals and companies to collaborate with trust and transparency. One of the best know applications of blockchains are the cryptographic currencies such as Bitcoin and others, but many other applications are possible. Blockchain technology is considered to be the driving force of the next fundamental revolution in information technology. Many implementations of blockchain technology are widely available today, each having its particular strength for a specific application domain. The tutorial provides the participants with insights and practical experience on Blockchain technology and applications in practice, as well as theory based exploration of possible business cases.

2. Block chain Work:

The function of a blockchain is straightforward. As it is a peer-to-peer network, a user needs to start a transaction. Once done, a block is allocated to the said transaction. The transaction block is also broadcasted to the network, and all the nodes in the network get the said information. The block is then mined and validated. It is also added to the chain, followed by a successful transaction.

3. Remarkable Benefits of Blockchain Technology

Blockchain technology is feature-rich. It is also extremely beneficial. For example, it lets the user do faster settlement compared to traditional methods. It is also immutable and more secure. When compared to a traditional network, blockchain technology is more capable and have improved network capacity. With decentralization built-in, it can be used to build a shared and distributed blockchain.

4. Public vs Private Blockchain Network

The slides discuss the difference between public and private blockchain network. The main difference between them is how they function. Public blockchain network is permissionless whereas the private blockchain is permissioned one. This means that the public blockchain is an open network which is not controlled by anyone. Anyone can access it. The private blockchain, on the other hand, is permissioned which means that there is an authority that manages who can use the network.

5. Centralized vs Decentralized vs Distributed Network: An Overview

There is a vital difference between centralized, decentralized and distributed network. That's what the slide is all about.

- Centralized: All the nodes come under a single authority
- Decentralized: There is no centralized authority and everyone can take part in the network.
- Distributed: Independent nodes interact with each other. Also, each node is interconnected.

Want to know more? Check out the ultimate blockchain guide.

6. Must know blockchain Terms

It is hard for a beginner to know blockchain terms and this can easily make them confused. This slide list 10 important must know blockchain terms out there. For example, it covers bitcoin, airdrop, App, ICO and others. Learn about all the popular blockchain definitions in detail.

7. Smart Contract

7.1 Smart Contract Explained

Smart contracts are similar to a legal document and create terms between two parties. The two parties that are dealing with using smart contracts. Also, the contracts use public ledger for storage purposes. Smart contracts are triggered when a condition is met, and are completely autonomous. It just executes based on the code that defines the pre-condition. To make sure that they work as intended, they are analyzed and managed by regulators. It is also helpful in understanding trends and predicts market uncertainties.

7.2. Smart Contract Work

Smart contract works between two parties. It is used to do buyer and seller matchmaking. Furthermore, it can be used for transactions. Banks and other institutes can use smart contracts to receive assets and distribute it.

The three key features of smart contracts include registered, automated settlement of contracts and there is no need for the third party.

7.3 Advantage of Smart Contracts

The slide discusses the advantages of smart contracts. The few advantages include total transparency, no paperwork, total transparency, trustworthy, guaranteed outcomes and so on.

7.4 Disadvantages of Smart Contracts

Smart contracts are not free from disadvantages. It does have some disadvantage. Few of the major disadvantages include error, confidentiality and rouge contracts.

Smart contracts are made up by humans. This makes them open for human-errors. Also, they are not 100% confidential. It can leak some vital info which can be read by a 3rd party. Not only that, there are rouge contracts that can act like a real one and make fraud possible.

7.5 Smart Contracts Use Cases

There are many uses cases for smart contracts. They can be used in different verticals, and can also be used to make things efficient. Few of its main uses cases include trading activities, supply chains, copyright protection, real estate market, government voting, and so on. It also has very useful use cases in Internet-of-things(IoT) where it can help protect the network as well.

Smart contracts are hard to grasp? Don't worry, check out the detailed guide on smart contracts

8. Verticals of Blockchain Transformations

Currently, the whole industry is going through a blockchain transformation. Its impact is seen everywhere. That's why the slide shares the nine verticals of blockchain transformation. They include the following:

1. Technology
2. Law and Crime
3. Government Service
4. Finance
5. Human Rights
6. Entertainment
7. Media
8. Transportation
9. Contracts

Digital transformation is already here. Check more about it here.

8. 2017-2018 Leading Sectors

There are also many sectors that are leading in the blockchain technology implementation. The two sectors that are leading the most include Fintech and supply chains. However, that's not all as there are other sectors which are slowly but steadily adopting blockchain technology including healthcare, shipping, retail, insurance, and mining. With time, we will see almost every sector to use some form of blockchain technology. Few sectors will see better implementation due to better suitability than other sectors.

Enterprises are very keen on implementing blockchain. We covered them in detail in this post.

10. Web 3.0: The Successor of Web 2.0

Blockchain will also begin Web 3.0 which is the 3rd generation of the internet. The internet will take advantage of the blockchain network and bring a truly decentralized network that is way more advanced than what we have right now. The current web is more focused on server-based databases and hence provide limited functionality. The new internet will be more focused on users which means that it will be better connected and offer a secure platform for everyone out there.

11. Web 3.0 Benefits

This slide continues with the topic of Web 3.0 and lists Web 3.0 benefits. As discussed earlier, Web 3.0 offers better functionality and features. It will be permissionless which means that there will be no centralized authority controlling it. It will also be free from any monopoly and will also provide tons of privacy to the users. The network is secure, and the data ownership stays with the end users who can keep it to themselves or sell it to the media companies. The Web 3.0 will also be ubiquitous and offer a semantic web.

12. Remember: Centralized vs Decentralized Internet

The core difference between the centralized and decentralized network is the absence of central authority. It is a decentralized internet which has its unique data flow, new business models and dApps. The slide also presents the difference in a visual way.

13. Conclusion

This leads us to the end of the blockchain fundamentals presentation. If you are looking for a blockchain presentation ppt, then you can download the presentation and save it to your machine. This slide not only covers the introduction to the blockchain but also introduces the reader to different new concepts, ideas, and information. You can also do blockchain presentation pdf download and use it as a reference for learning further advanced topics. Furthermore, you can share the blockchain slides pdf.

We request you to share the article with all your friends and help them know about blockchain. Also, don't forget to share your thoughts about the slide below. We are listening.

REFERENCE

Leo van Moergestel and Ander de Keijzer Tutorial Abstract-

<http://ceur-ws.org/Vol-2105/10000439.pdf>

<https://101blockchains.com/blockchain-fundamentals-presentation/#prettyPhoto>

<https://www.ibm.com/blogs/blockchain/2018/02/top-five-blockchain-benefits-transforming-your-industry/>

BLOCKCHAIN TECHNOLOGY IN SUPPLY CHAIN MANAGEMENT

VARADHA PALLY VINAY REDDY
245116733065, B.E(IV/IV) Sem-1
Dept. of CSE, MVSr Engineering College, Hyderabad
MAIL-ID: varadhapallyvinay26@gmail.com.

ABSTRACT:

Blockchain technology is emerging in every field. Blockchain is chain of blocks which are connected with cryptographic hash of the previous blocks with distributed, decentralized and immutable properties with each block containing fields like block no, data, timestamp etc. Supply Chain Management is collection of activities, organizations, suppliers etc, which maintain the flow of information about the products from it's initial stages to until it reaches the consumer. With the implementation of blockchain technology in supply chain it provides data transparency, traceability, reliability across the supply chain. The paper discourses the different case studies that we studied in which blockchain technology is implemented and produced efficacious results . The paper shows creation of Simulated general blockchain and how its implemented in supply chain which provides transparency, immutability and distributed properties.

KEYWORDS: Supply chain, Blockchain, Transparency, Network effects and Scalability.

1.INTRODUCTION:

One of the most appealing benefits of using blockchain for data is that it allows the data to be more interoperable. Due to this, it becomes easier for companies to share information and data with manufacturers, suppliers, vendors, and customers. Transparency in Blockchain helps reduce delays and disputes while preventing goods from getting stuck in the supply chain. As each product can be tracked in real-time, the chances of misplacements are rare. If any, we can easily find them.

Blockchain offers scalability due to it's distributed (shared public ledger) network through which any large database is accessible from multiple locations from any parts of the world. It also provides higher standards of security and the ability to customize according to the data feed and needs of the user. Moreover, blockchains can be created privately too, which will allow the data to be accessed explicitly between the parties who have permission for it.

The value of adopting blockchain technology can be taken from the fact that it has the potential to connect different ledgers and data points while maintaining the data integrity among multiple participants.

The properties of transparency and immutability of blockchain technology make it useful for eliminating fraud in the supply chain and maintaining the integrity of the system.

Other than these, a few other benefits of adopting Blockchain technology in the supply chain industry are:

- 1.Reduce or eliminate fraud and errors
- 2.Minimize courier costs
- 3.Reduce delays from paperwork
4. Identify issues faster
- 5.Increase consumer and partner trust

How blockchain helps in enhancing supply chain management?

An effective supply chain management depends on several parameters of the chain like transparency and privacy at any instant of time which includes where goods are at any given time, and the origin of all the parts of the chain. By using blockchain technology-based records, we can store and share that information for every component by using shared public ledger so that we can track the flow, how it was created and where it is at any moment.

For business, as it depends on trust or reliability, they provide to the consumers by providing authentic products with transparency makes them more trustworthy. Benefits: By using blockchain, we can efficiently improve traceability and transparency, and a consumer can access to information of the product from it's the initial stage to until it reaches the consumer. Measurement of supply chain management performance is often described in terms of objectives such as quality, speed, dependability, cost, and flexibility.

2. MOTIVATION

We often visit supermarkets for our daily products, In that super I have enthralled seeing apples from US, intrigued by those there must be some technology which provides trust which shows they are authentic apples(which are from US).This made a constant driving force in exploring blockchain technology in supply chain management. I have been doing research in this field for the last 1 year.

3 RELATED WORK

A prominent objective of supply chain management is also to reduce risks. Among the various risks that different organizations faces include relational risks such as a business partner's engagement in opportunistic behavior (e.g.,cheating, distorting information)[1](Baird & Thomas, 1991; Bettis & Mahajan, 1985). According to Svensson (2000)[1], the sources of risk in supply chains can be classified into two main categories,

namely, atomistic or holistic. To deal with atomistic sources of risk, a selected and limited part of the supply chain need to be looked at to assess risk. This approach is suitable for components and materials that are of low-value, less complex, and easily available. On the other hand, holistic sources of risk require an overall analysis of the supply chain to assess risk. This approach is preferable for high-value, complex, and rare components and materials.

Application 1: FOOD SAFETY

Food safety is one of the prominent areas where the research is going rapidly. Approximately, for one percent reduction in foodborne disease in the United States, it saves the US economy \$700 million.

CASE 1: FOOD SUPPLY CHAIN

Alibaba teamed up with AusPost, Blackmores, and PwC to explore the use of blockchain to fight food fraud, which involves selling lower quality foods and often with counterfeit ingredients. The four companies aim to develop a “Food Trust Framework” to help improve integrity and traceability on the global supply chains. They are working to develop a pilot blockchain solution model that participants across the [2] the supply chain can use

With the increase of complexity involved in the food supply chain it is becoming a difficult job to trace the food through it’s chain.

Consumers are increasingly becoming aware and are demanding transparency in terms of the food they consume. Presently, about only 12 percent of consumers trust the brands that they purchase food from a company which produces that product. In a consumer driven society, the consumers has to know information of their food while 94 percent of consumers state that it is highly essential for them to learn about all the information related to the food products that they buy.

Blockchain resolves the issues of a convoluted supply chain by providing neutrality in the platform. Since there are no third parties involved in the transaction authorization i.e. which is a decentralized and everything works based on a consensus, both, the users and the operators of the system had to follow a set of rules to keep the system working in a efficient way.

Benefits of Blockchain in Food Safety:

- 1.Enhanced food safety.
- 2.Less Food waste.
3. Detect food fraud.

We are in dire need of reducing food waste because one-third of all food that's produced on the planet goes to waste.

CASE 2:WALMART

Walmart in collaboration with IBM developed blockchain technology to monitor the consumer products by adding a RFID sensor tags to it's products.so that when a consumer wants to know about the product information like where it was produced, what are the ingredients that are used, expiry date and the suppliers etc. On May 31 2017,Walmart released the results of using blockchain technology in food supply chain and it reported that it reduced the time taken to track the food product form days to minutes[4].

And blockchain is also used in tracking the pork in china to know the illegal selling of pork[5].With the use of barcodes, sensors in supply chain provided the relevant data about the storing information about the product which produced efficacious results in transparency and traceability across the supply chain.

CASE 3: Everledger

Everledger, an London based startup implemented blockchain technology in the tracking of diamonds to eliminate any illegal selling in the markets, It produced propitious results in providing transparency across the supply chain which lead to reduced risks and frauds[6].Evenledger also used blokchain technology in tracking the wines by adding RFID tags for each bottle. As the bottle moves through the supply chain it stores the information of suppliers,location etc.which is used to track the bottle at any given time[7].

CASE Modum:

Modum, an swiss startup in collaboration with University of Zurich implemented blockchain technology to ensure safe delivery of medicines[8].

As we know that many medicines needs to be transported within the temperatures that medicines needs to be kept(sustained), humidity etc. With the conventional record keeping requires a tedious effort of storing information and monitoring the products. When the medicine reaches the destination, the information about the medicine is transferred to the ethereum based blockchain in which it is checked with all the requirements of the product with the standard requirements stored in smart contract[9]. If any medicine, not satisfying the requirements in smart contract are eliminated.so that we can get the medicine in proper conditions.

CASE 4:Intel's solution in food supply chain

Intel with the use of Sawtooth hyper ledger technology tracked the seafood supply chain[10]. In the sawtooth hyperledger technology, when a fisherman caught a fish it is stored in shared public ledger and upon successive supply of seafood is also stored like information about suppliers, location with the use of IoT sensors etc. With the use of above technology illegal supply of seafood is reduced greatly, which in turn increases income of the country.

Application 2: TagItSmart Safeguards Wine Supplies with a Blockchain Logistics Tool

It is estimated that nearly 30,000 bottles of illegitimate wine are sold every hour in China. Many of these wines are mixed with a variety of dangerous additives that can be detrimental to health. Original in conjunction with TagItSmart, conducted the first phase of their pilot program, which tracked more than 15,000 unique wine bottles[11]. The eventual goal is to put an end to illegitimate wine using blockchain transparency capabilities. By a simple scan of a QR code which will be placed on each bottle, consumers will be able to know every detail regarding their purchase.

OriginTrail's decentralized, the blockchain-based protocol provides a foundation for tracking wine from the vineyard to the point of sale as a safety measure alongside TagItSmart's anti-counterfeit technology, utilizing photochromic ink together with unique QR codes. The TagItWine project uses TagItSmart platform to implement a pilot system for brand protection and anti-counterfeiting in the wine industry. TagItWine project is led by the University of Donja Gorica, and the solution is being tested in collaboration with the wine producer Plantaze from Montenegro, an up-and-coming wine region in Southern Europe.

Plantaze's premium and ultra-premium wines, consisting mostly of red and sparkling wines, are sold in more than 35 countries all over the world. The company just expanded its distribution to the Chinese market, which highly values European wines. That is why it was actively seeking a way to prove the veracity and authenticity of origin for each bottle.

Connecting Advanced Anti-Counterfeit Systems With Blockchain-Based Protocol:

In this pilot project, each wine bottle will get a unique digital identity and is equipped with a smart tag that uses a unique QR code with additional specific information printed using photochromic ink. By applying a relatively strong light source (such as the flash on a mobile phone) the photochromic properties of the ink enable the customer to see previously invisible information (like a particular character) that, when paired with the unique QR code and checked on the original platform, can be used to fully authenticate a specific bottle.

The original protocol keeps track of historical information about each bottle by interacting with the TagItSmart platform. Throughout the life cycle of the wine bottle, consumers can check its authenticity and get the details about that particular bottle of wine. Each scan of

the QR code is logged, and the information about that particular bottle is updated and used by the solution's heuristics. For example, a consumer can check a bottle with their mobile phone while in a retail store – even when a smart tag is valid – and the anti-counterfeiting system will then identify a problem – like whether or not a bottle with that same tag has been previously labeled as purchased or consumed – if there is one. This pilot was performed in two phases, the data integration phase – which has been successfully completed – and consumer-facing interface implementation, which will introduce original checks in the TagItSmart application (available soon)[11]. Plantaze produces over 40 different wines, made of 26 wine grape varieties. It is the leading wine producer in the region with the production of 22 million kilos of wine grapes and sales of over 15 million wine bottles annually in 35 countries of the world. Original will start by tracing its premium product lines for export.

IMPLEMENTATION:

In this paper, I developed a general blockchain network with the use of a postman framework. Postman, an HTTP client is a user-friendly interface that is used to interact with the blockchain. Flask - a web framework, is used to develop a web application that encompasses (contain) blockchain technology. I created a transactions.json file, which stores the sender and receiver address and the information to send. The methods used in creating the blockchain network are:

Get_chain request: This request invokes the get_chain method to get the entire chain of blocks that we have mined till now.

Mine_block request: This request invokes the mine_chain method to mine the block i.e. to add new blocks to the chain by solving the valid hash with the use of proof of work mechanism. The first block in the chain is called "Genesis Block".

Firstly, I created a block that has to be mined where proof of work mechanism operates in which miners across the blockchain network participates to find the valid hash for the above created block. If a miner solves the cryptographic puzzle (valid hash) then, all the other miners in the network have to validate it. If the block is valid, it is added to the blockchain network

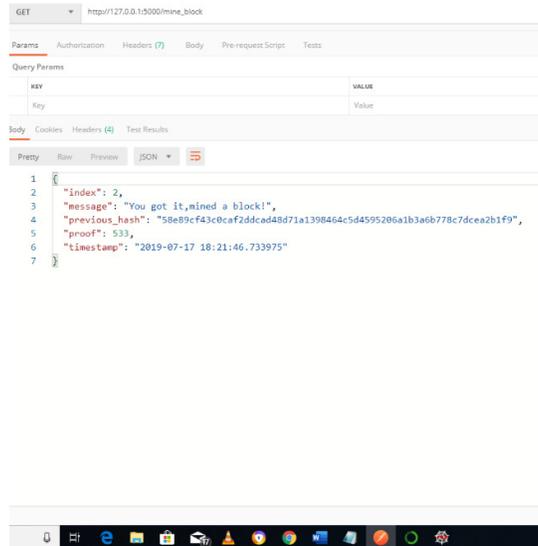


Fig 1.

Fig 1 shows output of mine_block method which gives message as “you have just mined the block”

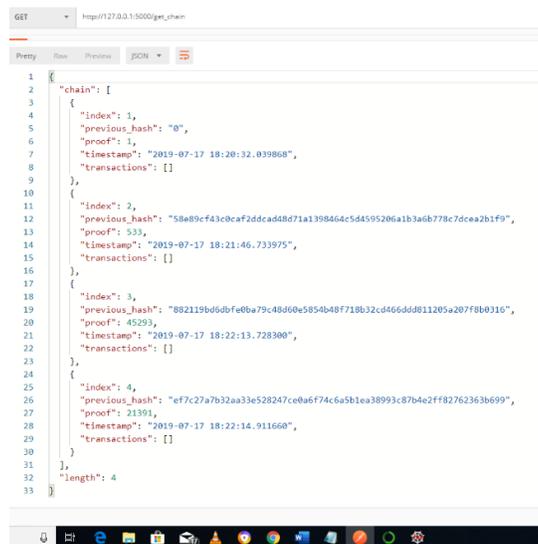


Fig 2

Fig 2 shows that after 4 mine_blocks requests then if we make get_chain request it gives the output as chain of blocks which are connected with previous blocks with Prev_hash field.

```
GET http://127.0.0.1:5001/mine_block

Pretty Raw Preview JSON

1 {
2   "index": 2,
3   "message": "Congratulations, you just mined a block!",
4   "previous_hash": "db5f4978ded48988835b980894a1907b74de023e83440cac725f00bdfbaf93f",
5   "proof": 533,
6   "timestamp": "2019-07-17 18:24:32.060554",
7   "transactions": [
8     {
9       "amount": 1,
10      "receiver": "vineeth",
11      "sender": "688f372dc2e4641986e11ba5fa5d7aa"
12    }
13  ]
14 }
```

Fig 3

Fig 3 shows output of mine_block method which gives message as “you have just mined the block” when request is made on “Vineeth” block.

```
POST http://127.0.0.1:5001/connect_node

Params Authorization Headers (9) Body Pre-request Script Tests
none form-data x-www-form-urlencoded raw binary GraphQL BETA JSON(application/json)

1 {
2   "nodes": [
3     "http://127.0.0.1:5002",
4     "http://127.0.0.1:5003"
5   ]
}

Body Cookies Headers (4) Test Results

Pretty Raw Preview JSON

1 {
2   "message": "All the nodes are now connected. The Blockchain now contains the following nodes:",
3   "total_nodes": [
4     "127.0.0.1:5003",
5     "127.0.0.1:5002"
6   ]
7 }
```

Fig 4

We have created a json file where we have address of the all the nodes in the chain i.e 3 in this case. In fig 4, if we make connect_node request then it gives output as “All nodes are connected”.

```
POST http://127.0.0.1:5001/add_transaction

Params Authorization Headers (9) Body Pre-request Script Tests
none form-data x-www-form-urlencoded raw binary GraphQL BETA

1 {
2   "sender": "vineeth",
3   "receiver": "vineeth",
4   "amount": 10
5 }

Body Cookies Headers (4) Test Results

Pretty Raw Preview JSON

1 {
2   "message": "This transaction will be added to Block 3"
3 }
```

Fig 5

We also created a json file which has the information about sender,receiver and amount for the transaction.In Fig 5,we made transaction between “Vinay” and “Vineeth” it gives output as “transaction will be added to block 3”

```
GET http://127.0.0.1:5001/get_chain

1 {
2   "sender": "vinay",
3   "receiver": "vineeth",
4   "amount": 10
5 }
6 }

body Cookies Headers (4) Test Results

Pretty Raw Preview JSON

1 {
2   "chain": [
3     {
4       "index": 1,
5       "previous_hash": "0",
6       "proof": 1,
7       "timestamp": "2019-07-17 18:20:50.593759",
8       "transactions": []
9     },
10    {
11     "index": 2,
12     "previous_hash": "db5f4978ded4898835b980894a1907b74de023e83440cac725f00bdfbaf93f",
13     "proof": 533,
14     "timestamp": "2019-07-17 18:24:32.060554",
15     "transactions": [
16       {
17         "amount": 1,
18         "receiver": "vinay",
19         "sender": "688f372dcd2e4641986e11ba5fa5d7aa"
20       }
21     ]
22   }
23 ],
24 "length": 2
25 }
```

Fig 6

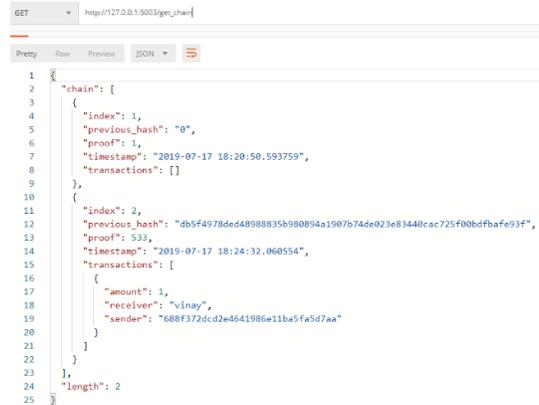
Fig 6 shows that If we make get_chain request from “vinay”(5000) node

```
GET http://127.0.0.1:5003/replace_chain

1 {
2   "message": "The nodes had different chains so the chain was replaced by the longest one.",
3   "new_chain": [
4     {
5       "index": 1,
6       "previous_hash": "0",
7       "proof": 1,
8       "timestamp": "2019-07-17 18:20:50.593759",
9       "transactions": []
10    },
11    {
12     "index": 2,
13     "previous_hash": "db5f4978ded4898835b980894a1907b74de023e83440cac725f00bdfbaf93f",
14     "proof": 533,
15     "timestamp": "2019-07-17 18:24:32.060554",
16     "transactions": [
17       {
18         "amount": 1,
19         "receiver": "vinay",
20         "sender": "688f372dcd2e4641986e11ba5fa5d7aa"
21       }
22     ]
23   }
24 ]
25 }
```

Fig 7

In fig 7 ,Due to the transparency and distributed property of the blockchain even the transaction is done between “vinay”(5000) and “vineeth”(5001), when the transaction is made on 5003 node we got the output as “node had different chains so the chain was replaced by the largest one”.



```
GET http://127.0.0.1:5003/get_chain|
Pretty Raw Preview JSON
1
2 "chain": [
3   {
4     "index": 1,
5     "previous_hash": "0",
6     "proof": 1,
7     "timestamp": "2019-07-17 18:20:50.593759",
8     "transactions": []
9   },
10  {
11    "index": 2,
12    "previous_hash": "db5f4978ded48988835b980894a1907b74de023e83440cac725f00bdfbaf93f",
13    "proof": 533,
14    "timestamp": "2019-07-17 18:24:32.060554",
15    "transactions": [
16      {
17        "amount": 1,
18        "receiver": "vinay",
19        "sender": "688f372dc2e4641986e11ba5fa5d7aa"
20      }
21    ]
22  }
23 ],
24 "length": 2
25 ]
```

Fig 8

Fig 8, shows the entire chain of the network when request is made by get_chain method due to the distributed and transparency properties of the supply chain.

Information passing mechanism:

we have created 3 nodes which are similar as if there were 3 computers; which were named as Vinay, with the address "http://0.0.0.0:5001/" ; "http://0.0.0.0:5002/" as Vineeth and "http://0.0.0.0:5003/" as Praveen in a nodes.json file.

When the Praveen has requested a mine_block method it gives the output as :

```
{ "index": 2,"message": "Congratulations, you just mined a block!"}
```

```
"previous_hash": "e0147a52b45c60c1b5356cbea6c
```

```
1f9bf3c1d8e20283f9e48ce6d7c4e38355ea8",
```

```
"proof": 533,
```

```
  "timestamp": "2019-07-04 18:12:02.761162",
```

```
  "transactions": [ {"amount": 1,"receiver": "praveen",
```

```
"sender":"ffc059f9bf9641e1b5afbb3c29585a82"}]]
```

With the use of transactions.json file which contains :

```
{"sender": "", "receiver": "", "information": }
```

When we use a add_transaction request with

```
{"sender": "vineeth", "receiver": "praveen", "info": "Hello!How do you do?"}
```

which gives output as

```
{"message": "The transaction will be added to Block 5"}
```

Then we have to request mine_block to validate

the hash with proof of work mechanism designed with output as :

```
{ "index": 4,
```

```
  "message": "Congratulations, you just mined a block!",
```

```
"previous_hash": "a94ef0d9cad0c3407394235db4f5
```

```
b2e80352251d2e9eefbfa19a4e8ca7fd007",
```

```
  "proof": 21391,
```

```
"timestamp": "2019-03-12 01:04:09.404833",
```

```
  "transactions": [{"info": "Hello!How do you do?"
```

```
    "receiver": "praveen",
```

```
    "sender": "vineeth"
```

```
  }, { "info": "Are you kidding?",
```

"receiver": "Hadelin",

"sender": "fea50ce20bc04090bd5b9d1bae0f7b13"}
}

4. LIMITATIONS OF USING BLOCKCHAIN

The Blockchain in Transport Alliance (BiTA) welcomed large industry leaders such as FedEx, BNSF Railway, UPS, and JD Logistics as new members. Together, they discussed the challenges that they would face because of integrating blockchain technology to the supply chain industry. These include network effects, the current lack of blockchain standards in the industry, the issues concerning blockchain and scalability, and the scarcity of technical talent.

Blockchain provides efficacious results in terms of transparency, traceability and distribution but it faces many issues which has to be addressed for efficient implementation of blockchain technology in supply chain management.

Limitations:

- a. Complexity: The amount of complexity involved in maintaining blockchain network through the supply chain is a daunting task.
- b. Lack of Scalability: Blockchain implementation requires high energy costs, which is one of the key reasons for it's slow pace of the development. Due to different economic status, many countries couldn't implement because of entire computerization and energy costs.
- c. Network affects : With the use of blockchain technology, the no of transactions that are processed is around(5 to 10) but with an conventional banking system provides approximately 2000 transactions or even more.
- d. Computational Costs: To make transaction within the network it requires huge energy which results in high energy consumption. It is estimated that to for a mining rig the amount of electricity consumption is approximately equal for New Zealand country per day consumption.

6. CONCLUSION

In a consumer driven society, providing trust is the prominent objective of any company, Blockchain technology with it's distributed, decentralized and immutable properties provides trust, reliability to the consumers by providing details about the products in every step of the supply chain.

Due to immutability property once the data is entered can't be changed which makes it as one time data insertion due to this property we have to be very careful about the information of the product. Being unequivocal at the transaction fees in the blockchain

mining makes it as unreliable and the consensus algorithms that are used, And there is huge scope of research in which blockchain technology should be embraced which has the ability to revolutionize the supply chain etc. By integration of blockchain, IoT, Machine Learning which provides propitious results for the security concerns in every field.

REFERENCES:

- 1.Goran, Svensson. (2000). A Conceptual Framework for the Analysis of Vulnerability in Supply Chains. *International Journal of Physical Distribution & Logistics Management*. 30. 731-750. 10.1108/09600030010351444.
2. Kshetri, Nir. "1 Blockchain's roles in meeting key supply chain management objectives." *International Journal of Information Management* 39 (2018): 80-89.
3. Bindi, T. (2017). Alibaba and AusPost team up to tackle food fraud with blockchain. <http://www.zdnet.com/article/alibaba-and-auspost-team-up-to-tackle-food-fraud-withblockchain/>. Bowen, F. E., Cousins, P. D., Lamming, R. C., & Faruk, A. C. (2001). The role of supply management capabilities in green supply. *Production and Operations Management*, 10(2), 174–189.
- 4.Kharif, O. (2016). Wal-Mart tackles food safety with trial of blockchain. Bloomberg. Retrieved from <https://www.bloomberg.com/news/articles/2016-11-18/wal-marttackles-food-safety-with-test-of-blockchain-technology>.
- 5.Higgins, S. (2017a). Walmart: Blockchain food tracking test results are 'very encouraging'. June 5 <http://www.coindesk.com/walmart-blockchain-food-tracking-test-resultsencouraging/>
- 6.<https://www.everledger.io/>
- 7.Mathieson, M. A. (2017). Blockchain starts to prove its value outside of finance. *Computer weekly*<http://www.computerweekly.com/feature/Blockchain-starts-to-prove-itsvalue-outside-of-finance>
- 8.Allen, M. (2017). How blockchain could soon affect everyday lives. Retrieved from http://www.swissinfo.ch/eng/joining-the-blocks_how-blockchain-could-soon-affecteveryday-lives/43003266.
- 9.Campbell, R. (2016). Modum.io's temperature-tracking blockchain solution wins accolades at kickstarter accelerator 2016. *Bitcoinmagazine*. Retrieved from <https://bitcoinmagazine.com/articles/modum-io-s-temperature-tracking-blockchain-solution-wins-accoladesat-kickstarter-accelerator-1479162773/>.

10. Del Castillo, M. (2017a). Intel demos seafood tracking on sawtooth lake blockchain. CoinDesk. Retrieved from <http://www.coindesk.com/intel-demos-seafood-trackingsawtooth-lake-blockchain/>.

11. <https://openledger.info/insights/blockchain-in-the-supply-chain-use-cases-examples/>

A Novel Architecture to Cloud based SCADA systems

Dr. Raman Dugyala(raman.vsd@gmail.com), N Hanuman Reddy(raju9009@gmail.com), Mr. Shrawan Kumar(shrawan.kumar@vardhaman.org)
Vardhaman College of Engineering
Hyderabad,India.

Abstract— SCADA systems are useful for real time monitoring of systems that are geographically widely distributed from a central location. Sensors and actuators of the system are located in the field and are accessed by the server using a data communication network. The companies involved in GENERATION TRANSMISSION or DISTRIBUTION of electrical energy are usually coordinated by a GRID operator for mutual advantage. To this end each company establishes a system for continuous monitoring of distributed data. A typical system comprises of Application, communication and database servers, a dedicated data communication network and a set of field devices (RTUs) that capture data from sensors and interfaces with communication server. Since this system comprises of distributed applications, REST – representational state transfer –architecture is adapted in general. Several vendors who offer cloud services make use of the REST architecture to provide services accessible by universal clients. Any dynamic changes in the grid that can implicate the system are to be incorporated in cloud architecture time to time. In our approach we propose to shift to internet-based data communication in place of the existing LAN architecture as our approach deals with cloud-based services. In addition, we propose to use IoT devices to capture data and operate actuators remotely. This architecture groups the instances of IoT core service, DynamoDB service and Lamda service and all their sub services as Cloud Resources in addition to deploying an AWS web service that will be used by end-users for viewing data or external users who may require data from this system.

Keywords— IoT, , Cloud, REST, RTUs, Smart Grid, Power, Energy

I. BACKGROUND

A. Application scenarios

Supervisory control and Data Acquisition systems are useful for real time monitoring of systems that are geographically widely distributed from a central location. Sensors and actuators of the system are located in the field and are accessed by the server using a data communication network.

B. Power Utilities GRID operation

In a power grid many independent companies operate. The companies can be involved in GENERATION TRANSMISSION or DISTRIBUTION of electrical energy and usually coordinated by a GRID[2] operator. These networks are very large and operate cooperatively for mutual advantage. Generally, a 15-minute schedule is prepared for the day for each generating unit (power to be generated) on the basis of the projected load demand by all distribution companies by the grid operator. Transmission companies provide necessary infrastructure capacity (transmission lines) for carrying the power. Distribution companies provide the necessary customer support to users and also responsible for billing and collection. In turn the generation and distribution companies get paid by them for their services.[1]

Main concerns in this operation are continuous balancing of actual generation and load at all times and adhering to power quality parameters such as maintaining voltage and frequency,

ensuring stability of the network for possible contingencies like faults leading to sudden change in load, loss of generation or transmission lines. To this end each company (in particular grid operator, transmission companies and distribution companies) establishes a system for continuous monitoring of distributed data. Each of them will acquire their own data as well as data from other systems as needed. Naturally the applications that use this data are different for grid operator transmission companies and distribution companies. As can be seen an architectural design that frees data representation from applications that use data will be very beneficial so also standardization of data exchange protocols.

C. Architecture of existing systems

A typical system comprises of

- servers for application/communication/database configured as a LAN
- a dedicated data communication network
- a set of field devices (RTUs) that capture data from sensors and interfaces with communication server.

On this LAN, situated physically at a central monitoring station, required number of clients are added for user interaction. Client side programs make use of the database or for configuring the system. Custom protocols are used for data transfer between RTU and communication server though some standards (Modbus) exist. Sensors usually are IEDs that are capable of providing data on a serial communication line. Application layer protocol (ICCP) is used for data exchange between systems which needs building necessary interfaces for the lower layers of communication stack. Application server is responsible for coordinating communication server for data transfer between RTUs and the system, validating the data received and populating it in data base.

D. Trends in COMPUTER APPLICATION Technologies

REST – representational state transfer –architecture is becoming very popular for adaptation in the distributed applications. The same is well presented in the Ph D thesis of Flemming as a set of constraints followed in building an application.

- Application will be of client- server type, clients (application programs/users) consuming server side services
- They will be stateless so that there will be no implications of ordering of execution at server multiple requests emanating from different users
- For efficiency the response to requests can be cached at either server or client and cache coherency is maintained by uniform protocol[4][5][6]
- All applications use a uniform interface between servers and clients. Typically, this is HTTP protocol over TCP/IP. That means only GET POST methods are used and application specific parameters are included in HEADERS and BODY of the HTTP message
- Applications can use CODE ON DEMAND
– scripts transferred as part of response by server executed at client.

- There can be layers of servers to generate the response.

CLOUD SERVICES - Incorporating the principles of REST there are several vendors who offer cloud services. In general, these services abstract RESOURCES globally accessible by NET clients. A typical resource can be an instance of compute service belonging to a user to instantiate and run several VIRTUAL MACHINES, a DATASET organized as an RDBMS or a TABLE, or indeed any service instance which finds a generic use. An abstract resource has a REPRESENTATION at the server and clients can perform ACTIONS on it which essentially modifies the representation. They can also dynamically create and use sub resources. For example, a DATABASE SERVICE[3] instance can manage multiple data bases (sub resources); an IOT service can manage multiple 'things' and 'topics'; compute service instance can manage multiple VMs. One can see a close resemblance to object oriented paradigm. Using cloud services one can build an application by creating abstract resources (cloud components) on the cloud and such components interact through actions.

E. Changes in grid and implications to systems

The following changes in the grid can affect the SCADA systems

- Introduction of renewable energy sources
- Ever increasing demand
- More energy storage systems
- Increased use of electric power for transportation
- Load management
- Demand response -power trading
- Active compensation – better control

II. OVERVIEW OF PROPOSED ARCHITECTURE

Main features of the proposed architecture:

- Shift to INTERNET for data communication and application developed as a set of interacting cloud components
- IOT devices for data capture and remote operation of actuators
- Cloud services to be used to create 'RESOURCES': AWS IOT-CORE instance, AWS DynamoDB table, AWS LAMDA functions, AWS S3(Glacier) instance, WEB application deployed on cloud

- Loose coupling among all components of the application
- Data definition independent of its consumption by various applications or users
- Lamda functions to expose data, as smart grid compatible XML documents, available as resources that can be invoked by all internal (having permission to AWS account) applications.
- A WEB application deployed on AWS cloud to data requirement of external users (not having permission to AWS account of the application).

III. ARCHITECTURE DESCRIPTION

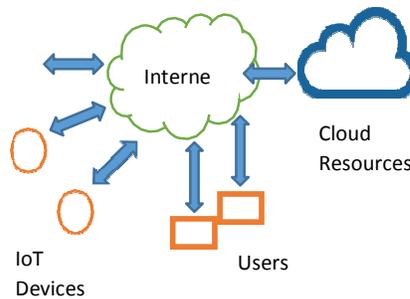


Fig. 1. Proposed architecture.

A. IoT Devices

As can be seen in the above diagram, Internet being pervasive, field devices can be connected to the NET wherever they are located geographically. These IOT devices are built using ARM cortex boards and so can capture data from the sensors. They can also issue digital outputs –commands – to actuators. IOT[8] devices interact with cloud resource – instance of AWS-IOT core service. Periodically they publish data using MQTT protocol to a topic- ‘ data’. They subscribe to another topic –‘command data’ to receive messages (and act) from users/applications who publish to this topic. Thus all IOT devices in field will interact with only one cloud resource namely instance of AWS IOT service.

B. DynamoDB Table

Another important cloud resource is a DynamoDB Table - ‘data table’, which holds the data acquired from the field devices. DynamoDB Table provides for efficient storage on cloud and access to retrieve data based on a combined partition key and range key. We

use deviceID as partition key and time as range key. IOT service provides a ‘rule engine’ that can be configured to insert the received data into the table automatically when a ‘message’ is received by the service from any device to the ‘ data’ topic.

C. Lamda Service

AWS lamda service provides for making available ‘lamda functions’ (cloud resource) that can be invoked by any application or triggered for execution on occurrence of a configured event. These functions are stateless as any other cloud service and have the advantage of setting up execution environment automatically. One main advantage of such server less function execution is automatic scaling of cloud resources with the frequency of invocation. Application development involves designing the lamda functions that take input request and return output response. Request and response can be JSON documents for internal users and XML documents for external users. Lamda functions can interact with any of the cloud resources belonging to the same AWS account –specifically the ‘data table’ in our case. Internal users are given necessary permissions to invoke these functions as required.

AWS[7] provides management of such permissions using AWS IAM service by defining users and roles.

D. AWS Web Application

External users not belonging to the owner of data- need not have access to the AWS account. For providing data to them a WEB application is deployed on the cloud (platform service from AWS). This will be a web application for browser users who have predefined functionality such as display of data or xml web service for programmatic interface.

Applications are developed either to be deployed on cloud or executed on client devices connected to the NET. An operator at monitoring center will be an external user accessing the URL of the WEB application deployed on the cloud. Deployed WEB Application provides for viewing the real time data or for issuing a command to operate an actuator in the field. Deployed WEB application uses lamda functions to generate response. For viewing data, lamda function queries data table to fetch required data. For operating an actuator, WEB application creates a 'message' and publishes it to topic – 'command data'. The field devices having subscribed to this topic interprets the received message and operate the actuator (code for this is part of ARM processor of the device). In fact, each of the device is a 'thing' of the IOT service and we make use of the thing shadow for modifying the device state – which facilitates provision for device connectivity - and thing attributes for configuring the device for its sensors and actuators. IOT service 'job' can be used for reconfiguring a device.

While data for the day required for monitoring the GRID is maintained in the data table, such data can be moved and efficiently stored in a bucket of S3 storage service (Glacier class) instance which can be retrieved by any analytic application.

E. Cloud Resources

Summary of Cloud Resources of the application and their interaction

- Instance of IOT core service, its sub resources- things: several field devices of same type having attributes that describe its configuration, - topics: 'sacda data' and 'command'
- Instance of DynamoDB[9] service, its sub resources table: 'data table'
- Instance of Lamda service, its sub resources: lamda functions to retrieve required data, to create and publish command messages, to validate received data from field devices, to retrieve data from archive storage as a stream
- WEB application [10] deployed on cloud: This will be used by end-users for viewing data or external users who may require data from this system. Such data will be provided as a web-service accessible from a public endpoint on the internet.

F. Reconfigurable field device

Built using ARM cortex M3. Application on ARM executes in FREE RTOS environment. A periodic task collects sensor data, forms MQTT [11][12][13][14]message and publishes to topic. Digital data state changes in the field can cause an event which in turn configured to invoke a lamda function (lamda service) that interacts with designated application to handle the change of state.

All such configuration[15] details are encoded as an xml document and device reconfigures itself when it is published by 'job' of a cloud application.

IV. CONCLUSION

In this paper, we have proposed a novel approach to cloud based SCADA systems making use of IoT Devices, Internet in place of the existing LAN architecture. Also Discussed the shortfalls of the existing systems viz., inelasticity to changes in the grid and their aftereffects. In addition made use of lamda service and deployment of AWS web application that will be used by both end-users and external users for viewing data from this system.

REFERENCES

- [1] Use of SCADA Data for Failure Detection in Wind Turbines
K. Kim, G. Parthasarathy, O. Uluyol, and W. Foslien
- Honeywell S. Sheng and P. Fleming National Renewable Energy Laboratory Presented at the 2011 Energy Sustainability Conference and Fuel Cell Conference Washington, D.C. August 7-10, 2011
- [2] Bruhadeswar Bezawada, Kishore k, Raman Dugyala, Rui Li “Symmetric Key based Secure Resource Sharing” Fifth International Symposium on Security in computing and communications (SSCC-17) ISSN:1865:0929 paper was published in Springer-CCIS Journal.2017.
- [3] Research and Application of a SCADA System for a Microgrid Shuangshuang Li, Baochen Jiang *, Xiaoli Wang and Lubei Dong School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209, China; lishuangshuang_sdu@163.com (S.L.); wxl@sdu.edu.cn (X.W.); donglubei@mail.sdu.edu.cn (L.D.) * Correspondence: jbc@sdu.edu.cn; Tel.: +86-186-6935-5366 Academic Editor: Manoj Gupta Received: 8 February 2017; Accepted: 29 March 2017; Published: 31 March 2017
- [4] Shen, Z.Q.; Deng, W.; Pei, W.; Mu, L.S.; Ouyang, H. Design and implementation of microgrid SCADA platform. *Adv. Mater. Res.* 2013, 732–733, 1358–1364. [CrossRef]
- [5] K Vinay Kumar, D Raman, “Secure Overlay Cloud Storage With Access Control And Assure Deletion” ,*International Journal of Engineering & Science Research*, Vol 4, Issue7, @ e-ISSN 2277-2685, p-ISSN 2320-976 July 2014.
- [6] Palma-Behnke, R.; Ortiz, D.; Reyes, L.; Jiménez-Estévez, G.; Garrido, N. A social SCADA approach for a renewable based microgrid—The Huatacondo project. In *Proceedings of the IEEE Power and Energy Society General Meeting*, Detroit, MI, USA, 24–28 July 2011; p. 7.
- [7] Mehta, B.R.; Reddy, Y.J. *Industrial Process Automation Systems: Design and Implementation*; Elsevier Publishers: Amsterdam, The Netherlands, 2015; Volume 7, pp. 237–300.
- [8] D Raman ,BojjaVamshi Krishna “Ensuring Security Services for Data Storing and Data Sharing in Cloud Computing”, *International Journal of Science and Research (IJSR)*, ISSN: 2319-7064, Volume 2 Issue 2, February 2013.
- [9] Pampashree; Ansari, M.F. Design and implementation of SCADA based induction motor control. *Int. J. Eng. Res. Appl.* 2014, 4, 5–18.
- [10] Cao, K. SCADA system research of the grid. *China New Technol. Prod.* 2013, 5, 23. (In Chinese).

- [11] Win, K.T.Z.; Tun, H.M. Design and implementation of SCADA system based power distribution for primary substation (control system). *Int. J. Electron. Comput. Sci. Eng.* 2014, 3, 254–261.
- [12] Chen, Y.N.; Pei, W. Design and implementation of SCADA system for Micro-grid. *Inf. Technol. J.* 2013, 12, 8049–8057. [CrossRef]
- [13] Raman Dugyala, N Hanuman Reddy, N Chandra Shaker reddy, J Phani prasad “A Roadmap to Security in IoT” *International Journal of Applied Engineering Research* ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 8270-8272
- [14] Lázár, E.; Eitz, R.; Petreus, D.; Pătăraș, T.; Ciocan, I. SCADA development for an islanded microgrid. In *Proceedings of the 21st IEEE International Symposium for Design and Technology in Electronic Packaging, Brasov, Romania, 22–25 October 2015*; pp. 147–150.
- [15] Mollah, M.B.; Islam, S.S. Towards IEEE 802.22 based SCADA system for future distributed system. In *Proceedings of the 1st International Conference on Informatics, Electronics and Vision, Dhaka, Bangladesh, 18–19 May 2012*; pp. 1075– 1080.

CERTIFICATELESS DATA INTEGRITY CHECKING AND DATA SHARING WITH SENSITIVE INFORMATION HIDING IN CLOUD STORAGE

K Naveen, M.Tech (CS), JNTU Anantapur, India, naveen00567@gmail.com

C. Shoba Bindu, Professor of CSE, JNTU Anantapur, India, shobabindhu@gmail.com

Abstract— Cloud Storage services allow the users to store their data remotely, where user can easily access the data from anywhere and share the data to others. Once the data is outsourced to the cloud the user shouldn't have control over the data. Due to this availability, integrity and privacy of the data will be in the risk. Hence, to check the integrity of the outsourced cloud data in cloud several Remote data integrity checking techniques have been proposed. In some cloud storages like EHR (Electronic Health Records), the EHR file may consists sensitive data shouldn't revealed to others while sharing the cloudfile. If we encrypt the entire shared file the sensitive information hiding can be achieved but it will make this shared file unable to be used by others such as researchers. Hence, there is a need to address data sharing with sensitive data hiding in remote data integrity checking (RDIC) approach. The data integrity checking schema uses sanitizer to hide the sensitive data while sharing the file into the cloud by generating signatures for the corresponding sensitive data blocks using certificateless cryptography which overcomes the Key Escrow problem. These Schema which ensures secure from the Key replacement and more efficient.

Key Words — Cloud Storage, Data integrity, Sanitizer, Key Escrow.

I. INTRODUCTION

Cloud computing has developed as the distributed computing model, which offers large group of shared resources like memory, storage, systems, applications, and handling capacity to the clients. In view of the demand, they release and maintains the resources and users will pay based on the no. of resources utilized. The engineering of distributed computing includes two stages to be specific Front and Back end.

The Front-end mainly consists of single users, multiple organizations and cloud platforms. Back-end contains of group of different data hubs linked over a network with many software applications, system programs, and different storage platforms. However, there chance of data corruption in cloud or loss in the cause of failures of hardware, human mistakes and unavoidable software bugs in cloud. Hence, there is a need to address remote data integrity checking approach when the data is uploaded to cloud.

Once the data is migrated to cloud, they can be shared by many users in different platforms like iCloud, Drop box, and Google Drive etc. There are many cloud storage like EHR, where the EHR file contains personal data (telephone number, ID number and patient's name, etc.)

it shouldn't be revealed to others while sharing the file. For example, if the EHRs are directly shared among researchers from the cloud there will be a chance of having the personal data of patient and hospital.

In remote data integrity auditing schemes, dataowner has to generate the signatures for the data blocks which has to be uploaded in the cloud. Later those signatures are used to validate the data blocks in cloud, i.e. whether they really contains same data blocks in the integrity checking phase and then the data blocks with their corresponding signatures will be uploaded by the data owner to the cloud.

Our contributions we designed a Certificateless RDIC where data sharing with sensitive data hiding in remote data integrity checking has been provided. In this work, dataowner has to generate the signatures for the data blocks which has to

be uploaded in the cloud. Later those signatures are used to validate the data blocks in cloud. This scheme overcomes the key replacement attack and key escrow problem compared to ID based cryptography. The performance analysis specifies that proposed method is secure and efficient.

II. LITERATURE SURVEY

J. Yu [4] et al., presented mainly on the ID- based signatures. These paper designed an intrusion resilient id-based signature (IR- IBS). These method generates the secrete key using homomorphic design for key updating. These method also specifies the implementation of the IRIBS methods.

J. Yu and H. Wang et al. [5], to maintain security of the key is one of the major problems with the cloud storage services for auditing. These schemas presented a secure key exposure resilient auditing for secure cloud. These method specifies the security & definition model for data integrity auditing.

G. Ateniese et al. [6], designed a model for ID-based RDIC and the security for the model. These model includes the security for both cloud server and the TPA. These model is very secure because it overcomes the leakage of the data in ID- based integrity in the phase of the verification. The results of these model shows that more secure over the generic group.

A. Fu et al. [7] these paper mainly deals about cloud storage services where file are easily manipulate the data and share the data. This method is mainly intended for shared data. In this method consists of t group administrator for recovery of the key and it removes the misuse of single admin of authority and it trust on the cloud storage and finds data changes. If the current data blocks are corrupted then it regenerates the original data blocks.

Y. Zhang et al. [8], these paper presented cloud storage auditing scheme. These scheme is which addresses how to get more efficiency on user revocation

without considering the no. of the total file blocks. These method updates technique for generation of private key. If the sessions of the user is not updated, these method performs correctly for the integrity checking of the revoked user.

III. SYSTEM AND SECURITY MODEL

A. System Model

The model of Certificateless remote data integrity checking (CRDIC) and data sharing in the cloud is specified as in Fig 1. It include five types of different entities

i .Cloud: These entity has huge data storage capacity for the client. In these distributed storage the user data will be uploaded to cloud and it will be shared with others

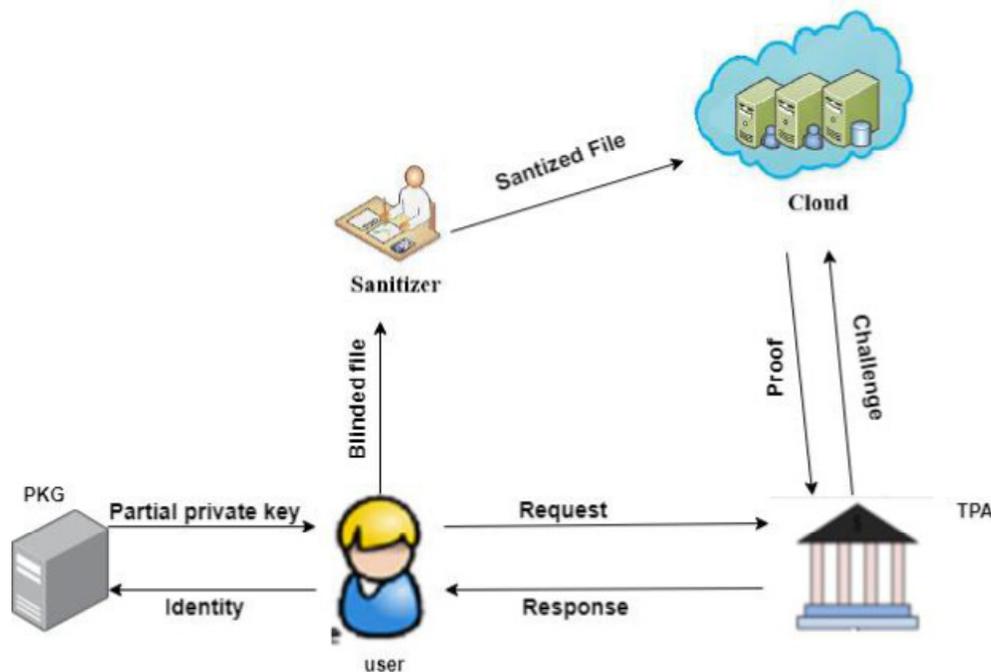


Fig .1 System model

ii. User: The user will be the responsible in the organization, to store the datafile in cloud.

- iii. Sanitizer: The entity which is responsible to perform the sanitization for sensitive data blocks and respective signatures are uploaded to the cloud.
- iv. PKG: These Partial private key Generator is trusted entity. PKG which is used to generate Pparam and partial private key to the clientuser using their UID.
- v. TPA: It is responsible for checking the integrity to the stored data in securecloud storage instead of users.

B. Security Model

We assume that TPA will be honest and secure in Cloud storage systems. Although it is honestly working on the whole integrity

checking process, there may be interest to reveal the sensitive information from the data received in the process. And, if cloud service provider also untrusted. These will show two challenges for security concerns

i.e. data integrity checking in cloud such as privacy over the serviceprovider. We specify the security over the service provider. These model which has two definitions Challenger (\check{C}) and adversary (\check{A}) which specifies user or Third party auditor and cloud serviceprovider respectively.

The proposed Remote data integrity checking with sensitive data hiding is secure over the Cloud Service Provider considered when there is possibility of adversary to win the game is very low. These two definitions are communicated in five phases shown below:

Phase (i) the \check{C} executes the algorithm- $setup(n)$ to get public parameters(Pparam) and the master secret key (mssk), and \check{C} forwards the Pparam to the \check{A} with the mssk as secret.

Phase (ii) The \check{A} recurrently creates some requests like key extract, sign query and hash query to obtain valid tags in the \check{C} . The \check{C} validates tag in each block and sent to the \check{A} .

Phase (iii) The \check{C} computes the challenged message for the specified datablock in the cloud file and sent to the adversary.

Phase (iv) from the challenge, the \check{A} checks the response as the proof for the challenged data blocks which are requested to challenge.

Phase (v) Then, the \check{C} validates proof and the proof is valid the \check{A} will be the winner of the game.

IV. REMOTE DATA INTEGRITY CHECKING WITH SENSITIVE DATA HIDING IN CLOUD

The proposed remote data integrity checking (RDIC) with sensitive data hiding in cloud storage defined as:

It includes the following algorithms: Setup(n), ExtractPartial-Privatekey, SetSecretvalue, Set privatekey, SignGen, Santization, ChallengeGen, ProofGeneration, ProofCheck. These algorithms are initiated in three phases User initialization Phase, Santization Phase and Data Integrity Phase.

A. User Initialization Phase

Setup(n) these algorithm setup which assumes securityparameter as n and gives output as master secrete key and public parameter (Pparam) which is executed by the PKG. These PKG choose three hash functions which are specified as:

$$H_1: \{0,1\}^* \times Z_q^{n \times m} \rightarrow Z_q^m, H_2: \{0,1\}^* \times Z_q^{n \times m} \rightarrow Z_q^m$$

and keeps master secret key as Secrete.

Extract-Partial-Privatekey-(Pparam, mssk ,UID) from the Pparam, the mssk and the user identity (UID), PKG runs the SampleBased (A, TA, s, H1 (UID)) to get response matrix as PUID $\in Z_q^{m \times k}$ and the user will sets his partialprivate key as puid = PUID using the matrix response.

Set-Secrete-value (Pparam, UID) From the input Pparam and the UID, the user casually selects a matrix

$$S_{UID} \in Z_q^{m \times k} \text{ (which must satisfy } \|S_{UID}\| \leq b, \text{ where } b \text{ is a positive integer)}$$

and obtains as secret value suid = SUID. SetPrivate-key (Pparam, Puid, Suid) These algorithm assumes the Pparam and the partial secret key (suid) and the Suid as the input and the user computes his full private key SK = (PUID,SUID).

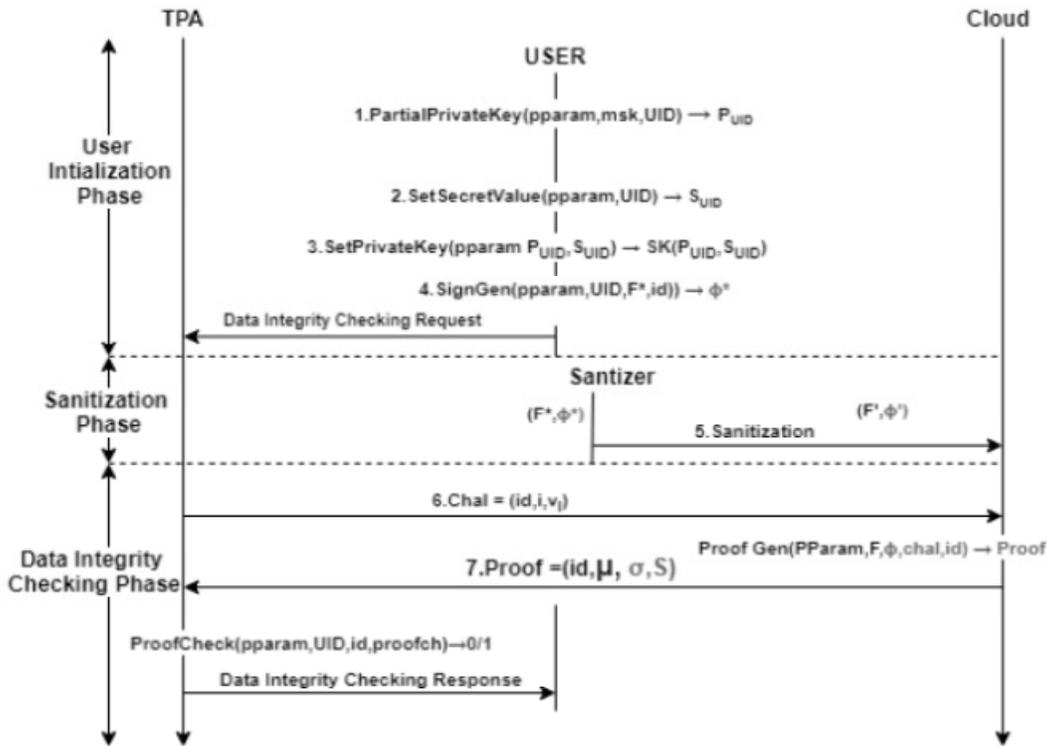


Fig .2 Flow diagram of remote data integrity checking with sensitive data hiding in cloud

SignGen ($F, id, ssk, name$)

(i) The UID randomly picks secret value $P_{UID} \in Z_q$, and evaluates the gr. Then for blinding the UID randomly picks a index K

$\in Z^*$ as these secrete key is used as input for of pseudo random function f . The UID works as the secrete index K to calculate the blinding factor $\alpha_i = fk(i, name)$ ($i \in K$) where blinding factor is to blind the sensitive data blocks which has personal data, where $name \in Z^*_p$ is a random value which is used as the file identifier.

(ii) To hide the personal data which is considered as sensitive data blocks should not revealed to sanitizer, the user is responsible to implement the blinding of sensitive data blocks in the original file F sent before sanitizing. The index of these sensitive fields are in taken as K . The UID calculates the blinding factor for the data block using these $m_i = m_i + \alpha_i$ for every block of the original

file F .

(iii) The user UID creates a transformation value $\beta = u'$ is taken to generate the signature in Sanitization algorithm. User sends $\{F^*,$

$\phi^*, \tau, K\}$ and β to the sanitizer, and removes these messages from the local storage. Furthermore, when the user UID needs to

recover his file F , he can request the

sanitizer. And after that the sanitizer gets the file and forwards the blinded record F^* to user. The original file F can be recovered user using the blinding factor.

B. Sanitization Phase

Sanitization (F^*, ϕ^*)

(i) The sanitizer verifies correctness of the file tag τ_0 by validating $SSign_{ssk}(\tau_0)$ is a legal signature. Then it is correct signature, the sanitizer continues τ_0 where the file identifying name $name$ and validating the gr_{UID} and gr , and forwards to next Stage.

(ii) The sanitizer separately validates the rightness of signature σ_i ($i \in [1, n]$) as follows:

$$e(\sigma_i, g) = e(g_1, g_2) \cdot e(\mu' \prod_{j=1}^l \mu_j^{ID_j}, g^{r_{ID}}) \cdot e(H(name||i) \cdot u^{m_i^*}, g^r). \quad (1)$$

If these above (1) equation is not satisfied, then the sanitizer takes as invalid signature; else it moves to next stage.

(iii) The sanitizer confirms rightness for the value β by verifying the $e(u, gr) = (\beta, g)$ hold or not. If the mentioned equation (2) withholds, then blocks of blinded data will be sanitized by the sanitizer which holds the organization's sensitive data.

$$\sigma'_i = \left\{ \begin{array}{ll} \sigma_i (\beta)^{m_i - m_i^*} & i \in K \\ \sigma_i & i \in [1, n] \text{ and } i \notin K \end{array} \right\} \\ = g_2^x (\mu' \prod_{j=1}^l \mu_j^{ID_j})^{r_{ID}} (H(name||i) \cdot u^{m_i^*})^r \quad (2),$$

(3)

(iv) These sanitizer forwards sanitized file and signatures $\{F', \Phi'\}$ to the cloudserver, and TPA receives the file tag τ_0 from the sanitizer. At the end, messages are removed from the local storage.

C. Data Integrity Phase

ChallengeGen (id, F) the UID sends the checking request to TPA, then validate the integrity of the cloudfile F. Then, TPA selects a subset I of the set [1, l] to be $I =$

{ci}. For every $i \in I$, TPA casually chooses a value and generates the challenged request as $\text{Chall} = \{\text{id}, i, v_i\}$. Then, the challenge is forwarded to the secure cloud from the TPA.

ProofGeneration (Pparam, F', Φ' ,chall, UID) after getting the challenge request from TPA, the cloudserver selects the data blocks and respective signatures of the data blocks. Then, the cloudserver executes the ProofGeneration algorithm to get the proof for the challenge block.

Proof-Check (Pparam, UID, id, Proof, Chall) When TPA gets the proof, then the TPA evaluates $\gamma = H_2(S)$ and checks either the calculation γ holds or not .If γ is valid then confirms the proof else proof is flagged as invalid.

V. PERFORMANCE ANALYSIS

A. Computation cost

To calculate the performance of proposed method, we used .NET Programming language & NTRU-Cryptosystem library & implemented on Green Cloud Open Source Simulator, with the system configuration of

2.30 GHz with 4GB RAM IntelCore i3 processor and Ubuntu 18.04 operating system. We consider file F size 2 Gigabytes is distributed into $l = 1,00,000$ blocks, $|n| = 20$ bits and $|q| = 60$ bits.

Our proposed method which includes only matrix to vector or matrix to matrix multiplications only, while present methods are considered using various pairing and exponential calculations. Our Proposed method, computation cost of SignGeneration algorithm is 1150.00 milliseconds, and Challenged requests takes 800 random blocks. The computed cost taken for the ProofGeneration is 2876.11milliseconds.

To verify the proof, time taken for algorithm is 1265.18 milliseconds. We evaluate our method with Wenting shen et al. Method i.e. shown in Table 1.

Phases	Wenting shen et al.	Our method
Type of Signature	Identity-based signatures	Certificateless signatures
SignGen	2413.67	1150.00
ProofGen	1892.56	800.23
ProofVerify	3420.00	1265.18

Table 1: Comparison of Cost (ms)

We compare our Certificateless Remote Data integrity checking and sensitive information hiding in securecloud with identity based data integrity checking method by Wenting shen et al. using pairing-based cryptography, shown in Fig 3.

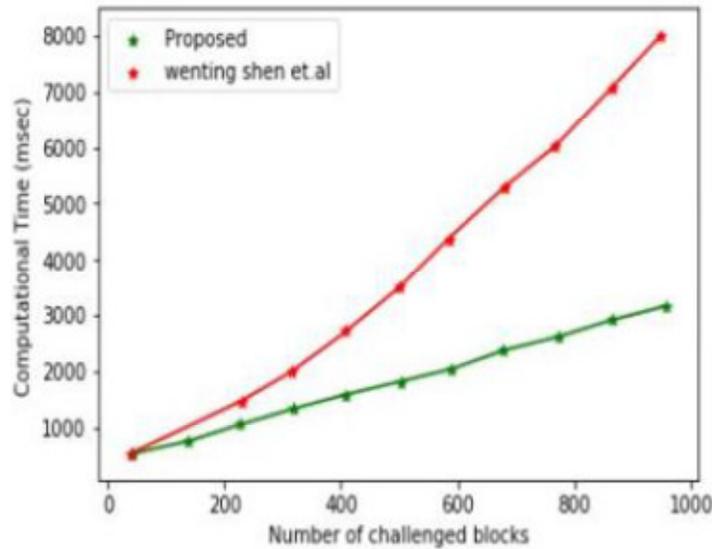


Fig 3: Comparison Of computation Cost Between Wenting Shen et al and proposed method

B. Communication cost

The remote data integrity checking with sensitive information hiding process is started by TPA by forwarding $chall = \{id, i, v_i\}$ whenever $i \in I$ for the secure cloud, then response for

the challenge is sent from the server as c to the TPA. The cost of communication for the generated challenge is taken a $r(|q|)$ bits and cost for message is $(\mu', \sigma', S) \in Z_q^m$.

Performance in different Phases: To calculate the performances in different phases and comparing with wenting shen and our proposed method, consider no. of data blocks as 100 and the no. of sanitized data blocks as 5. As shown in Fig. 4, privatekey generation (PKG) and private key verification (PKV) takes the less time in our proposed method. The time taken for signature generation (SG) and sensitive information sanitization (SIS) respectively are 1.765s and 0.035s. So our method can achieve the signature verification with less time compared to wenting shen et al. method.

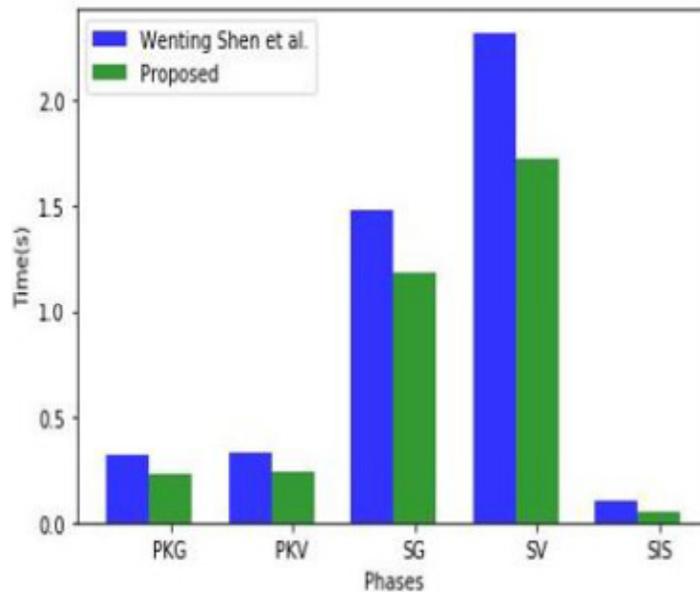


Fig 4: Comparison Of computation Cost Between Wenting Shen et al in different phases

VI. CONCLUSION AND FUTURE WORK

The proposed method guarantees the integrity for outsourced data over untrusted cloud with sensitive information hiding through sanitization technique. It is implemented in three phases i.e. User initialization phase, Sanitization phase and Data integrity checking phase and these are implemented using Lattices. The performance analysis of our proposed method is explained over theoretical analysis, which is then validated by experimental results. Additionally, a relative study has been carried on the proposed and existing methods and the results shows that our method takes less time which makes efficient and effective with cost for the computation with large file size. These approach mainly

concentrates on the validation of proof in the TPA but the data recovery is not focused. Hence there is scope to extend these to provide recovery for the data also.

VII. REFERENCES

- [1] Wenting Shen, Jing Qin, Jia Yu, Rong Hao, and Jiankun Hu, "Enabling Identity-Based Integrity Auditing and Data Sharing with Sensitive Information Hiding for Secure Cloud Storage", 2018.
- [2] C. Sasikala, C. Shoba Bindu, "Certificateless remote data integrity checking using lattices in cloud storage", *Journal of Neural Computing and Applications*, Springer, June 2018.
- [3] Jiguo Li, Hao Yan, and Yichen Zhang, "Certificateless public integrity checking of group shared data on cloud storage," accepted in *IEEE Transactions on Services Computing*.
- [4] J. Yu, R. Hao, H. Xia, H. Zhang, X. Cheng, and F. Kong, "Intrusion-resilient identity-based signatures: Concrete scheme in the standard model and generic construction," 2018.
- [5] J. Yu and H. Wang, "Strong key-exposure resilient auditing for secure cloud storage," 2017. [6] G. Ateniese, X. Huang, W. Susilo, Y. Dai, and G. Min, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage," 2017.
- [7] A. Fu, S. Yu, Y. Zhang, H. Wang, and C. Huang, "Npp: A new privacy-aware public auditing scheme for cloud data sharing with group users," 2017.
- [8] Y. Zhang, J. Yu, R. Hao, C. Wang, and K. Ren, "Enabling efficient user revocation in identity-based cloud storage auditing for shared big data", 2018.
- [9] W. Shen, G. Yang, J. Yu, H. Zhang, F. Kong, and R. Hao, "Remote data possession checking with privacy preserving authenticators for cloud storage," 2017.
- [10] Y. Li, Y. Yu, G. Min, W. Susilo, J. Ni, and K. R. Choo, "Fuzzy identity-based data integrity auditing for reliable cloud storage systems," 2017.
- [11] J. Shen, J. Shen, X. Chen, X. Huang, and W. Susilo, "An efficient public auditing protocol with novel dynamic structure for cloud data," 2017. [12] Haiyang Yu, Yongquan Cai, Shanshan Kong, Zhenhu Ning, Feixue and Han Zhong, "Efficient and Secure Identity-Based Public Auditing for Dynamic Outsourced Data with Proxy." 2017.

Secure and Efficient Access Control For Multi-Authority Cloud Storage

Dr.P.Venkateswara Rao¹, Dr.V.Sucharita², P.Maneepa³

¹Professor,CSE, Narayana Engineering College, Gudur, Nellore Dist, AP.

²Professor, CSE, Narayana Engineering College, Gudur, Nellore Dist, AP

³Scholar, CSE, Narayana Engineering College, Gudur, Nellore Dist, AP.

Abstract: Cloud storage allows both individuals and enterprises to share their data over the Internet in an cost effective manner. This also brings difficult challenges to the access control of shared data since few cloud servers can be fully trusted. Cipher-text policy attribute-based encryption enables the data owners themselves to place fine-grained and cryptographically-enforced access control over outsourced data. In this paper a secure and cost-effective attribute-based data access control for cloud storage systems. In this paper we constructed a Multiauthority CP-ABE scheme that features, the system does not need a fully trusted central authority and all the attribute authorities independently issues the secret keys for users. It also allows the each attribute authority can add or remove any user dynamically from its domain such that those revoked users cannot access subsequently outsourced data. And the cloud servers can update the encrypted data from the current time period to the next one such that the revoked users cannot access those previously available data. The updation of secret keys and cipher text is performed in a public way.

Keywords – Cloud Storage, CP-ABE, Revocation, Data Access Control, Multi Authority.

I.INTRODUCTION

Cloud storage is a computing model in which the data is stored on the remote servers which can access from the internet. It is maintained and operated by a cloud service provider on a storage servers. However the new methodology challenges the traditional data access control scenarios, where a fully trusted server is in charge of access control mechanisms, since the transferred data (e.g: patients health care records)might be sensitive and valuable for data owners, and few servers can be fully trusted. Thus, to protect the security of outsourcing data, data user would like to place access policies over the outsourcing data. There are different solutions in which securing sharing data in the cloud storage systems and attribute-based access control helps to employs the Cipher-text Policy Attribute-Based Encryption (CP-ABE).

This new approach allows each and every user to describe a set of personal attributes and holds a secret key which is distributed by an authority according to his/her Attributes. The data owner places an access policy over those attributes and then encrypts data which is to be supplied out under this policy. Consequently, after

outsourcing the encrypted data to cloud servers, only the outsourced data. This effectively helps to those users whose attributes satisfy the access policy can decrypt prevents unauthorized users from obtaining the outsourced data.

The parameters of these schemes are dependent of the attribute universe which shows that whenever system completes the process of initialization the attributes are completely fixed. Where as in reality, the condition of adding new attributes into system is very common in use.

II. PROPOSED SYSTEM FOR SECURITY MODEL

The system consists of the following entities which plays an important roles.

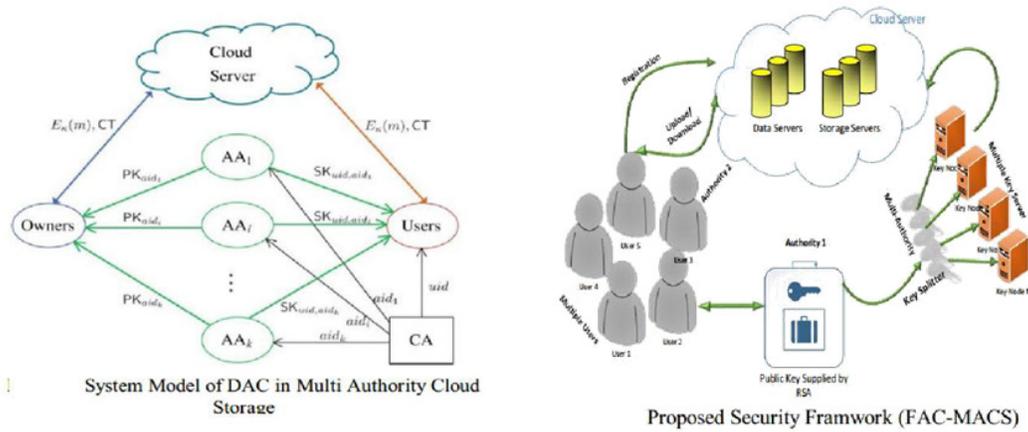
1. Attribute authorities (AAs)
2. Data owners or vendors (owners)
3. Cloud server (server)
4. Data consumers (users)

The Certificate authority is shown as the in-charge of the public parameters which are global. He/she either saves the secret key or generates any for authorities and the users. The Attribute authority, independently manages his/her attribute universe For checking the attribute validity it is used. Followingly, incharge of generating the keys which are used for the updation of user.

Data Owner(DO)is one of the important entity that outsources the user data which is stored in cloud by the service-providers. An access policy is defined for the users based on their attributes and by calling our proposed model multiauthority CP-ABE. Then the data owner sends the encrypted data to cloud servers. Each user owns a unique global-identifier in the system, and possesses a set of attributes and the corresponding secret key which consists of all secret-key components which are issued by different Attribute Authorities (AAs). If the particular user is not revoked by authority then he/she can use the publically defined update key in order to update the corresponding secret key component.

The Semitrusted server is used to store and perform updation on the data which is encrypted by data owners. The updation process can be done by using the parameter which is public and without the involvement usage of secret information/data.

The multiauthority CP-ABE model is developed to enhance the efficiency of previous schemes. It also fixes the update key is same to all non-revoked users. Hence anon-revoked data user can share the key with revoked user.



In this theme the following challenges can be seen:

1. Compromising of authorities which is (N-2) do not allow the shutdown of system.
2. User's data should be protected in a secure way by the single-authority.
3. A feasibility is shown here with a security analysis and better performances.

III. LITERATURE SURVEY

Author proposed a revocation of attributes with the multi authority CP-ABE model. It also proved to improve the system efficiency[1].

Author proposed a revocable multi-authority CPABE scheme that could support efficient attribute revocation and constructed an effective data access control scheme for multi-authority cloud storage systems. Author also proved that this scheme was provable secure in the random oracle model. The revocable multi-authority CPABE is trustworthy technique, which can be applied in any remote storage systems and online social networks [2].

Author proposed the requirement that the new parameters should be defined publically by the authority. In addition to that the owner or the assigned authority should be able to generate the new key for the data which is encrypted. Thus it is only needed to limiting the secret key usage to use for only decryption process[3].

Author proposed a new authority scheme for data access which is only of single-authority. To overcome the revocation of keys problem author constructed the CPABE which supports the single authority and also address the revocation demerit. Their methodology is built on product of primes of three with a lower efficiency of implementation [4].

Author proposes that the attacks of the user un-encrypted data. This scheme is practical, and all parameters grow at most logarithmically with the total number of time periods. An efficient scheme such as the random oracle can extend to achieve the attacks against the chosen text and also can unbounded the number of time periods. The new scheme such as encryption of binary tree can be shown[5].

Author proposed schemes achieved fine-grained privilege control and identity anonymity while conducting privilege control depends on user's identity. Also conducted security and performance analysis which shows that Anony Control both secure and efficient for cloud storage system. More important is, this system can tolerate up to $N - 2$ authority compromise, which is mostly prefer specially in Internet-based cloud computing environment. This inherits the security from the Anony Control which is secure, but overhead communication is incurred during the 1-out-of-n oblivious transfer [6].

Authors designed a secure data sharing scheme Mona for dynamic groups in an untrusted cloud. In Mona, users are able to share data with others in the group without revealing identity privacy to the cloud. Also, Mona is efficient in user revocation and new user joining. More specially, efficient user revocation can be achieved by public revocation list without updating the private keys of the other remaining users, and new users can directly decrypt files stored in the cloud without their participation. Moreover, the storage overhead and the encryption computation cost are constant. By analysis it is proved that proposed scheme was satisfy the security requirements and efficiency [7].

Author proposes a new threshold multi-authority CP-ABE access control scheme TMACS, in public cloud storage, in which all AAs jointly manage the whole attribute set and share the master key α . Taking advantage of (t, n) threshold secret sharing, by interacting with any t AAs, a legal user can generate his/her secret key. Thus, TMACS avoids any one AA being a single-point bottleneck on both security and performance. The analysis results show that author's access control scheme is robust and secure. It can easily find appropriate values of (t, n) to make TMACS secure when less than t authorities are compromised, also robust when no less than t authorities are alive in the system. Further, based on efficiently combining the traditional multi-authority scheme with TMACS, construct a hybrid scheme that is more suitable for the real scenario. This scheme addresses attributes coming from different authorities, security and system-level robustness [8].

Author found that, if a revoked user wants to access the unauthorized content whose access policy can be satisfied by his/her revoked attributes. Here it analyzes the shortcoming of DAC-MACS in dealing with attribute revocation. Author's proposed attack algorithm to transform the new-version ciphertext to the old-version one if he/she can collude with the cloud service provider to get enough ciphertext update keys. The security vulnerability exists because DAC-MACS wrongly use a bidirectional re-encryption scheme in the ciphertext updating procedure. This vulnerability allows any party to re-encrypt the ciphertext between old-version and new-version, only if he/she can get the CUKs between these two versions [9]. Various data access control scheme and Attribute based encryption methods are compared as shown in table 1.

Table 1: Comparison between various data access control scheme with Attribute-Based Encryption

Dataaccesscontrol techniques	Advantages	Disadvantages
Threshold multi-authority ciphertext-policy(CP)ABE accesscontrol sceme(TMACS)	1) It satisfies the scenario of attributes from different AAs 2) It can achieve security and system-level robustness.	Reusing of the master key shared among multiple authorities (AAs).
Comments and corrections of CP-ABE	Analyze the shortcoming of DAC-MACS in dealing with attribute revocation, main construction proved it Secure	Security vulnerability

DST Sponsored National Conference on Recent Advancements on
Computer Science (CONRACS 2019)- 26 to 28 July 2019

<p>Privilege control scheme3</p> <p>AnonyControl</p> <p>AnonyControl-F</p>	<p>1) Able to protect user's privacy against single authority.</p> <p>2) Tolerant against authority</p>	<p>1)Data confidentiality</p> <p>2)Person information defined by each user's attributes set is at risk</p> <p>3) Resilient in security breach.</p>
<p>Attribute revocable multi-authority CP-ABE scheme</p>	<p>1) It incurs less communication cost and computation cost, and is secure</p> <p>2) It can achieve both backward and forward security</p>	<p>Lack of efficiency</p>
<p>Secure multi-owner data sharing scheme MONA</p>	<p>1. Reduced the computation overhead to encrypt files and cipher text size.</p> <p>2. The ciphertext size is constant and independent of revocation users.</p>	<p>1)User Compute revocation parameters To protect the Confidentiality</p> <p>2) computation overhead of the encryption</p>

IV. RESULT

Attributes of different systems are managed by the multiple different authorities. Different access-policies are implemented by multiple-authorities with different attributes. In Cloud Storage systems for accessing of data is controlled by this methodology.

This project work is implemented using the coding language java and java-script .Thus the implementation of project outcome with results are shown in the following fig[1],[2],[3] and [4] .



Figure-1



Figure-2



Figure-3



Figure-4

V.CONCLUSION

A cost-effective model with a multi-authority attribute-based access control scheme for sharing the data in cloud storage systems. With a scalable no of User revocations and publically updating of the cipher-text. Also the defined security properties such as the forward security and backward security, can also with stand the decryption of key exposure. The security of the random oracle model is proven.This model can be aptable for online networks and many other applications.

VI. FUTURE WORK

One of the promising future works is to introduce the efficient user revocation mechanism on top of proposed anonymous ABE. Supporting user revocation is an important issue in the real application, and this is one of the greatest challenges in the applications of ABE schemes. Making this scheme compatible with existing ABE schemes, support efficient user revocation.

VII. REFERENCES

- [1] K. Yang, X. Jia, K. Ren, B. Zhang, and R,Xie,“DAC-MACS: Effectivedataaccesscontrolformultiauthoritycloudstoragesystems,”*IEEETrans.Inf. Forensics Security*, vol. 8, no. 11, pp. 1790–1801, Nov. 2013.
- [2] Kan Yang and Xiaohua Jia, “Expressive, Efficient, and Revocable Data Access Control for Multi-Authority Cloud Storage”, *IEEETransactions on parallel and distributed systems*, VOL. 25, NO. 07, July 2014

[3] Q. Li, J. Ma, R. Li, X. Liu, J. Xiong, and D. Chen, "Secure, efficient and revocable multi-authority access control system in cloud storage," *Comput. Security*, vol. 59, pp. 45–59, 2016.

[4] A. Sahai, H. Seyalioglu, and B. Waters, "Dynamic credentials and ciphertext delegation for attribute-based encryption," in *Proc. Adv. Cryptol.— CRYPTO 2012*. New York, NY, USA: Springer, 2012, pp. 199–217.

[5] R. Canetti, S. Halevi, and J. Katz, "A forward-secure public-key encryption scheme," *J. Cryptol.*, vol. 20, no. 3, pp. 265–294, 2007.

[6] Taeho Jung, Xiang-Yang Li, Zhiguo Wan, and Meng Wan, "Control Cloud Data Access Privilege and Anonymity with Fully Anonymous Attribute-Based Encryption", *IEEE transactions on information forensics and security*, VOL. 10, NO. 01, January 2015.

[7] Hideaki Ishii, Roberto Tempo, and Er-Wei Bai, "Mona: Secure Multi-Owner Data Sharing for Dynamic Groups in the Cloud", *IEEE Transactions on parallel and distributed systems*, VOL. 24, NO. 06, June 2013.

[8] Wei Li, Kaiping Xue, Yingjie Xue, and Jianan Hong, "TMACS: A Robust and Verifiable Threshold Multi-Authority Access Control System in Public Cloud Storage", *IEEE Transactions on parallel and distributed systems*, VOL. 24, NO. 06, October 2015.

[9] Jianan Hong, Kaiping Xue and Wei Li, "Comments on "DAC-MACS: Effective Data Access Control for Multi-authority Cloud Storage Systems"/Security Analysis of Attribute Revocation in Multi-authority Data Access Control for Cloud Storage Systems", *IEEE transactions on information forensics and security*, VOL. 10, NO. 06, June 2015.

Review on Implementation of Artificial Intelligence for Optimal Task Scheduling in Cloud Environments

Ramakrishna Goddu^{1} and Kiran Kumar Reddi²*

¹Research scholar, Department of computer science, Krishna University, Machilipatanam, Andhra Pradesh, India

²Assistant professor, Department of computer science, Krishna University, Machilipatanam, Andhra Pradesh, India

*Email: ramakrishnagoddu1@gmail.com

Abstract: Task scheduling in cloud environments deals with the allocation of resources to execute the task in cloud environments. Task scheduling is a np-hard problem as has number of solutions during the resource allocation. To solve this problem researchers developed several kinds of algorithms. But no single techniques is globally accepted because of the desirability of the cloud environments. This paper deals with a detailed review analysis on implementation of artificial intelligence techniques to obtain the resource allocation plans to execute the tasks.

Keywords: virtual machines, cloud computing, task scheduling, artificial intelligence and optimization

1. Introduction

Cloud computing provides sharing of computing resources and data storage, and allows its users to access information to utilize its services over the internet and central remote servers on demand [1]. It allows users to pay amount based on their usage only. This feature of cloud computing attracts so many small scale industries and other upcoming companies. The resources are shared between cloud clients through virtualization technique. Virtualization divides single computational resource into multiple independent execution environments, also known as virtual machines[2]. Once the partition is done, each logical server behaves like physical server that runs the operating system automatically, this is the main flexibility in virtualization concept. In cloud computing, the resource allocation to tasks is the main objective of management system, because the virtual machines efficiency reflects the performance and cost of the system. An inefficient virtual machine allocated a resource will decrease performance and increases cost of the system [3]. The primary objective is to obtain optimal output and to solve large scale computation problems by using resource infrastructure of resource allocation[8]. The cloud provider can guarantee the QOS(quality of service) to its users with maximum resource utilization and minimum power consumption. QOS denotes the degree of satisfaction of a customer for a service depending on the level of performance, reliability, and availability [9]. To implement QOS in cloud computing, task scheduling algorithm is used [10]. Assigning jobs to particular resources at particular time is known as scheduling and scheduling of tasks plays major role in cloud computing. Improper scheduling may lead to reduction of system performance. The main objective of task scheduling is to increase performance and to decrease the task completion time by developing an efficient scheduling algorithm. Based on the priority, the task is allocated to the virtual machine and then it is mapped to suitable physical machine.

The virtual machine is selected according to the task necessities like CPU power and cost, after that the task scheduler allocates task to the selected virtual machine in which the task is to be executed successfully[3].

1.1 Why optimization of task scheduling

Task scheduling optimization in a distributed heterogeneous computing environment is a NP-hard problem which plays a major role in optimizing cloud utilization and QOS. As cloud task scheduling is an NP-hard problem we should increase the efficient use of the shared resources to achieve optimal task scheduling. To attain optimal task scheduling, so many meta-heuristic algorithms have been developed, such as max-min, min-min, PSO, GA, ACO, AIS, simulative annealing etc.

2. Task scheduling using Artificial neural networks (ANN)

Artificial neural networks also known as connectionist systems are computing systems that are developed based on the concept of biological neural networks, but not exactly similar. Dr. Robert Hecht-Nielsen, the inventor of the first neurocomputer stated a neural network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

ANNs are made up of several nodes as shown in Fig.1, which reflect the behaviour of biological neurons of human brain. To interact with each other the neurons are connected in the form of links. The nodes take input data and perform simple operations based on the data present in it. The output of these operations is forwarded to other neurons. The output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The output obtained at each node is known as node value. Each link is assigned with some weight. The weight increases or decreases the strength of the signal at a connection.

Artificial neurons can suffer a threshold due to which the signal is sent only if the aggregate signal crosses that threshold. Artificial neurons are a collection of layers. Different layers perform different types of operations on their inputs. Signals travel from the starting layer i.e. input layer, to the ending layer i.e. output layer, probably after crossing the layers several number of times.

The main objective of the ANN was to find solutions to the problems in the same way how human can do. Now a days Artificial neural networks have been used to perform multiple tasks such as, speech recognition, social network filtering, machine translation, playing board and video games, computer vision and medical diagnosis.

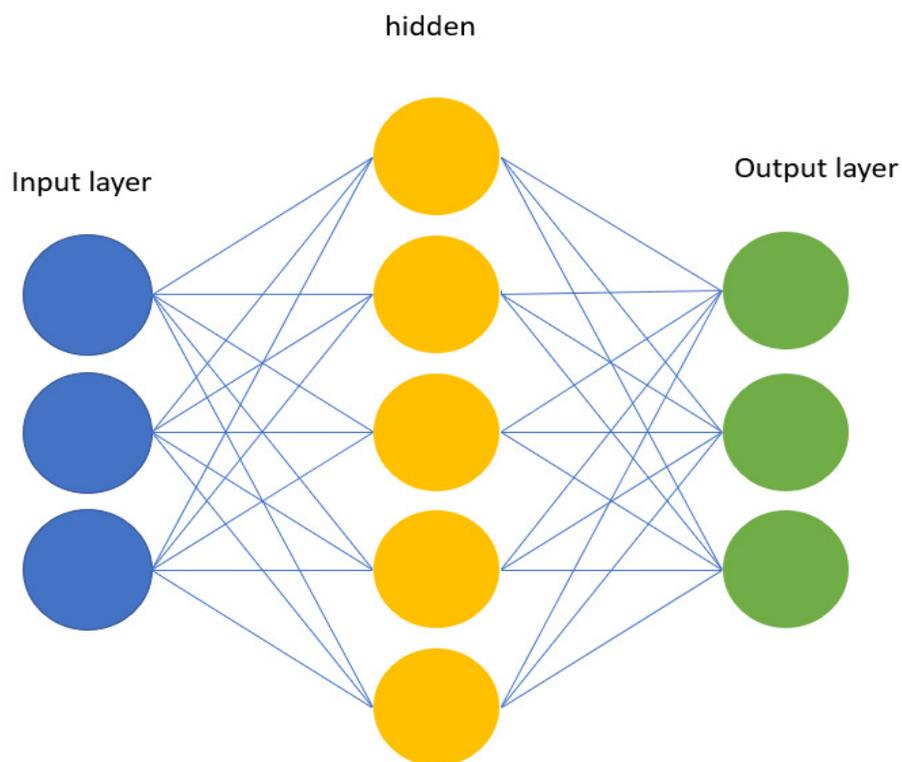


Fig. 1 ANN architecture

Kumar and Dinesh [11] introduced the concept of job scheduling using fuzzy neural network algorithm. The objective is to optimize the QOS parameters such as bandwidth, processing time and memory. Based on these parameters the tasks are classified and given to the fuzzier where input value are converted to 0's and 1,s. These converted input values are sent to the neural network. The neural network then will map the resources with the tasks. And then de-fuzzification is done to convert fuzzy values into original values. Results shown that the turnaround time is decreased and system performance was increased.

3. Task scheduling using Min-Min

In min-min algorithm the task with minimum execution time (i.e, min) is selected among all the tasks. That task is allocated to the resource which has minimum completion time (min). The process continues until all the tasks are executed. As this algorithm completes small tasks first there will be a greater delay of larger tasks.

Chen et al [12] developed user-priority guided min-min scheduling algorithm for load balancing in cloud computing. In this paper to decrease the makespan and increase the resource utilization the priority of the tasks is decided by the user and then min-min algorithm is applied which improves the load balance on resources.

Liu et al [13] proposed an improved min-min algorithm in cloud computing. In addition to min-min algorithm they had considered three constraints such as quality of service, dynamic priority model, and cost as main objectives to satisfy the user needs and increase resource utilization by giving priority to the tasks with minimum completion time.

Maipan and Mishra [14] proposed an extended min-min scheduling algorithm in cloud computing. Unlike min-min which gives priority to the tasks with minimum completion time, the enhanced min-min allocates tasks based on the difference between max-min execution time of tasks. Results shown that the proposed algorithm works better than the existing min-min and improved min-min algorithms in terms of makespan.

4. Task scheduling using max-min

To overcome drawback of the min-min algorithm max-min was introduced. In max-min algorithm the largest task (max) i.e, the task with maximum execution time is sent first to the resource with minimum completion time (min) for execution. But in this algorithm some times the makespan may increase because larger tasks are executed first as well as the waiting time of smaller tasks is also high.

Bhoi and Ramanui [15] introduced enhanced max-min task scheduling algorithm in cloud computing. They had presented the modified work of improved max-min algorithm. Where in improved max-min the selection is done based on expected execution time rather than completion time, the enhanced max-min also follows the same criteria but the difference is that the enhanced max-min allocates task with average execution time to resource with minimum completion time whereas improved max-min allocates largest task to slowest resource. Results shown that overall makespan and load balance among resources is reduced.

Elzeki et al.[16] proposed improved max-min algorithm in cloud computing which is a modified form of max-min algorithm. The algorithm was developed based on RASA algorithm and basic max-min concept. In this proposed algorithm the selection process is based on expected execution time rather than completion time. Experimental results shown that proposed algorithm best performs than RASA and basic max-min.

Mao et al.[17] developed max-min task scheduling algorithm for load balancing in cloud computing. They proposed a novel approach for load balance in elastic cloud by using max-min task scheduling algorithm. In this algorithm a task status table is maintained to store the load of virtual machines and their expected completion of tasks which allocates workload between nodes and load balance is known.

5. Task scheduling using PSO:

Particle Swarm Optimization (PSO) is an optimization method which is inspired by the functionalities of bird schooling or fish schooling. PSO is more advantageous and easier in its

implementation as compared to genetic algorithm. In PSO, a swarm of birds are considered as population and each bird is treated as an individual. Each individual has its own position, velocity and fitness values while moving towards the destination position. The fitness value of the particle/individual is to be evaluated in order to update its position and velocity. Thereby, particles of the population move in the search space after knowing the global best position ever achieved earlier.

Let us consider swarm of birds are flocking for the food and there are 'n' of birds are also known as individuals are in the swarm as represented with

$$\{Set_{swarm}\}$$

Each individual has its own position and velocity as represented with

$$\begin{aligned} &\{Set_{position}\} \\ &= \{X_1, X_2, \dots, X_n\} \\ &\{Set_{velocity}\} \\ &= \{V_1, V_2, \dots, V_n\} \end{aligned}$$

During the fly, each individual imagine to move to its next best position. It means, for a swarm with 'n' number set will have the same number on position best values. Among all the position values, the swarm decides the global best position and every individual tries to approach the global best position as represented in Figure.1.

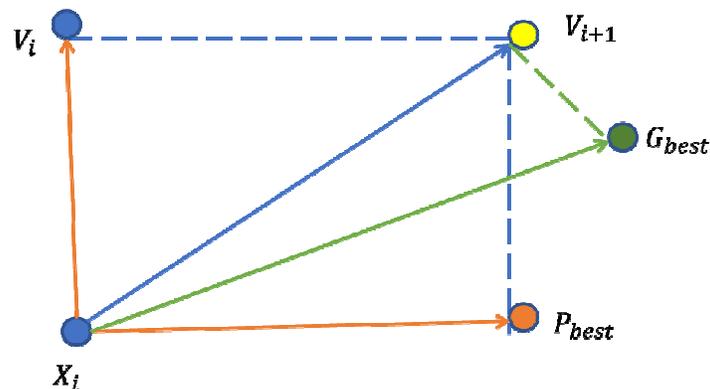


Figure.1 Particle movement representation in PSO algorithm.

If X_i & V_i represent the position and velocity of particle in i^{th} iteration, the particle will update its position and velocity i.e. X_{i+1} & V_{i+1} after finding the G_{best} as represented in equations (1) & (2).

$$V_{i+1} = V_i + C_1 * R_1 * (P_{best} - X_i) + C_2 * R_2 * (G_{best} - X_i) \quad (1)$$

$$X_{i+1} = X_i + V_{i+1} \quad (2)$$

P_{best} = Position best of the current particle

G_{best}

= Global best of the swarm which is not achieved earlier

C_1 & C_2
= Cognitive parameters

R_1 & R_2

= Random parameters

X_i & V_i

= Position and velocity of individual particle in i^{th} iteration

X_{i+1} & V_{i+1} = Updated Position and velocity of individual particle in $(i + 1)^{th}$ iteration

The pseudo code of the basic PSO algorithm is represented in Figure.2

```

Algorithm-1: PSO Pseudo Code

Step-1: Start PSO
Step-2: Initialize (Population, Positions & Velocities)
Step-3: Obtain      for each individual
Step-4: While (end criteria) {
Step-5:     Evaluate fitness of each individual
Step-6:     Obtain      of swarm
Step-7:     Update Positions & Velocities of each individuals as
              per equations (1) & (2)
              } end while
Step-8: End PSO
    
```

Figure 2. PSO pseudo code

Particle swarm optimization is a social inspired evolutionary algorithm, which has been implemented for solving various engineering problems [4-6]. Past studies considered particle swarm optimization (PSO) based task scheduling is more efficient for dynamic task scheduling. Juana et al.[8] proposed a PSO based task scheduling algorithm to overcome the problems of cloud computing. They developed cost vector model which measures scheduling schemes cost and solution was developed based on input task

and QOS parameters. This method is proven to be effective but it is with more complexity. Krishnasamy[18] proposed a hybrid particle swarm optimization task scheduling algorithm which decreases the average operation time and increases the usage of resources and provides the resources according to the user tasks. Alkayal et al.[3] developed an PSO based multi object task scheduling by introducing a new approach of ranking. Here, the tasks are allocated to virtual machines according to the rank, which decreases the waiting time and increase system performance. To find a solution to the problems of cloud computing in task scheduling, Dordaie & Navimipour [19] proposed a hybrid particle swarm optimization and hill climbing algorithm which was properly scheduled, but it takes more time for task completion.

A.I.Awad et al. [20] introduced a new mathematical model using load balancing mutation a particle swarm optimization(LBMPSO) for task scheduling to achieve reliability and availability which are cumbersome parameters in cloud computing. This LBMPSO model maintains load balancing while distributing tasks to the resources available and the failure tasks are rescheduled. This feature made the model reliable.RK Jena [21] proposed task scheduling algorithm by the means of multi objective nested particle swarm optimization(MOPSO) which mainly focused on reducing power consumption and processing time. Results shown that when compared to BRS and RSA the MOPSO is efficient with multi objectives.Entisar S. Alkayal et al. [22] developed a model for optimized task scheduling based on particle swarm optimization with multi objective optimization for resource allocation. In this they introduced a new concept of ranking strategy , based on the results obtained from this ranking strategy the scheduling is done that finds the best virtual machine which suits the task. Results shown that this algorithm improves throughput and decreases latency.

6. GTS-Grouped Task Scheduling

Hendgamal EI Din Hassan Ali [23] introduced grouped task scheduling algorithm(GTS) algorithm for scheduling tasks into services which are QOS driven. This algorithm combines tasks with similar attributes that forms a category. These categories helps to know priority of tasks and then these tasks are scheduled to available resources. Tasks with high value of attributes are firstly scheduled and secondly tasks with less execution time is scheduled first i.e., with low latency. GTS performs optimal scheduling when compared to min-min and TS algorithms.Xiaonianwu et al.[24] proposed a QOS driven based task scheduling algorithm in cloud computing. This algorithm schedules the tasks based on their priority and the time taken to complete the task, i.e, it allocates the resource which takes less time to execute the given task. This work obtains load balancing and optimal performance

7. Task scheduling using GA (Genetic algorithm)

Genetic algorithm (GA) is a metaheuristic stimulated by the process of natural selection that was inspired by evolution theory . Genetic algorithms are vastly used to find optimal solutions to both constrained and unconstrained optimization and search problems based on bio-inspired operators such as mutation, crossover and selection. This algorithm reflects the process of natural selection where the fittest individuals(parents) are selected for reproduction in order to generate a offspring(children) of the next generation.

Initially the fittest individuals from a population are selected. They generate offspring which receive the characteristics of the parents and will be supplementary to the next generation. Depending on the fitness of the parents, parents having better fitness, produce better offspring which is improved than parents and have a better possibility at living. This process will be iterated until, either a maximum number of generations are developed, or a optimal fitness level has been reached for the population.

Five phases are considered in a genetic algorithm.

1.Initializing Population: In GA set of individuals is called as Population. Each individual is a solution to the problem you want to solve. An individual is a set of parameters called as Genes. Genes are combined as a string to generate a chromosome. Usually, binary values are used (string of 1s and 0s) to represent genes of a individual.

2.Fitness: The fitness function calculates how a individual is capable to contest with other individuals. The fitness value of each individual is calculated. The probability of selecting a individual for reproduction is based on its fitness value.

3.Selection: Tin the selection process the fittest individuals are selected and allows to pass their genes to the next generation. Based on their fitness values two pairs of individuals (parents) are selected. Individuals with high fitness have more chance to be selected for reproduction.

4.Crossover: In crossover for each pair of parents to be reproduced, a crossover point from within the genes is selected randomly. Offsprings are generated by exchanging the genes of parents among themselves until they reach the cross over point as shown in Figure 3. The obtained new offspring are added to the existing population.

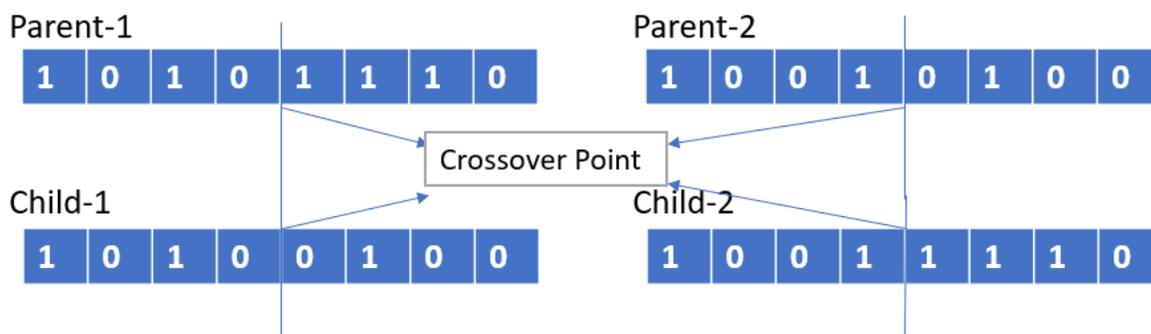


Fig.3 Representation of cross over genetic operator

5. *Mutation*: The new offspring formed, are subjected to mutation with a low random probability. This indicates that some of the bits in the bit string can be reversed. Mutation maintains diversity within the population to prevent premature convergence (Figure 4).

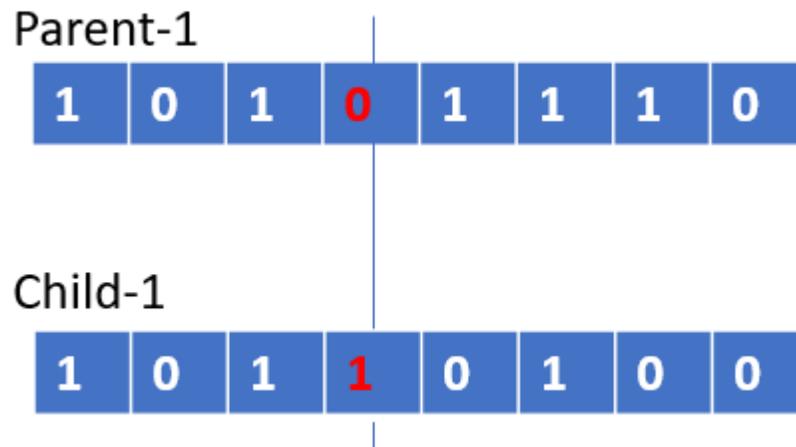


Fig.4 Representation of Mutation genetic operator

The flow process of genetic algorithm is represented in Fig.5.

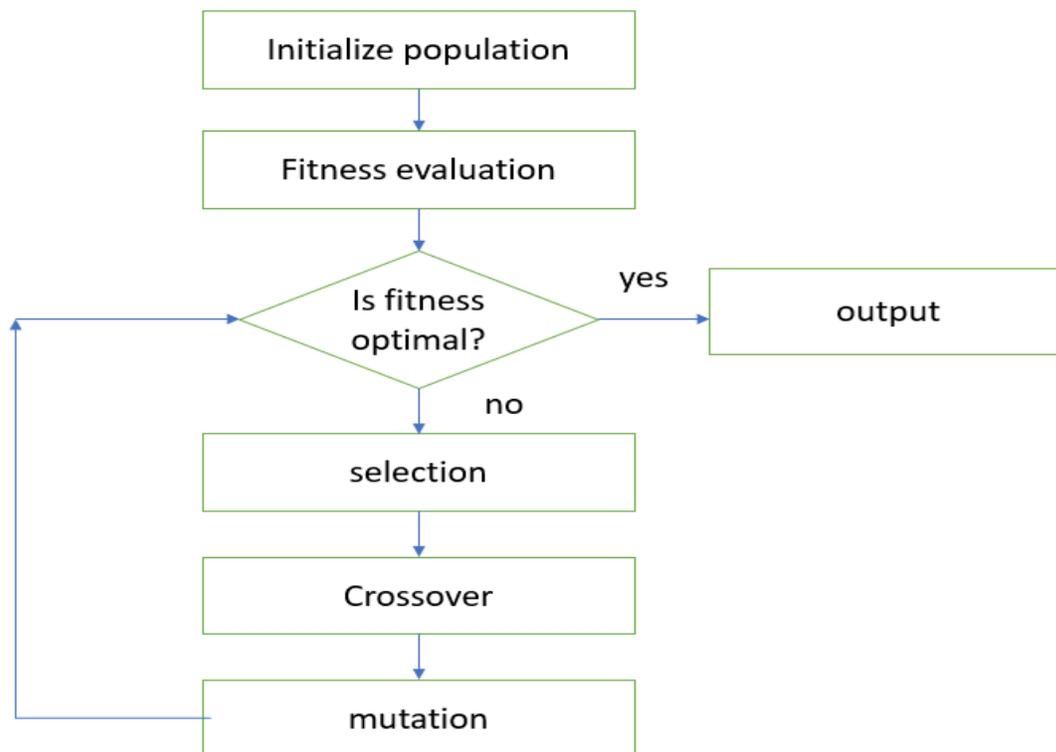


Fig.5 Flow process of GA

Yuji , guiyiwei [25] proposed genetic algorithm based techniques to solve task scheduling problems in cloud computing. In this research, the entire group of tasks in the job que are examined and then scheduling decision is taken by reducing the make span in order to achieve better load balancing. Sung ho jang, tac young kim et al.[26] introduced task scheduling technique with the implementation of genetic algorithm in cloud environments. The algorithm focused on user benefits such as QOS and profits to cloud providers. Whereas, the previous developed models concentrated on either or combination of minimum execution time or work load. This work schedules the with preference of user satisfaction by evaluating the GA based fitness function. This process is iterated until an optimal task schedule achieved. They compared the results with previous works and found it performed well. Shaminderkaur, Amandeep verma[27] developed a modified genetic algorithm by addressing a fitness to achieve minimum completion time in regard to a single user only. Experiment results showed that the developed algorithm achieves good performance under heavy loads. Chun-yanliu, et al.[28] described a task scheduling algorithm with the help of hybridised genetic ant colony systems. The authors focused on finding a solution to slow convergence problem caused by initial pheromone deficiency of ACO and then GA is integrated to find the optimal solution.

8. Task scheduling using ACO (Ant Colony Optimization):

ACO is a metaheuristic and probabilistic method that is used to find optimal solutions to different optimization problems. ACO was developed by observing the real world behaviour of ants.

In general the ants start searching for food, in that trail they move randomly in a path and once the food is found they return back to their colony by leaving the pheromone trail. When other ants come across the pheromone trails, there is a possibility to follow the same path. If they do so, they then leave their own pheromone while bringing their food back. As time goes on, the pheromone trail starts to evaporate, by reducing its intensity. A short path, by comparison, has more intensity which more ants travelled frequently, and thus the pheromone density becomes higher on shorter paths than longer ones. when more ants select the same path, the pheromone intensity increases and it gets stronger. Since the ants leave pheromones every time they bring back food, shortest path becomes high pheromone concentrated defining the optimal solution. Meanwhile, some ants are still randomly search for closer food sources. A similar approach can be used to find near-optimal solution to the traveling salesman problem.

The dynamic behaviour of the ant colony algorithm made it best suit to works such as graphs with changing topologies.

Medhat Tawfek et al.[29] proposed an ant based task scheduling in cloud environments. The researchers concentrated on minimizing the make span by implementing ACO. Finally the

obtained results are compared with previous algorithms and found efficient. Qiangguo [30] addressed a algorithm to optimize task scheduling problem in cloud computing based on ant colony algorithm. The algorithm initializes the pheromone and is updated along with the developed heuristic function. They performed experiment by comparing results with Min-Min and found the algorithm obtained better make span, less cost and load balance. Liyunzuo, et al.[31] introduced a multi-objective optimization scheduling method based on ant colony system. The research used the performance and budget constraint functions to examine the past to give the feedback on the quality. This feature prevents the ACO from local optimal solution and then from the feedback, quality of the solution is obtained.

9. Task scheduling using AIS (Artificial Immune System):

Artificial immune system (AIS) was inspired by the natural immune systems which adopts many properties of natural immune systems such as diversity, distributed computing, fault tolerance, dynamic learning, and self-monitoring. The AIS represents learning from the situation, once an agent from that environment is come across, an immune response is set into motion as a cascade of events that attempts to recognize the agent as “friend” or “enemy.” This recognition is remembered for future encounters, such that friendly agents are ignored and foes result in a reaction that in extreme cases, such as anaphylaxis, result in serious physical harm. The artificial immune system is designed along these same principles, and exhibits similar qualities to the in vivo system. These systems have several characteristics that make them particularly useful for data mining. First, they learn from interaction with an environment, and that learning occurs under reinforcement such that learning is rewarded or punished, depending on the response to environmental stimuli. Second, these systems have memory, in that over time they remember environmental inputs from prior experience, so there is an element of recognition. Third, they are adaptive systems, such that response to environmental stimuli can change over time. Finally, they are diverse, in that there are multiple and distributed receptors in the immune pathway. These features translate nearly directly to a computational model of the immune system.

Waanengshu, Weiwang And Yunji [32] wang Proposed a novel energy efficient resource allocation algorithm based on immune clonal optimization for green cloud computing. The authors considered the cost and energy consumption as parameters to minimize. In immune algorithms performs the clonal selection process to obtain the optimal solution. Yuanjunlaili, Linzhang et al.[33] introduced an energy adaptive immune – genetic algorithm for task scheduling algorithm in cloud manufacturing systems. The developed technique improve the searching diversity which depends on immune strategy and capable of existing crossover and

mutation probability. This algorithm maintains the balancing between search diversification and intensification. R.k.jena[34] addressed an energy efficient algorithm for task scheduling algorithm in cloud computing by using clonal selection algorithm. By taking into consideration, optimization of energy consumption and make span. The clonal operator is an antibody and its affinity is tested to send the new antibodies which have high affinity. The proposed method performs better than scheduling algorithms and random task plans.

10. Task scheduling using Fuzzy Systems:

The word fuzzy is defined as the things that are not clear or ambiguous. In general many times we across the situations that we cannot decide the solution is true or false. As a solution to this problem fuzzy systems provides a very precise and flexible solution. In this way, we can consider the inaccuracies and uncertainties of any situation.

In general Boolean system considers truth value as 1 and 0 as false value. In contrast, fuzzy systems doesn't consider absolute truth or false values such as 1 and 0 . Fuzzy logic, takes values in between 1 and 0 which might be partially true or partially false.

Fuzzification is the process of converting inputs i.e. crisp numbers into possible fuzzy sets. Crisp inputs are the inputs that are given to sensors and sent for processing, such as temperature, pressure, rpm's, etc.

Fuzzy control technique represents how human thinks and implement it into a control system. fuzzy control provides acceptable reasoning than accurate reasoning. uncertainties can be handled easily with the help of fuzzy logic. This system can work with any type of inputs whether it is vague, slanted or noisy input information.

The structure of Fuzzy Logic Systems is easy and understandable. Fuzzy logic comes with mathematical concepts of set theory and reasoning. It provides a very efficient solution to complex problems in all fields of life as it resembles human reasoning and decision making but fuzzy logic comes with a drawback that it works with precise and imprecise data so most of the time accuracy is negotiated.

Kong et al.[35] described a dynamic task scheduling algorithm with fuzzy prediction in virtualized data centres. Virtualization comes with some provocations such as task scheduling and resource management. This algorithm considers availability and responsiveness to solve task scheduling. A Fuzzy based prediction tool is developed to represent uncertain work load and availability of vogue of virtualized server nodes by using typ-1 and type-2 fuzzy logic systems.

Javanmardi, et al. [36] proposed a novel approach in which they modified the existing genetic algorithm with the help of fuzzy theory. This concept helped in reducing the iteration of producing population. Two chromosomes with different QOS constraints are taken and for

those fitness value is obtained by using fuzzy theory. They had shown that by using this approach about fifty percentage of cost and time can be reduced.

Shojafar et al. [37] proposed a hybrid approach which is a combination of fuzzy theory and genetic algorithm(FUGE). The traditional genetic algorithm(GA) is modified and fuzzy theory is applied to develop a steady state in GA to increase the makespan. In FUGE the jobs according to their lengths are assigned to the suitable resources based on the virtual machine speed, memory and bandwidth. Results shown this approach best performs than other cloud scheduling algorithms in terms of time and cost of execution and average degree of imbalance.

Mehranzadeh and Hashemi [38] developed A Novel-Scheduling Algorithm for Cloud Computing based on Fuzzy Logic controller. The controller takes into consideration the number of requests received by host and Tasks precedence value to find the optimal results.

Sethi and Jena [39] described Efficient load Balancing in Cloud Computing using Fuzzy Logic. In this proposed approach the speed of the processor and load on the virtual machine are considered and fuzzy logic was applied to balance the load.

11. Task scheduling using Hybrid moth search

Mohamed Abd Elaziz et al.[40] proposed a paper on task scheduling based on hybrid moth search algorithm using differential equation (MSDE) to minimize the processing time and increase the throughput. The moth search algorithm(MSA) was developed by taking into consideration the behaviour of moths to move towards light in nature for global optimization, the DE algorithm is used for local search to improve the capability of MSA.

12. Conclusion

An elaborative study has been presented on implementation of artificial intelligence techniques for solving task scheduling presented in this review article. From the study, it is found that most of the works dealt with the genetic algorithm, particle swarm optimization and ant colony optimization. However, a few research works have been devoted to implementation of hybrid techniques for task scheduling.

From the review analysis, it is observed that a suitable algorithm is to be implemented which has less number of tuning parameters so as to obtain the solution without the mathematical complexity.

References

- [1] Tilak, S., & Patil, D. (2012). A survey of various scheduling algorithms in cloud environment. *International Journal of Engineering Inventions*, 1(2), 36-39.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

- [2] Buyya, R., Vecchiola, C., & Selvi, S. T. (2013). *Mastering cloud computing: foundations and applications programming*. Newnes.
- [3] Alkayal, E. S., Jennings, N. R., & Abulkhair, M. F. (2016, November). Efficient task scheduling multi-objective particle swarm optimization in cloud computing. In *2016 IEEE 41st Conference on Local Computer Networks Workshops (LCN Workshops)* (pp. 17-24). IEEE.
- [4] Kassarwani, N., Ohri, J., & Singh, A. (2019). Performance analysis of dynamic voltage restorer using improved PSO technique. *International Journal of Electronics*, 106(2), 212-236.
- [5] Arumugam, P., Panchapakesan, M., Balraj, S., & Subramanian, R. C. (2019). Reverse Search Strategy Based Optimization Technique to Economic Dispatch Problems with Multiple Fuels. *Journal of Electrical Engineering & Technology*, 1-7.
- [6] Ghosh, P., Karmakar, A., Sharma, J., & Phadikar, S. (2019). CS-PSO based Intrusion Detection System in Cloud Environment. In *Emerging Technologies in Data Mining and Information Security* (pp. 261-269). Springer, Singapore.
- [7] Vinothina, V., Sridaran, R., & Ganapathi, P. (2012). A survey on resource allocation strategies in cloud computing. *International Journal of Advanced Computer Science and Applications*, 3(6), 97-104.
- [8] Juan, W., Fei, L., & Aidong, C. (2012). An Improved PSO based Task Scheduling Algorithm for Cloud Storage System. *Advances in Information Sciences and Service Sciences*, 4(18), 465-471.
- [9] Ardagna, D., Casale, G., Ciavotta, M., Pérez, J. F., & Wang, W. (2014). Quality-of-service in cloud computing: modeling techniques and their applications. *Journal of Internet Services and Applications*, 5(1), 11.
- [10] Wu, X., Deng, M., Zhang, R., Zeng, B., & Zhou, S. (2013). A task scheduling algorithm based on QoS-driven in cloud computing. *Procedia Computer Science*, 17, 1162-1169.
- [11] Kumar, V. V., & Dinesh, K. (2012). Job scheduling using fuzzy neural network algorithm in cloud environment. *Bonfring International Journal of Man Machine Interface*, 2(1), 01-06.
- [12] Chen, H., Wang, F., Helian, N., & Akanmu, G. (2013, February). User-priority guided Min-Min scheduling algorithm for load balancing in cloud computing. In *2013 National Conference on Parallel computing technologies (PARCOMPTECH)* (pp. 1-8). IEEE.
- [13] Liu, G., Li, J., & Xu, J. (2013). An improved min-min algorithm in cloud computing. In *Proceedings of the 2012 International Conference of Modern Computer Science and Applications* (pp. 47-52). Springer, Berlin, Heidelberg.
- [14] Maipan-uku, J. Y., Mishra, A., Abdulganiyu, A., & Abdulkadir, A. (2018). An Extended Min-Min Scheduling Algorithm in Cloud Computing. *i-manager's Journal on Cloud Computing*, 5(2), 20.
- [15] Bhoi, U., & Ramanuj, P. N. (2013). Enhanced max-min task scheduling algorithm in cloud computing. *International Journal of Application or Innovation in Engineering and Management (IJAIEEM)*, 2(4), 259-264.
- [16] Elzeki, O. M., Reshad, M. Z., & Elsoud, M. A. (2012). Improved max-min algorithm in cloud computing. *International Journal of Computer Applications*, 50(12).
- [17] Mao, Y., Chen, X., & Li, X. (2014). Max-min task scheduling algorithm for load balance in cloud computing. In *Proceedings of International Conference on Computer Science and Information Technology* (pp. 457-465). Springer, New Delhi.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- [18] Krishnasamy, K. (2013). Task Scheduling Algorithm Based On Hybrid Particle Swarm Optimization In Cloud Computing Environment. *Journal of Theoretical & Applied Information Technology*, 55(1).
- [19] Dordaie, N., &Navimipour, N. J. (2017). A hybrid particle swarm optimization and hill climbing algorithm for task scheduling in the cloud environments. *ICT Express*.
- [20] Awad, A. I., El-Hefnawy, N. A., &Abdel_kader, H. M. (2015). Enhanced particle swarm optimization for task scheduling in cloud computing environments. *Procedia Computer Science*, 65, 920-929.
- [21] Jena, R. K. (2015). Multi objective task scheduling in cloud environment using nested PSO framework. *Procedia Computer Science*, 57, 1219-1227.
- [22] Alkayal, E. S., Jennings, N. R., &Abulkhair, M. F. (2016, November). Efficient task scheduling multi-objective particle swarm optimization in cloud computing. In *2016 IEEE 41st Conference on Local Computer Networks Workshops (LCN Workshops)* (pp. 17-24). IEEE.
- [23] Ali, H. G. E. D. H., Saroit, I. A., &Kotb, A. M. (2017). Grouped tasks scheduling algorithm based on QoS in cloud computing network. *Egyptian informatics journal*, 18(1), 11-19.
- [24] Wu, X., Deng, M., Zhang, R., Zeng, B., & Zhou, S. (2013). A task scheduling algorithm based on QoS-driven in cloud computing. *Procedia Computer Science*, 17, 1162-1169.
- [25] Ge, Y., & Wei, G. (2010, October). GA-based task scheduler for the cloud computing systems. In *2010 International Conference on Web Information Systems and Mining* (Vol. 2, pp. 181-186). IEEE.
- [26] Jang, S. H., Kim, T. Y., Kim, J. K., & Lee, J. S. (2012). The study of genetic algorithm-based task scheduling for cloud computing. *International Journal of Control and Automation*, 5(4), 157-162.
- [27] Kaur, S., & Verma, A. (2012). An efficient approach to genetic algorithm for task scheduling in cloud computing environment. *International Journal of Information Technology and Computer Science (IJITCS)*, 4(10), 74.
- [28] Liu, C. Y., Zou, C. M., & Wu, P. (2014, November). A task scheduling algorithm based on genetic algorithm and ant colony optimization in cloud computing. In *2014 13th International Symposium on Distributed Computing and Applications to Business, Engineering and Science* (pp. 68-72). IEEE.
- [29] Janmohammadi, P., &Babazade, M. Resource Management in the Cloud Computing Using a Method Based on Ant Colony Optimization.
- [30] Guo, Q. (2017, April). Task scheduling based on ant colony optimization in cloud environment. In *AIP Conference Proceedings* (Vol. 1834, No. 1, p. 040039). AIP Publishing.
- [31] Zuo, L., Shu, L., Dong, S., Zhu, C., & Hara, T. (2015). A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing. *Ieee Access*, 3, 2687-2699.
- [32] Shu, W., Wang, W., & Wang, Y. (2014). A novel energy-efficient resource allocation algorithm based on immune clonal optimization for green cloud computing. *EURASIP Journal on Wireless Communications and Networking*, 2014(1), 64.
- [33] Laili, Y., Zhang, L., & Tao, F. (2011, December). Energy adaptive immune genetic algorithm for collaborative design task scheduling in cloud manufacturing system. In *2011 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 1912-1916). IEEE.
- [34] Jena, R. K. (2017). Energy efficient task scheduling in cloud environment. *Energy Procedia*, 141, 222-227.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

- [35] Kong, X., Lin, C., Jiang, Y., Yan, W., & Chu, X. (2011). Efficient dynamic task scheduling in virtualized data centers with fuzzy prediction. *Journal of network and Computer Applications*, 34(4), 1068-1077.
- [36] Javanmardi, S., Shojafar, M., Amendola, D., Cordeschi, N., Liu, H., & Abraham, A. (2014). Hybrid job scheduling algorithm for cloud computing environment. In *Proceedings of the fifth international conference on innovations in bio-inspired computing and applications IBICA 2014* (pp. 43-52). Springer, Cham.
- [37] Shojafar, M., Javanmardi, S., Abolfazli, S., & Cordeschi, N. (2015). FUGE: A joint meta-heuristic approach to cloud job scheduling algorithm using fuzzy theory and a genetic method. *Cluster Computing*, 18(2), 829-844.
- [38] Mehranzadeh, A., & Hashemi, S. M. (2013). A novel-scheduling algorithm for cloud computing based on fuzzy logic. *International Journal of Applied Information Systems (IJ AIS)*, 5(7).
- [39] Sethi, S., Sahu, A., & Jena, S. K. (2012). Efficient load balancing in cloud computing using fuzzy logic. *IOSR Journal of Engineering*, 2(7), 65-71.
- [40] Elaziz, M. A., Xiong, S., Jayasena, K. P. N., & Li, L. (2019). Task scheduling in cloud computing based on hybrid moth search algorithm and differential evolution. *Knowledge-Based Systems*, 169, 39-52.

Clustering EDP (Error Detection Program) Errors from Cloud Data Centers using Data Mining

¹Karunala Arunraj Bapuji, ²B. Anand Kumar, ³Vorsu Mallaiah, Prof.A Vinayababu
¹ 3 Research Scholar, Department of Computer Science, Acharya Nagarjuna University, Guntur.
² Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool,
⁴Director (IT), Stanley College of Engineering and technology, Hyderabad.
1 E-Mail : arunraj.karunala@gmail.com

Abstract: The dynamic nature of client requirements and sharing of resources in cloud computing lead to unexpected errors in the cloud and may lead to unavailability of service and server. To resolve these issues, we have developed a new mechanism called Error Detection Program (EDP) to resolve all the issues by using proactive and reactive algorithms in fault tolerance of IaaS of cloud service model. In The previous paper, the errors are limited to only text or excel file and data mining techniques are not applied on error files. This paper is proposed with application of data mining techniques on error files to store, prioritize and studying the properties of the errors using k-means classification algorithm

Keywords: EDP Classification, virtual machines, Data Mining in Cloud Computing, SLA, IaaS, K-Means in Cloud errors.

1. INTRODUCTION

Data mining is the process used to store, classify, cluster and dig the huge volume of data hails from various sources such as data bases, text files, social media, internet and shopping malls. Cloud computing is cutting edge technology which enables sharing of resource using the internet to the end use at various service levels based on the requirements of clients.

As per National Institute of Standards and Technology (NIST) cloud computing reference architecture, the cloud architecture contains five characteristics, four deployment models and three service models [2][3] discussed in flowing.

1.1 Five characteristics of cloud architecture are

- i. **On demand self-services:** Cloud resources like processor, RAM, power and storage to end user using could services without any human interaction.
- ii. **Broad network:** In cloud computing all the service and resources can be access using internet.
- iii. **Resources pooling:** In cloud computing same physical resources can be shred to the multiple users.
- iv. **Rapid elasticity:** End user can change resources as per requirement.

Measured Services: A Cloud is internet-based service, where service provider will charge as per customer usage. Service will be charged and measured based on processor time, storage space, internet bandwidth etc.

1.2 Types cloud (Deployed Models)

Public cloud: In public cloud all the shared resources will be provided to the end user through internet. In this public cloud same hardware resources are shared to multiple users across network. Service provider apply charges same resources according to the usage. Public cloud is most useful for startup and small companies because no need to install hardware, software and which provide immediate access to the server. On public cloud has less scope control to the end user on hardware infrastructure, application and services. [1]

Private cloud: Private cloud is maintained by internal IT team or service provider and it is best choice where business critical and data security is needed. Private cloud resides always behind the firewall and only authorized persons can access it. Private cloud provides full control in security, customized application, and performance measured issues. [1]

Hybrid cloud: Hybrid cloud is the grouping public and private cloud. In hybrid cloud all the precarious actions are performed in private cloud and non- precarious activities are performed in public cloud. In hybrid cloud maximum utilization of resources and services. It can be maintained in house or service provider and End-user has complete freedom to customize resources as per requirement. [1]

Community: In this cloud same resources are shared between different originations. Community cloud can be presented and maintained by one of the community IT team or third-party service provider. Budgets are shared across the origination, so those involved get some of the benefits of a private cloud at a lower price. [1]

1.3 Cloud service models

Infrastructure-as-a-Service [IaaS]: In IaaS fundamental resources like physical hardware, virtual machine, networking, and storage etc. provided to end-user. In this service model pre-install hardware and operating system provided to the end user. End user can install own apps in server and service provider grant access to the user to perform common operation. Service provider also supports the previous infrastructure of the customer. [5]

Platform-as-a-Service [PaaS]: PaaS provide improvement tools and runtime environment applications. PaaS will provide up-to-date services to end user for their application development. Service provider will test all the development tools before proving it to end-user, so that end-user will get benefit which will save their money and valuable time [5]. PaaS is good for startup or development companies.

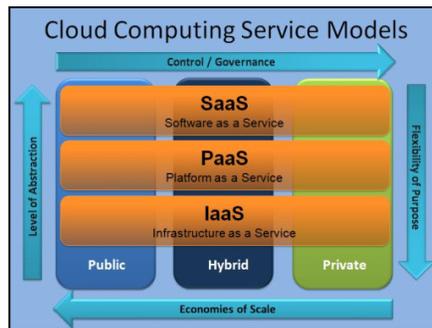


Figure 1.0 Cloud deployed and service model [9]

Software-as-a-Service [SaaS]: In SaaS application provided to the end-user as a service. Most commonly used service is office 365, salesforce and email application google apps etc. customers over the internet based on the type of subscription. In SaaS customer may not be know what kind operating-system and hardware used in the to run all these services. [5]

1.4 Data mining

Data mining is process of extracting useful information from raw data. In cloud computing most of data will be unstructured. Data mining can help in cloud computing extracting unstructured data to structured data. In cloud compute all the software's storage servers and network maintained are centralized. Data mining will help to maintain these entire centralized infrastructures and secure reliable all services for users of the cloud.

Data mining is highly efficient to collect structured data from unstructured data for a service or task. And data can be from various sources or platforms. Data mining uses two kinds of models 1. Descriptive 2. Predictive. [13] Descriptive model generally used to characterize the general properties of the data in database [1]. Predictive model if performing inference on current data to make prediction [13] [14]. Both models can archive different variety of data from raw data. Analyzing and extracting useful data in

various fields where human interaction is involved. This helps to all cloud customers to get their valuable information for just click on one button.

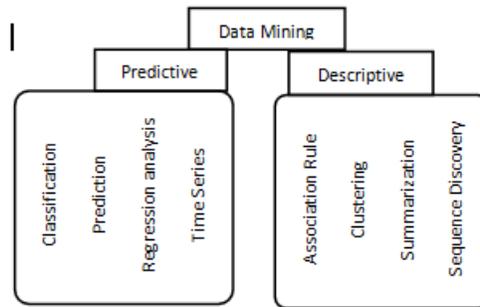


Fig.1.1. Data mining model

2 Literature Survey

In the previous paper K. Arunraj Bapuji¹, Prof K. Mruthyunjaya Rao, “data mining in cloud computing a proactive error detection program in cloud computing to identify the errors”. [13] We have focused on generating error arriving from cloud data centers at IaaS service level.

B. Anand Kumar, A.Vinayababu, V. Malliah, K. Madhukar Worked on “Generating Fault Tolerance Mechanism in Cloud Computing System using EDP”. The proposed paper concentrated on the EDP algorithm proactively recognizes faults, when reboot or update is required. To resolve the faults, EDP uses the rejuvenation mechanism of proactive fault tolerance system. If errors still exist, it uses task resubmission mechanism of reactive fault tolerance system [15].

In the recent years Mr Ghamdan Mohammed Qasem, Dr. Madhu B. K worked on proposal "Proactive Fault Tolerance in Cloud Data Centers for Performance Efficiency". They proposed Fuzzy min mix natural network classification approach when a VM (Virtual Machine) failure happens by monitoring the fail progress status of VM. Proposed method improves the FMNN and there by lowers computational complexity and increase performance. They implemented a simulation setup and the proposed method of algorithm will move VM to another server in advance when failures happen, and it predicts failures. They concluded that they use Cloudsim software for practical

possibility of prediction of VM failures. The proposed proposal shown better results to predict VM's from failures. [4]

In recent year Mr. Renu Asnani worked on "A distributed k-mean clustering algorithm for cloud data mining". The proposed paper is survey of different cloud data storage, and their cluster analysis for utilizing the data into various business intelligence applications. And author proposed new model of cluster analysis of data which provides the clustering as service. [11]

Research scholars in 2017 Archana, Mohan Kumar, worked on "Study on Fault Tolerance and Resilience in Cloud Computing Using Predictive Approach". The proposed paper concluded to use mirror and detecting faults using linkup approach in cloud. First faults or errors will be detected and resolved by best technique by reactive method (pre-emptive migration method). And this makes the system tolerate from the faults. The faults also can be proactively handled which helps to resolve errors associative with faults. [6]

Mr. Sachin Shinde and Bharat Tidke worked on "Improved K-means Algorithm for Searching Research Papers". They projected a paper which uses the search engine based on clustering and text mining. The improved architecture with improved k-means algorithm, which recommends a method for making the algorithm more operative and efficient, to get better clustering with reduced difficulty. It will search the base keyword of the content from the knowledge database. They conclude that showing correct results and better initial centroids and provide a resourceful way of transfer the data points to the suitable clusters, associated to original k-means algorithm. [12]

Research scholars in 2017 another article was proposed, Deepak Kochhar, Abhishek Kumar Jabanjalin Hilda, worked on "An Approach for Fault Tolerance in Cloud Computing Using Machine Learning Technique". They proposed using Naïve Bayes application classifier in proactive FT. Naïve Bayes classifier is used to categorize VM's into ones which are prone to errors and the ones which are not and apply proactive FT techniques. By using Naïve Bayes classifier with FT techniques and reliability is archived. The proposed model they use proactive FT method with Naïve Bayes classifier and they experienced issue with first node and applying FT systems to ensure enhanced accurate of the system and it can enhance the correct of the system, values

used to estimation the failures of component in node. Consistency of system is measured using MTBF factor. With Naïve Bayes application results, 85% node failures could be reduced. [7]

In recent years Mr. Iqjot Singh, Perna Dwivedi, Taru Gupta and P. G. Shynu, worked on "Enhanced K-means clustering with encryption on cloud" they projected this paper tries to solve the problematic of storing and managing big files over cloud by put on hashing on Hadoop in big-data and guarantee security while uploading and downloading files. [8]

3.0 PROPOSED MODEL

The proposed model contains two steps 1) EDP extract errors from cloud data centres 2) Clustering the EDP errors based on properties of errors using k-means enhanced algorithm.

3.1 EDP Errors from Various Cloud Data Centres

Cloud computing is one of the most complexes and challenging technology, where we have number of servers and services are up and running 24/7. All the servers and services are connected virtually on a same physical machine. Different types of network, operating system, and hardware errors occurred in real time, and this error leads to fault in the server and leads to the SLA (service level agreement). There are different types of fault layers are hardware, operating system, network, storage and application. Faults can happen either hardware or operating system level which makes the server or service unavailable.

3.2 K-means

K-Means algorithm groups in to 'n' objects k clusters, the object in one cluster having approximately equal mean with the other object in the same cluster, the thing on the cluster are calculated as follows. This proposed model implements the Bisecting k-means algorithms which is an efficient algorithm to cluster the EDP error database. This algorithm presented in the following implementation section.

```

*****
duplicate packages :
-----
*****

=====
conflict packages :
-----
Unsatisfied dependencies for libibnetdisc5-xxxxxxx.x86_64:
  libad.so.5()(64bit) is needed by (installed) libibnetdiscX-xxxxxxx.x86_64
  libad.so.5(AD 1.3)(64bit) is needed by (installed) libibnetdisc5-xxxxxxxxxxx.x86_64
  libad.so.3()(64bit) is needed by (installed) libibnetdisc5-xxxxxxxxxxx.x86_64
  libibumad.so.3(IBMAD 1.0)(64bit) is needed by (installed) libibnetdisc5-xxxxxxxxxxx.x86_64
  libosmcomp.so.3()(64bit) is needed by (installed) libibnetdisc5-xxxxxxxxxxx.x86_64
  libosmcomp.so.3(OSMCOMP 2.3)(64bit) is needed by (installed) libibnetdisc5-xxxxxxxxxxx.x86_64
Unsatisfied dependencies for kernel-syms-9999999999999999.x86_64:
  pesign-obs-integration is needed by (installed) kernel-syms-9999999999.x86_64
*****
rpm DB issue :
-----

/run/lvm/lvmetad.socket: connect failed: No such file or directory
WARNING: Failed to connect to lvmetad. Falling back to internal scanning.
/run/lvm/lvmetad.socket: connect failed: No such file or directory
WARNING: Failed to connect to lvmetad. Falling back to internal scanning.
/run/lvm/lvmetad.socket: connect failed: No such file or directory
WARNING: Failed to connect to lvmetad. Falling back to internal scanning.

```

Fig.3.1. Format of the real time cloud environment EDP generated errors

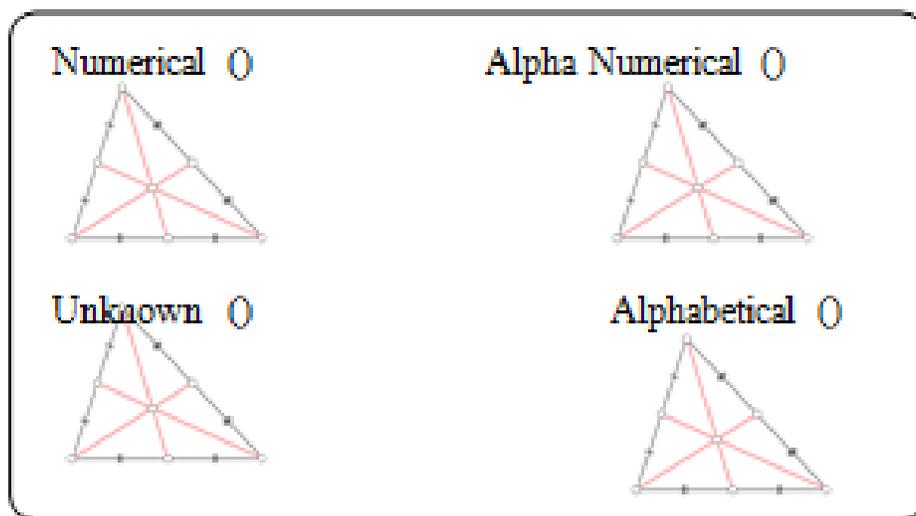


Fig.3.2 Categorization of EDP errors

Tables 1.1 Sample error data base

Sl No.	Error Code	Error Description
1	Error 137	Patching related error
2	Error 407	Patching related error
3	Error P1	Patching related error
4	Error unknown	Kernel related error

4.0 Implementation.

The EDP part of the cloud computing generate the errors in to the text or excel format. And then using the python scripting the above error format is imported to MySQL database. The snap short of the error data base presented in the above table 1.1. The next step is to implement the bisecting k-means algorithm to cluster error data points.

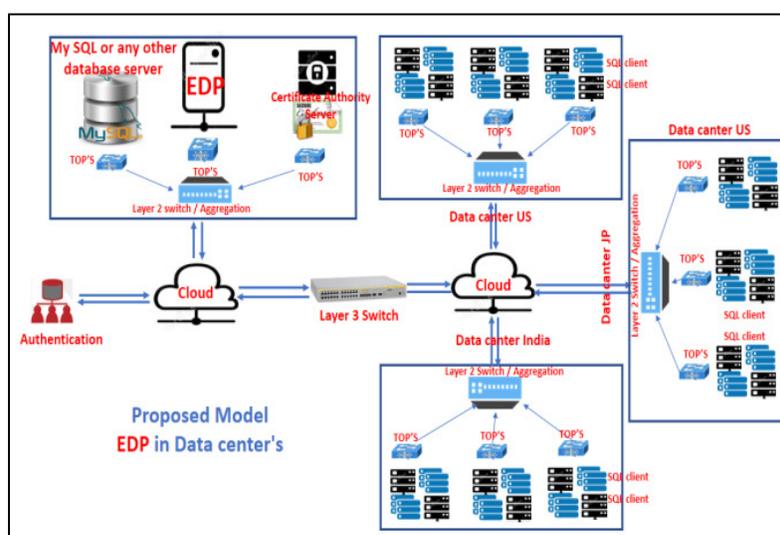


Fig.3.3. Proposed EDP program

Proposed Bisecting K-means Algorithm

1. Convert the error database in to MySQL database
2. Initialize the list of clusters to contain the cluster consisting of all error data points.
3. Repeat.
4. Remove a cluster from the list of clusters.
5. {Perform several “trail” by session of the chosen cluster.}
6. For $i=1$ to number of trails do Bisect the selected cluster using basic k-mean
7. End for
8. Select the two clusters from the bisection with the lowest total SSE.
9. Add these two clusters to the list of clusters.
10. Until the list of clusters contains k Clusters

5. Conclusion

In the real-world fault tolerance technique are very familiar. In this paper we proposed a solution for all UNIX flavors proactively identifying the errors and storing them in database servers and clustering using bisecting K-Means for further analysis. However, in future we will extend this work using digging techniques to make the ranking of errors and prediction.

6. References:

- [1] <https://www.telappliant.com/blog/cloud-service-types-and-deployment-models/>
- [2] Anju Bala, indeer chnna, “Fault tolerance challenges, techniques and implementation in cloud computing”. IJCSI international Journals of computer science issue, vol. 9, Issue. 1, January 2012.
- [3] Youssef M. Essa Bigdata Consultant, EMC, Cairo, Egypt, "A Survey of Cloud Computing Fault Tolerance: Techniques and Implementation" IJCA (0975 – 8887) Volume 138 – No.13, March 2016
- [4] Ghamdan Mohammed Qasem, Dr. Madhu B. K "PROACTIVE FAULT TOLERANCE IN CLOUD DATA CENTERS FOR PERFORMANCE EFFICIENCY" ijipam.eu, ISSN: 1311-8080 (printed version); ISSN: 1314-3395.
- [5] <https://www.ibm.com/cloud>
- [6] Archana, Mohan Kumar, “Study on Fault Tolerance and Resilience in Cloud Computing Using Predictive Approach”, IJIACS ISSN 2347 – 8616 Volume 6, Issue 10 October 2017.

- [7] Deepak Kochhar, Abhishek Kumar Jabanjalin Hilda, "AN APPROACH FOR FAULT TOLERANCE IN CLOUD COMPUTING USING MACHINE LEARNING TECHNIQUE", IJPAM, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)., Volume 117 No. 22 2017, 345-351.
- [8] Iqjot Singh, Perna Dwivedi, Taru Gupta and P. G. Shynu, "Enhanced K-means clustering with encryption on cloud" IOP Conf. Series: Materials Science and Engineering 263 (2017) 042057 doi:10.1088/1757-899X/263/4/042057.
- [9] <http://cloudcomputingnet.com/cloud-computing-models/>
- [10] B. Anand Kumar¹, A.Vinayababu, V. Malliah³, K. Madhukar⁴, IJRASET, ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue XII, Dec 2018
- [11] Renu Asnani, "A distributed k-mean clustering algorithm for cloud data mining", (IJETT) – Volume 30 Number 7 - December 2015.
- [12] Sachin Shinde, Bharat Tidke, "Improved K-means Algorithm for Searching Research Papers " IJCS&CN, Vol 4(6),197-202 ISSN:2249-5789
- [13] Nikit Jain, Vishal srinivas "Data Mining techniques: a survey paper" ijret : international journal of research in engineering and technology, volume : 02 issue : 11 NOV – 2013
- [14] Dr. M.H Dunham, "Data Mining, Introductory and Advanced Topics", Prentic Hall, 2002.
- [15] B. Anand Kumar¹, K. Madhukar², V. Malliah³, Vinay Babu" Generating Fault Tolerance Mechanism in Cloud Computing System using EDP - PART 2" (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887.

Providing Authentication for Big Data in Cloud Computing

**M Vijaya laxmi, Assistant Professor, CSE Department, AAR Mahaveer Engineering College,
Bandlaguda, Telangana .**

**M. Sri laxmi, Assistant Professor , CSE Department, AAR Mahaveer Engineering College,
Bandlaguda, Telangana .**

Abstract:

Most organizations agree that data should be at the heart of decision-making. To transform data into insight, a big data storage infrastructure that gives meaning to unstructured and dark data, and can perform when the window of action is milliseconds. In this paper we are providing high security through authentication. In existing system organizations are using cloud services with security but within the organization data breach occurs via hacking, a disgruntled employee, or careless username/password security, in business data can be compromised. And more data breaches are happening to more supposedly secure companies all the time. Cloud service will support third party authentication. The third party will be trusted by both the cloud service and accessing user. Third party authentication will add an additional security layer to the cloud service.

1.Introduction

Data management optimizes data performance and costs through capacity planning, storage utilization and data placement. All by moving data between storage systems without disrupting users or applications. Cloud data is accessible from anywhere on the internet, meaning that increases and maintenance of data will be more complex it menaces storing ,managing ,and accessing of data through secure authentication channel is important in any organizations.

1.1Cloud computing

Cloud Computing means a type of computing in which services are delivered

through the Internet. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advent of this technology, the cost of computation, application hosting, content storage and delivery is reduced

significantly. Enterprises would need to align their applications, so as to exploit the architecture models that Cloud Computing offers. Some of the typical benefits are listed below:

1. Reduced Cost

There are a number of reasons to attribute Cloud technology with lower costs. The billing model is pay as per usage; the infrastructure is not purchased thus lowering maintenance. Initial expense and recurring expenses are much lower than traditional computing.

2. Increased Storage

With the massive Infrastructure that is offered by Cloud providers today, storage & maintenance of large volumes of data is a reality. Sudden workload spikes are also managed effectively & efficiently, since the cloud can scale dynamically.

3. Flexibility

This is an extremely important characteristic. With enterprises having to adapt, even more rapidly, to changing business conditions, speed to deliver is critical. Cloud computing stresses on getting applications to market very quickly, by using the most appropriate building blocks necessary for deployment.

1.2 Security

It is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely. Data security not only involves the encryption of the data, but also ensures that appropriate policies are enforced for data sharing. In addition, resource allocation and memory management algorithms also have to be secure. The big data issues are most acutely felt in certain industries, such as telecoms, web marketing. In addition to the above methodologies, cloud service will support third party authentication. The third party will be trusted by both the cloud service and accessing user. Third party authentication will add an additional security layer to the cloud service. Real time access control will be a good security measure in the cloud environment. In addition to access control to the cloud environment, operational control within a database in the cloud can be used to prevent configuration drift and unauthorized application changes. Multiple factors such as IP address, time of the day, and authentication method can be used in a flexible way to employ.

1.3 Big data

Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as inconsistent component performance and availability. In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements. The Organizations have a new option to outsource their massive data in the cloud without having to worry about the size of data or the capacity of memory. However, moving confidential and sensitive data from trusted domain of the data owners to public cloud will cause various security and privacy risks. Furthermore, the increasing amount of big data outsourced in the cloud increases the chance of breaching the privacy and security of these data. Despite all the research that has been done in this area, big data storage security and privacy remains one of the major concerns of organizations that adopt the cloud computing and big data technologies. Thus to ensure better data security we need to focus on two major problems which are the access control and encryption policies.

2. Research Methodology

Security issues of these systems and technologies are applicable to cloud computing. For example, it is very important for the network which interconnects the systems in a cloud to be secure. Also, virtualization paradigm in cloud computing results in several security concerns. For example, mapping of the virtual machines to the physical machines has to be performed very securely. Data security not only involves the encryption, race allocation and memory management algorithms also have to be secure. The big data issues are move the data, but also ensures that appropriate policies are enforced for data sharing. In addition to the above methodologies, cloud service will support third party authentication. The third party will be trusted by both those cloud service and accessing user. Third party authentication will add an additional security layer to the cloud service. Real time access control will be a good security measure in the cloud environment. In addition to access control to the cloud environment, operational control within a database in the cloud can be used to prevent configuration drift and unauthorized application changes. Multiple factors such as IP address, time of the day, and authentication method can be used in a flexible way to employ.

3. Literature survey

1. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools Moreover, cloud computing, big data and its applications by *Venkata N.Inukolly*.
2. Security policies describe the demeanor of a system through specific rules and are becoming an increasingly popular approach for static and dynamic environment applications by *Sailaja Arsis*.
3. Storing the data safely and efficiently on Cloud is one of the biggest challenges in Cloud computing. There is no guarantee that data stored on Cloud is securely protected. We propose a method to build a trusted computing environment by providing a secure platform in a Cloud computing system. by *ArjunKumar and HooJae Lee*.
4. Cloud computing is revolutionizing many ecosystems by providing organizations with computing resources featuring easy deployment, connectivity, configuration, automation and scalability. This paradigm shift raises a broad range of security and privacy issues that must be taken into consideration. Multi-tenancy, loss of control, and trust are key challenges in cloud by *Ali Gholami and Erwin Lause*.
5. Cloud computing and big data are an ideal combination for this. Together, they provide a solution which is both scalable and accommodating for big data and business analytics. The analytics advantage is going to be a huge benefit in today's world. Imagine all the information resources which will become easily accessible. Every field of life can benefit from this information. Let's look at these advantages in detail by *Kamil Riaz*.

4.Conclusion

Data management optimizes data performance and costs through capacity planning, storage utilization and data placement. All by moving data between storage systems without disrupting users or applications. Cloud data is accessible from anywhere on the internet, meaning that be increases and maintenance of data will be more complex it menace storing ,managing ,and accessing of data through secure authentication channel is important in any organizations.by applying authentication concepts in big data we can Increasing data confidentiality in organizations with big data. Protecting data from unauthorized users.and also by using third party authentication can increase the data security in cloud computing environment.

5. References

1. Torry Harris, et al., "Cloud computing – An overview" Basic overview of cloud,2009.
2. M. Armbrust, et al., "A view of cloud computing," Communications of the ACM, 2010, 53, p. 50-58.
3. R. Buyya, et al., "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, 2009, 25, p. 599-616.
4. D.-G. Feng, et al., "Study on Cloud Computing security," RuanJianXueBao/Journal of Software, 2011, 22, p. 71-83.
5. K. Hwang, et al., "Cloud security with virtualized defense and reputation-based trust management," in 8th IEEE International Symposium on Dependable, Autonomic and Secure Computing, DASC 2009, December 12, 2009 - December 14, 2009, Chengdu, China, 2009, p. 717-722.
6. D. Zissis and D. Lekkas, "Addressing cloud computing security issues," Future Generation Computer Systems, 2012, 28, p. 583-592.
7. H. Takabi, et al., "Security and privacy challenges in cloud computing environments," IEEE Security and Privacy, 2010, 8, p. 24-31.
8. L.-Q. Tian, et al., "Node behavior trust evaluation based on behavior evidence in WSNs," in 2010 2nd International Conference on Future Computer and Communication, ICFCC 2010, May 21, 2010 - May 24, 2010, Wuhan, China, 2010, p. 1312-1317.
9. L.-q. Tian, et al., "Evaluation of user behavior trust in cloud computing," in 2010 International Conference on Computer Application and System Modeling, ICCASM 2010, October 22, 2010 - October 24, 2010, Shanxi, Taiyuan, China, 2010, p. 7567-7572.
10. L. Jiang, et al., "A new evidential trust model for open distributed systems," Expert Systems with Applications, 2012, 39, p. 3772-3782.
11. Y. Peng, G. Kou, et al., "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery," International Journal of Information Technology & Decision Making, 2008, 7, p. 639 682.
12. AkankashaDexit et al., "A trust evolution model for cloud computing" A review for the current academic students about cloud computing.

A COMPARATIVE ANALYSIS OF HIGH- PERFORMANCE COMPUTING ON THE CLOUD FOR SCIENTIFIC APPLICATIONS

Dr.Chinthireddy Prakash¹
Department of Management Studies
Siddhartha Institute of Engg & Technology,
Ibrahimpattam, Hyderabad

Dr.K.Rajeshwar Rao²
Department of CSE
Siddhartha Institute of Engg & Technology,
Ibrahimpattam, Hyderabad

Abstract

We propose the primary systematic cost estimation display for assessing cloud database costs in plain and encoded cases from an occupant's perspective in a medium-term period. It considers the changeability of cloud costs and the likelihood that the database workload may change amid the assessment time frame. This model is instanced concerning a few cloud supplier offers and related genuine costs. Obviously, versatile encryption impacts the costs identified with capacity size and system use of a database benefit.

1. INTRODUCTION

Distributed computing has been chosen as the consideration of researchers a forceful advantage to run HPC applications at a possibly low cost. However, as a replacement framework, it is unclear whether mists are prepared for running logical applications with a practical implement for every buck. This work gives a comprehensive valuation of EC2 cloud in round the corner. I initially divide the possibilities of the cloud by measuring the crude implementation of several administrations of AWS, for example, register, memory, system and I/O. In view of the findings on the crude accomplishment, and after that measure the implement of the logical applications executing in the cloud. Finally, in contrast to the implementation of AWS and a private cloud, with a definite end goal to discover the main driver of its limitations while running logical applications. This project plans to survey the capacity of the cloud to perform well, and furthermore to measure the cost of the cloud as far as both crude implementation and logical applications implement. Moreover, I assess different administrations including S3, EBS and Dynamo DB among numerous AWS benefits keeping in mind the end goal to survey the capacities of those to be utilized by logical applications and systems. This likewise assess a genuine logical registering application through the Swift parallel scripting framework at scale. Outfitted with both point by point benchmarks to gage expected implement and a definite

money related cost examination, I expect this paper will be a formula cookbook for researchers to enable them to choose where to send and run their logical applications between open mists, private mists, or half breed mists.

2. LITERATURE REVIEW

Writing study is the most authoritative stride in programming advancement handle. Before building up the device it is significant to decide the time element, economy and organization quality. Once these things are contented, ten subsequent stages are to figure out which working framework and dialect can be utilized for building up the device. Once the software engineers begin collecting the apparatus the developers require part of external support. This support can be learnt from senior software engineers, from book or from sites. Before structure the framework the above thought careful for building up the proposed outline.

2.1 Evaluating Interconnect and Virtualization Performance for High Performance Computing

Researchers are gradually bearing in mind distributed computing stages to fulfill their computational needs. Past work has established that virtualized cloud conditions can have critical implement affect. However there is as yet a controlled comprehension of the idea of overheads and the sort of operations that may do well in these conditions. In this subtle elements of benchmarking comes about that label the virtualization overhead and its effect on implement and additionally analyze the implement of different interconnect innovations with a view to empathetic the implement effects of different decisions. Our outcomes demonstrate that virtualization can have a remarkable effect upon implement, with no less than a 60% implement punishment.

2.2 A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing

Distributed computing is increasing today as a business framework that positions the requirement for keeping up costly reckoning equipment. Using virtualization, mists guarantee to address with the same shared arrangement of physical assets an expansive client base with various requirements. Along these lines, mists guarantee to be for researchers another option to bunches, lattices, and supercomputers. Notwithstanding, virtualization may incite huge implement punishments for the demanding logical figuring workloads. In this work showing an assessment of the usefulness of the present distributed computing administrations for logical registering. I examine the implement of the Amazon EC2 stage utilizing small scale benchmarks and pieces.

2.3 The Magellan Report on Cloud Computing for Science

Distributed computing has served the supplies of big business web applications during the previous couple of years. The expression “distributed computing” has been used to allude to numerous distinctive ideas (e.g., Map Reduce, open mists, private mists, and so on.), advances (e.g., virtualization, Apache Hadoop), and administration models (e.g., Infrastructure as-a-Service [IaaS], Platform-as-a-Service [PaaS], Software-as-a-Service [SaaS]). Mists have been seemed to give various key advantages including cost investment funds, fast versatility, convenience, and dependability. Distributed computing has been particularly productive with clients lacking significant IT framework or clients who have fast outgrown their current limit.

3. Existing System

The distributed computing worldview is effectively joining as the fifth utility , however this positive pattern is somewhat constrained by worries about data classification and indistinct expenses over a medium-long haul .I am occupied with the Database as a Service worldview (DBaaS) that represents a few research challenges as far as security and cost assessment from an inhabitant's perspective. Most outcomes concerning encryption for cloud-based administrations are in appropriate to the database worldview. Other encryption plans, which permit the implement of SQL operations over scrambled information, either experience the ill effects of implement cutoff points or they require the decision of which encryption plot must be received for every database section and SQL operations.

4. Proposed System

The proposed engineering ensures in a versatile way the best level of information secrecy for any database workload, notwithstanding when the arrangement of SQL questions progressively changes. The versatile encryption conspire, which was at first proposed for applications not alluding to the cloud, scrambles each plain section into numerous encoded segments, and each esteem is typified into various layers of encryption, so that the external layers ensure higher secrecy yet bolster less calculation abilities concerning the inward layers.

I propose the primary systematic cost estimation display for assessing cloud database costs in plain and encoded cases from an occupant's perspective in a medium-term period. It considers the changeability of cloud costs and the likelihood that the database workload may change amid the assessment time frame. This model is instanced concerning a few cloud supplier offers and related genuine costs. Obviously,

versatile encryption impacts the costs identified with capacity size and system use of a database benefit.

5. IMPLEMENTATION

- Adaptive encryption
- Metadata structure
- Encrypted database management
- Cost Estimation of cloud database services
- Cost model
- Cloud pricing models
- Usage Estimation

5.1.1 Adaptive Encryption

Tomcat is an open source web server created by Apache Group. Apache Tomcat is the servlet compartment that is utilized as a part of the official Reference Implementation for the Java Servlet and Java Server Pages innovations. The Java Servlet and Java Server Pages details are produced by Sun under the Java Community Process. Web Servers like Apache Tomcat bolster just web parts while an application server underpins web segments and in addition business segments (BEAs Web rationale, is one of the well-known application server). To build up a web application with jsp/servlet introduce any web server like JRun, Tomcat and so on to run your application.

5.1.2 Metadata Structure

Metadata incorporate all data that permits an authentic customer knowing the ace key to execute SQL operations over a scrambled database. They are sorted out and put away at a table-level granularity to diminish correspondence overhead for recovery, and to enhance administration of simultaneous SQL operations. I characterize all metadata data related to a table as table metadata. Give us a chance to portray the structure of a table metadata.

Table metadata incorporates the correspondence between the plain table name and the scrambled table name in light of the fact that each encoded table name is arbitrarily created. Besides, for every section of the first plain table it likewise incorporates a segment metadata parameter containing the name and the information sort of the comparing plain segment (e.g., number, string, and timestamp). Every segment

metadata is related to at least one onion metadata, the same number of as the quantity of onions identified with the segment.

5.1.3 Encrypted Database Management

The database head creates an ace key, and uses it to instate the engineering metadata. The ace key is then dispersed to honest to goodness customers. Each table creation requires the addition of another column in the metadata table. For each table creation, the head includes a segment by determining the section name, information sort and privacy parameters.

These last are the most essential for this venture since they incorporate the arrangement of onions to be related with the section, the beginning layer (signifying the genuine layer at creation time) and the field privacy of every onion. On the off chance that the executive does not indicate the secrecy parameters of a section, at that point they are naturally picked by the customer concerning an occupant's arrangement. Ordinarily, the default strategy accept that the beginning layer of every onion is set to its most grounded encryption calculation.

5.1.4 Cost Estimation of Cloud Database Services

An inhabitant that is keen on assessing the cost of porting its database to a cloud stage. This porting is a vital choice that must assess secrecy issues and the related expenses over a medium-long haul. Thus, I propose a model that incorporates the overhead of encryption plans and changeability of database workload and cloud costs. The proposed show is sufficiently general to be connected to the most well-known cloud database administrations, for example, Amazon Relational Database Service.

5.1.5 Cost Model

The cost of a cloud database service can be estimated as a function of three main parameters:

$$\text{Cost} = f(\text{Time}, \text{Pricing}, \text{Usage})$$

where:

- Time: recognizes the time interim T for which the inhabitant requires the administration.

- Pricing: alludes to the costs of the cloud supplier for membership and asset utilization; they regularly have a tendency to decrease amid T.
- Usage: signifies the aggregate sum of assets utilized by the inhabitant; it ordinarily increments amid T. Keeping in mind the end goal to detail the evaluating trait, indicate that cloud suppliers embrace two membership strategies: the on-request strategy enables an inhabitant to paper-utilize and to pull back its membership whenever; the reservation approach requires the occupant to confer ahead of time for a reservation period. Subsequently, I recognize charging costs relying upon asset utilization and reservation costs meaning extra expenses for responsibility in return for bring down pay-per-utilize costs. Charging costs are charged occasionally to the occupant each charging period.

5.1.6 Cloud Pricing Models

Well known cloud database suppliers embrace two diverse charging capacities, that I call direct L and layered T. Give us a chance to consider a bland asset x , I characterize as x_b its use at the b -th charging period and $p_x b$ its cost. On the off chance that the charging capacity is layered, the cloud supplier utilizes distinctive costs for various scopes of asset utilization. Give us a chance to characterize Z as the quantity of levels, and $[\hat{x}_1, \dots, \hat{x}_{Z-1}]$ as the arrangement of edges that characterize every one of the levels. The uptime and the capacity charging elements of Amazon RDS are straight, while the system utilization is a layered charging capacity. Then again, the uptime charging elements of Azure SQL is straight, while the capacity and system charging capacities are layered.

5.1.6 Usage Estimation

The uptime is effectively quantifiable; it is harder to evaluate precisely the utilization of capacity and system, since they rely on upon the database structure, the workload and the utilization of encryption. I now propose a procedure for the estimation of capacity and system use because of encryption. For clearness, I characterize s_p , s_e , s_a as the capacity use in the plaintext, scrambled, and adaptively encoded databases for one charging period. Thus, n_p , n_e , n_a speak to arrange use of the three designs. I expect that the inhabitant knows the database structure and the question workload and accept that every section an A stores r_a esteems. By signifying as VP_a normal stockpiling size of

each plaintext esteem put away in section and, I gauge the capacity of the plaintext database.

OUTPUT SCREENS

Home Page:

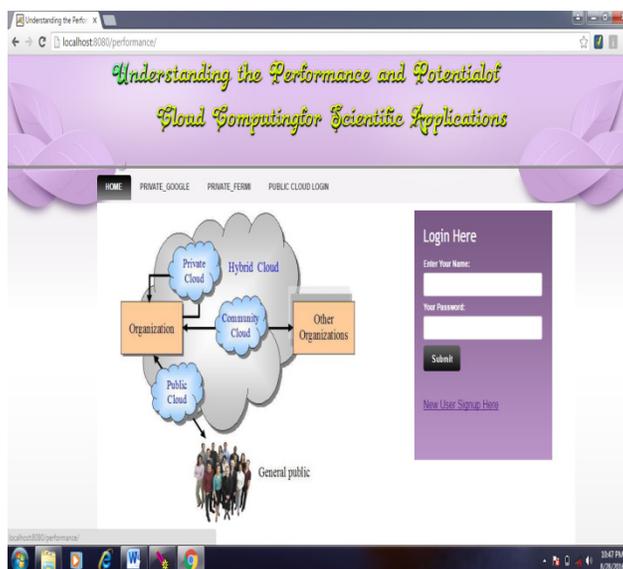


Fig 6.1: Home Page

Cloud User Register Page:

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

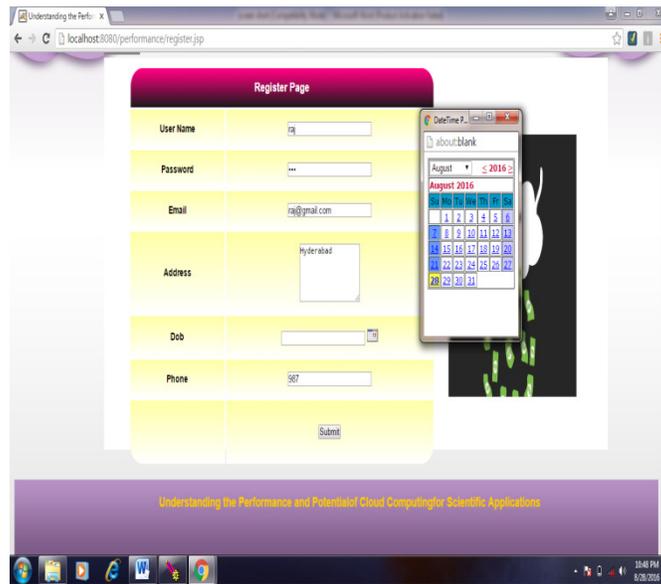


Fig 6.2: Cloud User Registration Page

User Login Page:

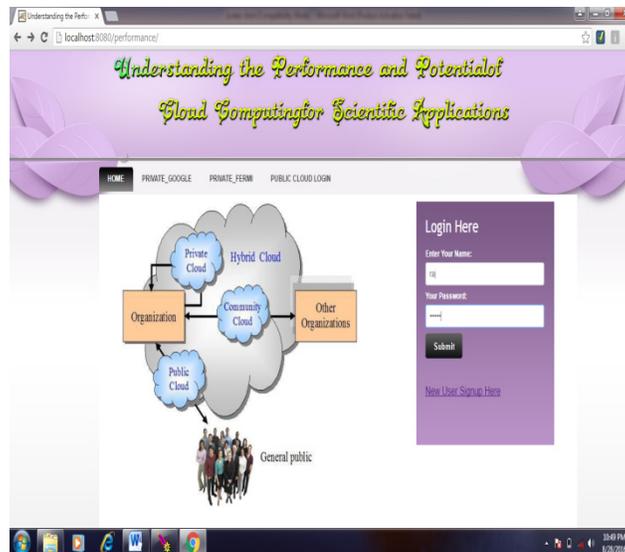


Fig 6.3: User Login Page

User Page:

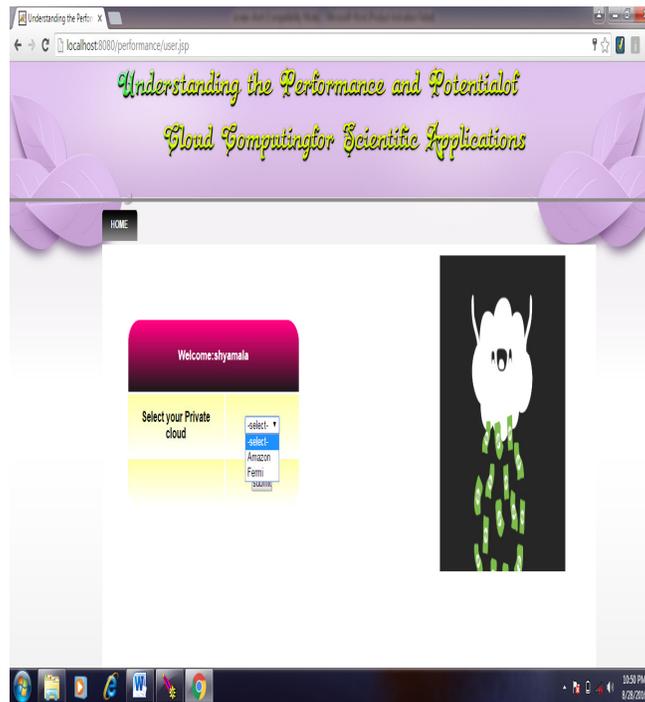


Fig 6.4: User Page

Amazon Cloud Login Page:

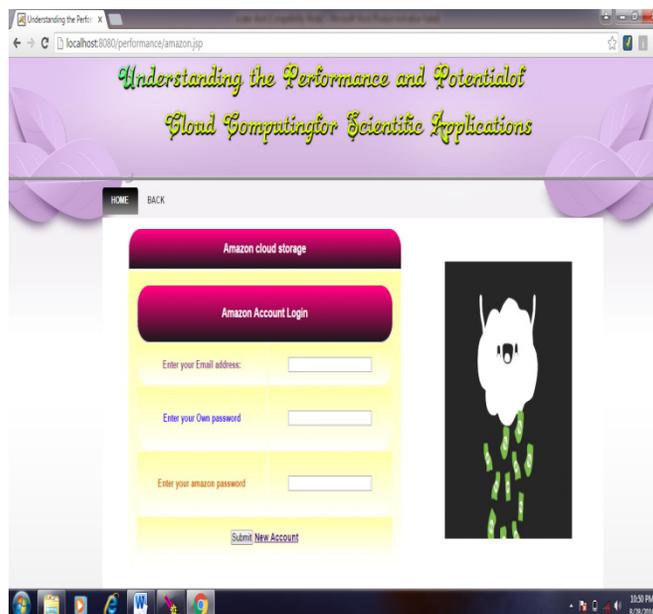
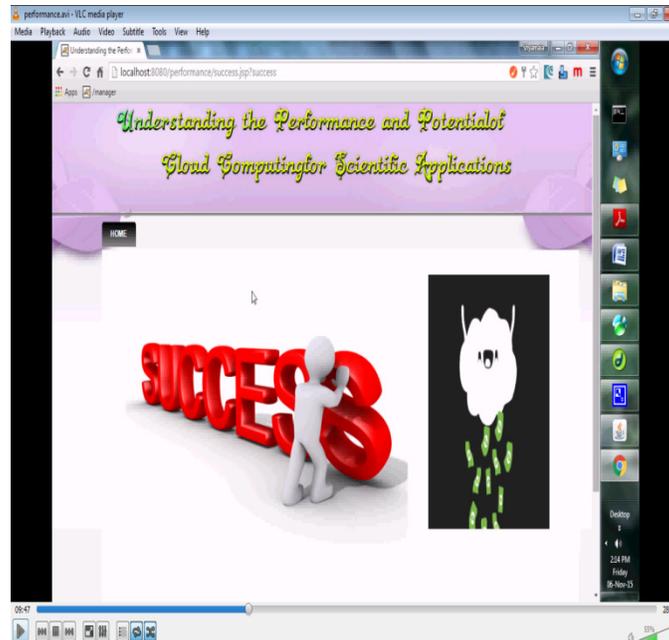


Fig 6.5: Amazon Cloud Login Page

Success Page:



8. CONCLUSION AND FUTURE ENHANCEMENT

I propose an Understanding the implement and Potential of Cloud Computing for Scientific Applications. Amazon EC2 gives capable occasions that are equipped for running HPC applications. I assessed the I/O implement of Amazon cases and capacity administrations like EBS and S3 give a more extensive perspective of EC2 by investigating the implement of cloud benefits that could be utilized as a part of present day logical applications. More logical systems and applications have transformed into utilizing cloud administrations to better use the capability of Cloud. Our work address the capacity administrations implement both on miniaturized scale benchmarks and in addition the implement while being utilized by information serious applications.

9. BIBIOLOGRAPHY

[1] Amazon EC2 Instance Types, Amazon Web Services, [online] 2013, <http://aws.amazon.com/ec2/instance-types/> (Accessed: 2 November 2013)

- [2] Amazon Elastic Compute Cloud (Amazon EC2), Amazon Web Services, [online] 2013, <http://aws.amazon.com/ec2/> (Accessed: 2 November 2013)
- [3] Amazon Simple Storage Service (Amazon S3), Amazon Web Services, [online] 2013, <http://aws.amazon.com/s3/> (Accessed: 2 November 2013)
- [4] Iperf, Souceforge, [online] June 2011, <http://sourceforge.net/projects/iperf/> (Accessed: 2 November 2013)
- [5] A. Petitet, R. C. Whaley, J. Dongarra, A. Cleary. "HPL", (netlib.org), [online] September 2008, <http://www.netlib.org/benchmark/hpl/> (Accessed: 2 November 2013)
- [6] J. J. Dongarra, S. W. Otto, M. Snir, and D. Walker, "An introduction to the MPI standard," Tech. Rep. CS-95-274, University of Tennessee, Jan. 1995
- [7] Release: Amazon EC2 on 2007-07-12, Amazon Web Services, [online] 2013, <http://aws.amazon.com/releasenotes/Amazon-EC2/3964> (Accessed: 1 November 2013)
- [8] K. Yelick, S. Coghlan, B. Draney, and R. S. Canon, "The Magellan report on cloud computing for science," U.S. Department of Energy, Tech. Rep., 2011
- [9] L. Ramakrishnan, R. S. Canon, K. Muriki, I. Sakrejda, and N. J. Wright. "Evaluating Interconnect and virtualization performance for high performance computing", ACM Performance Evaluation Review, 2012
- [10] P. Mehrotra, et al. 2012. "Performance evaluation of Amazon EC2 for NASA HPC applications" In *Proceedings of the 3rd work-shop on Scientific Cloud Computing* (ScienceCloud '12). ACM, New York, NY, USA, pp. 41-50
- [11] Q. He, S. Zhou, B. Kobler, D. Duffy, and T. McGlynn. "Case study for running HPC applications in public clouds," In Proc. of ACM Symposium on High Performance Distributed Computing, 2010
- [12] G. Wang and T. S. Eugene Ng. "The Impact of Virtualization on Network Performance of Amazon EC2 Data Center". In IEEE INFOCOM, 2010

Fuzzy Keyword Search over Encrypted Data in Cloud Computing

¹ Boga Jayaram,, ²Mysa Kalyana chakravarthy

¹ Assistant Professor, Department of Computer Science & Engineering, Balaji Institute of Technology & science, Telangana,India,

² Assistant Professor, Department of Computer Science & Engineering , St.Marry's Engineering College, Telangana, India

Email: ¹ jayaramboga@gmail.com , ² mkalyan8@gmail.com

Abstract—As Cloud Computing becomes prevalent, more and more sensitive information are being centralized into the cloud. For the protection of data privacy, sensitive data usually have to be encrypted before outsourcing, which makes effective data utilization a very challenging task. Although traditional searchable encryption schemes allow a user to securely search over encrypted data through keywords and selectively retrieve files of interest, these techniques support only *exact* keyword search. That is, there is no tolerance of minor typos and format inconsistencies which, on the other hand, are typical user searching behavior and happen very frequently. This significant drawback makes existing techniques unsuitable in Cloud Computing as it greatly affects system usability, rendering user searching experiences very frustrating and system efficacy very low. In this paper, for the first time we formalize and solve the problem of effective fuzzy keyword search over encrypted cloud data while maintaining keyword privacy. Fuzzy keyword search greatly enhances system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when *exact* match fails. In our solution, we exploit edit distance to quantify keywords similarity and develop an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads. Through rigorous security analysis, we show that our proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search

I. INTRODUCTION

As Cloud Computing becomes prevalent, more and more sensitive information are being centralized into the cloud, such as emails, personal health records, government documents, etc. By storing their data into the cloud, the data owners can be relieved from the burden of data storage and maintenance so as to enjoy the on-demand high quality data storage service. However, the fact that data owners and cloud server are not in the same trusted domain may put the outsourced data at risk, as the cloud server may no longer be fully trusted. It follows that sensitive data usually should be encrypted prior to outsourcing for data privacy and combating unsolicited accesses.

However, data encryption makes effective data utilization a very challenging task given that there could be a large amount of outsourced data files. Moreover, in Cloud Computing, data owners may share their outsourced data with a large number of users. The individual users might want to only retrieve certain specific data files they are interested in during a given session. One of the most popular ways is to selectively retrieve files through keyword-based search instead of retrieving all the encrypted files back which is completely impractical in cloud computing scenarios. Such keyword-based search technique allows users to selectively retrieve files of interest and has been widely applied in plaintext search scenarios, such as Google search [1]. Unfortunately, data encryption restricts user's ability to perform keyword search and thus makes the traditional plaintext search methods unsuitable for Cloud Computing. Besides this, data encryption also demands the protection of keyword privacy since keywords usually contain important information related to the data files. Although encryption of keywords can protect keyword privacy, it further renders the traditional plaintext search techniques useless in this scenario. To securely search over encrypted data, searchable encryption techniques have been developed in recent years [2]–[10]. Searchable encryption schemes usually build up an index for each keyword of interest and associate the index with the files that contain the keyword. By integrating the trapdoors of keywords within the index information, effective keyword search can be realized while both file content and keyword privacy are well-preserved. Although allowing for performing searches securely and effectively, the existing searchable encryption techniques do not suit for cloud computing scenario since they support only exact keyword search. That is, there is no tolerance of minor typos and format inconsistencies. It is quite common that users' searching input might not exactly match those pre-set keywords due to the possible typos, such as Illinois and Ilinois, representation inconsistencies, such as PO BOX and P.O. Box, and/or her lack of exact knowledge about the data. The naive way to support fuzzy keyword search is through simple spell check mechanisms. However, this approach does not completely solve the problem and sometimes can be ineffective due to the following reasons: on the one hand, it requires additional interaction of user to determine the correct word from the candidates generated by the spell check algorithm, which unnecessarily costs user's extra computation effort; on the other hand, in case that user accidentally types some other valid keywords by mistake (for example, search for "hat" by carelessly typing "cat"), the spell check algorithm would not even work at all, as it can never differentiate between two actual valid words. Thus, the drawbacks of existing schemes signifies the important

need for new techniques that support searching flexibility, tolerating both minor typos and format inconsistencies.

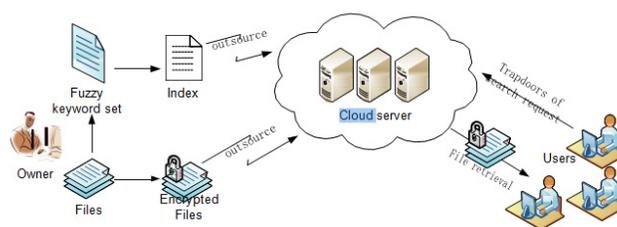
In this paper, we focus on enabling effective yet privacy-preserving fuzzy keyword search in Cloud Computing. To the best of our knowledge, we formalize for the first time the problem of effective fuzzy keyword search over encrypted cloud data while maintaining keyword privacy. Fuzzy keyword search greatly enhances system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. More specifically, we use edit distance to quantify keywords similarity and develop a novel technique, i.e., an wildcard-based technique, for the construction of fuzzy keyword sets. This technique eliminates the need for enumerating all the fuzzy keywords and the resulted size of the fuzzy keyword sets is significantly reduced. Based on the constructed fuzzy keyword sets, we propose an efficient fuzzy keyword search scheme. Through rigorous security analysis, we show that the proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search.

The rest of paper is organized as follows: Section II summarizes the features of related work. Section III introduces the system model, threat model, our design goal and briefly describes some necessary background for the techniques used in this paper. Section IV shows a straightforward construction of fuzzy keyword search scheme. Section V provides the detailed description of our proposed schemes, including the efficient constructions of fuzzy keyword set and fuzzy keyword search scheme. Section VI presents the security analysis. Finally, Section VIII concludes the paper.

I. RELATED WORK

Plaintext fuzzy keyword search. Recently, the importance of fuzzy search has received attention in the context of plaintext searching in information retrieval community [11]–[13]. They addressed this problem in the traditional information-access paradigm by allowing user to search without using try-and-see approach for finding relevant information based on approximate string matching. At the first glance, it seems possible for one to directly apply these string matching algorithms to the context of searchable encryption by computing the trapdoors on a character base

within an alphabet. However, this trivial construction suffers from the dictionary and statistics attacks and fails to achieve the search privacy. Searchable encryption. Traditional searchable encryption [2]–[8], [10] has been widely studied in the context of cryptography. Among those works, most are focused on efficiency improvements and security definition formalizations. The first construction of searchable encryption was proposed by Song et al. [3], in which each word in the document is encrypted independently under a special two-layered encryption construction. Goh [4] proposed to use Bloom filters to construct the indexes for the data files. To achieve more efficient search, Chang et al. [7] and Curtmola et al. [8] both proposed similar “index” approaches, where a single encrypted hash table index is built for the entire file collection. In the index table, each entry consists of the trapdoor of a keyword and an encrypted set of file identifiers whose corresponding data files contain the keyword. As a complementary approach, Boneh et al. [5] presented a public-key based searchable encryption scheme, with an analogous scenario to that of [3]. Note that all these existing schemes support only exact keyword search, and thus are not suitable for Cloud Computing.



. 1: Architecture of the fuzzy keyword search

Others. Private matching [14], as another related notion, has been studied mostly in the context of secure multiparty computation to let different parties compute some function of their own data collaboratively without revealing their data to the others. These functions could be intersection or approximate private matching of two sets, etc. The private information retrieval [15] is an often-used technique to retrieve the matching items secretly, which has been widely applied in information retrieval from database and usually incurs unexpectedly computation complexity.

III. PROBLEM FORMULATION

A. System Model

In this paper, we consider a cloud data system consisting of data owner, data user and cloud server. Given a collection of n encrypted data files $\mathcal{F} = (F_1, F_2, \dots, F_N)$ stored in the cloud server, a predefined set of distinct keywords $W = w_1, w_2, \dots, w_p$, the cloud server provides the search service for the authorized users over the encrypted data. We assume the authorization between the data owner and users is appropriately done. An authorized user types in a request to selectively retrieve data files of his/her interest. The cloud server is responsible for mapping the searching request to a set of data files, where each file is indexed by a file ID and linked to a set of keywords. The fuzzy keyword search scheme returns the search results according to the following rules: 1) if the user's searching input exactly matches the pre-set keyword, the server is expected to return the files containing the keyword1;

2) if there exist typos and/or format inconsistencies in the searching input, the server will return the closest possible results based on pre-specified similarity semantics (to be formally defined in section III-D). An architecture of fuzzy keyword search is shown in the Fig. 1.

B. Threat Model

We consider a semi-trusted server. Even though data files are encrypted, the cloud server may try to derive other sensitive information from users' search requests while performing keyword-based search over. Thus, the search should be conducted in a secure manner that allows data files to be securely retrieved while revealing as little information as possible to the cloud server. In this paper, when designing fuzzy keyword search scheme, we will follow the security definition deployed in the traditional searchable encryption [8]. More specifically, it is required that nothing should be leaked from the remotely stored files and index beyond the outcome and the pattern of search queries.

C. Design Goals

In this paper, we address the problem of supporting efficient yet privacy-preserving fuzzy keyword search services over encrypted cloud data. Specifically, we have the following goals: i) to explore new mechanism for constructing storage- efficient fuzzy keyword sets; ii) to design efficient and effective fuzzy search scheme based on the constructed fuzzy keyword sets; iii) to validate the security of the proposed scheme.

D. Preliminaries

Edit Distance There are several methods to quantitatively measure the string similarity. In this paper, we resort to the well-studied edit distance [16] for our purpose. The edit distance $ed(w_1, w_2)$ between two words w_1 and w_2 is the number of operations required to transform one of them into the other. The three primitive operations are 1) Substitution: changing one character to another in a word; 2) Deletion: deleting one character from a word; 3) Insertion: inserting a single character into a word. Given a keyword w , we let

$S_{w,d}$ denote the set of words w_j satisfying $ed(w, w_j) \leq d$ for a certain integer d .

Fuzzy Keyword Search Using edit distance, the definition of fuzzy keyword search can be formulated as follows: Given a collection of n encrypted data files $= (F_1, F_2, \dots, F_n)$ stored in the cloud server, a set of distinct keywords $W = w_1, w_2, \dots, w_p$ with predefined edit distance d , and a searching input (w, k) with edit distance k ($k \leq d$), the execution of fuzzy keyword search returns a set of file

IDs whose corresponding data files possibly contain the word w , denoted as FID_w : if $w = w_i \in W$, then return FID_{w_i} ; otherwise, if $w \notin W$, then return $\{FID_{w_i}\}$, where $ed(w, w_i) \leq k$. Note that the above definition is based on the assumption that $k \leq d$. In fact, d can be different for distinct keywords and the system will return $\{FID_{w_i}\}$ satisfying

$ed(w, w_i) \leq \min\{k, d\}$ if exact match fails.

I. THE STRAIGHTFORWARD APPROACH

Before introducing our construction of fuzzy keyword sets, we first propose a straightforward approach that achieves all the functions of fuzzy keyword search, which aims at providing an overview of how fuzzy search scheme works over encrypted data.

Assume $\Pi=(\text{Setup}(1\lambda), \text{Enc}(\text{sk},), \text{Dec}(\text{sk},))$ is a symmetric encryption scheme, where sk is a secret key, $\text{Setup}(1\lambda)$ is the setup algorithm with security parameter λ , $\text{Enc}(\text{sk},)$ and $\text{Dec}(\text{sk},)$ are the encryption and decryption algorithms, respectively. Let T_{wi} denote a trapdoor of keyword w_i . Trapdoors of the keywords can be realized by applying a one-way function f , which is similar as [2], [4], [8]: Given a keyword w_i and a secret key sk , we can compute the trapdoor of w_i as $T_{wi} = f(\text{sk}, w_i)$.

The scheme of the fuzzy keyword search goes as follows:

We begin by constructing the fuzzy keyword set $S_{w_i,d}$

for each keyword $w_i \in W$ ($1 \leq i \leq p$) with edit distance

d . The intuitive way to construct the fuzzy keyword set of w_i is to enumerate all possible words w_{ij} that satisfy the similarity criteria $\text{ed}(w_i, w_{ij}) \leq d$, that is, all the words with edit distance d from w_i are listed. For example, the following

is the listing variants after a substitution operation on the first character of keyword CASTLE: AASTLE, BASTLE, DASTLE, , YASTLE, ZASTLE. Based on the resulted fuzzy keyword sets, the fuzzy search over encrypted data is conducted as follows:

1) To build an index for w_i , the data owner computes trapdoors $T_{w_{ij}} = f(\text{sk}, w_{ij})$ for each $w_{ij} \in S_{w_i,d}$ with a secret key sk shared between data owner and authorized users. The data owner also encrypts FID_{w_i} as $\text{Enc}(\text{sk}, \text{FID}_{w_i} \parallel w_i)$. The index table $(T_{w_{ij}}, \text{Enc}(\text{sk}, \text{FID}_{w_i} \parallel w_i))$ $w \in W$ and

files are outsourced serv

storage; 2) To search with w , the authorized user computes the trapdoor T_w of w and sends it to the server; 3) Upon receiving the search request T_w , the server compares it with the index table and returns all the possible encrypted file identifiers $\text{Enc}(\text{sk}, \text{FID}_{w_i} \oplus w_i)$ according to the fuzzy keyword definition in section III-D. The user decrypts the returned results and retrieves relevant files of interest.

This straightforward approach apparently provides fuzzy keyword search over the encrypted files while achieving search privacy using the technique of secure trapdoors. However, this approach has serious efficiency disadvantages. The simple enumeration method in constructing fuzzy keyword sets would introduce large storage complexities, which greatly affect the usability. Recall that in the definition of edit distance, substitution, deletion and insertion are three kinds of operations

in computation of edit distance. The numbers of all similar words of w_i satisfying $\text{ed}(w_i, w_j) \leq d$ for $d = 1, 2$ and

3 are approximately $2k$, $2k^2$, and $4k^3$, respectively.

For example, assume there are 104 keywords in the file collection with average keyword length 10, $d = 2$, and the output length of hash function is 160 bits, then, the resulted storage cost for the index will be 30GB. Therefore, it brings forth the demand for fuzzy keyword sets with smaller size.

V. CONSTRUCTIONS OF EFFECTIVE FUZZY KEYWORD

SEARCH IN CLOUD

The key idea behind our secure fuzzy keyword search is two-fold: 1) building up fuzzy keyword sets that incorporate not only the exact keywords but also the ones differing slightly due to minor typos, format inconsistencies, etc.; 2) designing an efficient and secure searching approach for file retrieval based on the resulted fuzzy keyword sets.

A. Advanced Technique for Constructing Fuzzy Keyword Sets

To provide more practical and effective fuzzy keyword search constructions with regard to both storage and search efficiency, we now propose an advanced technique to improve the straightforward approach for constructing the fuzzy keyword set. Without

loss of generality, we will focus on the case of edit distance $d = 1$ to elaborate the proposed advanced technique. For larger values of d , the reasoning is similar. Note that the technique is carefully designed in such a way that while suppressing the fuzzy keyword set, it will not affect the search correctness.

Wildcard-based Fuzzy Set Construction In the above straightforward approach, all the variants of the keywords have to be listed even if an operation is performed at the same position. Based on the above observation, we proposed to use a wildcard to denote edit operations at the same position.

The wildcard-based fuzzy set of w_i with edit distance d is denoted as $Sw_{i,d} = \{Sw_{i,0}, Sw_{i,1}, \dots, Sw_{i,d}\}$, where $Sw_{i,\tau}$ denotes the set of words w_{ij} with τ wildcards. Note each

wildcard represents an edit operation on w_i . For example,

for the keyword CASTLE with the pre-set edit distance 1, its wildcard-based fuzzy keyword set can be constructed as $SCASTLE,1 = \{CASTLE, *CASTLE, *ASTLE, C*ASTLE, C*STLE, \dots, CASTL*E, CASTL*, CASTLE*\}$. The total

number of variants on CASTLE constructed in this way

is only $13 + 1$, instead of $13 \cdot 26 + 1$ as in the above exhaustive enumeration approach when the edit distance is set to be 1. Generally, for a given keyword w_i with length A the size of $Sw_{i,1}$ will be only $2A + 1 + 1$, as compared to $(2A + 1) \cdot 26 + 1$ obtained in the straightforward approach.

The larger the pre-set edit distance, the more storage overhead can be reduced: with the same setting of the example in the straightforward approach, the proposed technique can help reduce the storage of the index from 30GB to approximately 40MB. In case the edit distance is set to be 2 and 3, the size of $Sw_{i,2}$ and $Sw_{i,3}$ will be $C_1 + C_1 \cdot C_1 + 2C_2$ and $C_1 + C_3 + 2C_2 + 2C_2 \cdot C_1$. In other words, the number is only $O(Ad)$ for the keyword with length A and edit distance d .

B. The Efficient Fuzzy Keyword Search Scheme

Based on the storage-efficient fuzzy keyword sets, we show how to construct an efficient and effective fuzzy keyword search scheme. The scheme of the fuzzy keyword search goes as follows:

1) To build an index for w_i with edit distance d , the data owner first constructs a fuzzy keyword set $S_{w_i,d}$ using the wildcard based technique. Then he computes trapdoor set T_{w_i} for each $w_i \in S_{w_i,d}$ with a secret key sk shared between data owner and authorized users. The data owner encrypts FID_{w_i} as $Enc(sk, FID_{w_i} || w_i)$. The index table $(T_{w_i} | w_i \in S_{w_i,d}, Enc(sk, FID_{w_i} || w_i))$ $w_i \in W$ and encrypted data files are outsourced to the cloud server for storage;

2) To search with (w, k) , the authorized user computes the trapdoor set $T_{w,r} | w_r \in S_{w,k}$, where $S_{w,k}$ is also derived from the wildcard-based fuzzy set construction. He then sends $\{T_{w,r} | w_r \in S_{w,k}\}$ to the server;

3) Upon receiving the search request $\{T_{w,r} | w_r \in S_{w,k}\}$, the server compares them with the index table and returns all the possible encrypted file identifiers $Enc(sk, FID_{w_i} || w_i)$ according to the fuzzy keyword definition in section III-D. The user decrypts the returned results and retrieves relevant files of interest.

In this construction, the technique of constructing search request for w is the same as the construction of index for a keyword. As a result, the search request is a trapdoor set based on $S_{w,k}$, instead of a single trapdoor as in the straightforward approach. In this way, the searching result correctness can be ensured.

I. SECURITY ANALYSIS

In this section, we analyze the correctness and security of the proposed fuzzy keyword search scheme. At first, we show the correctness of the schemes in terms of two aspects, that is, completeness and soundness.

Theorem 1: The wildcard-based scheme satisfies both completeness and soundness. Specifically, upon receiving the request of w , all of the keywords w_i will be returned if and only if $ed(w, w_i) \leq k$.

The proof of this Theorem can be reduced to the following Lemma:

Lemma 1: The intersection of the fuzzy sets $S_{w_i, d}$ and $S_{w, k}$ for w_i and w is not empty if and only if $ed(w, w_i) \leq k$. Proof: First, we show that $S_{w_i, d} \cap S_{w, k}$ is not empty when $ed(w, w_i) \leq k$. To prove this, it is enough to find an element in $S_{w_i, d} \cap S_{w, k}$. Let $w = a_1a_2 \dots a_n$ and $w_i = b_1b_2 \dots b_m$, where all these a_i and b_j are single characters. After $ed(w, w_i)$ edit operations, w can be changed to w_i according to the definition of edit distance. Let $w^* = a^*1 a^*2 \dots a^*m$, where $a^*i = a_j$ or $a^*i = _$ if any operation is performed at this position. Since the edit operation is inverted, from w_i , the same positions containing wildcard at w^* will be performed. Because $ed(w, w_i) \leq k$,

w^* is included in both $S_{w_i, d}$ and $S_{w, k}$, we get the result that

$S_{w_i, d} \cap S_{w, k}$ is not empty.

Next, we prove that $S_{w_i, d} \cap S_{w, k}$ is empty if $ed(w, w_i) > k$. The proof is given by reduction. Assume there exists an w^* belonging to $S_{w_i, d} \cap S_{w, k}$. We will show that $ed(w, w_i) \leq k$, which reaches a contradiction. First, from the assumption that $w^* \in S_{w_i, d} \cap S_{w, k}$, we can get the number of wildcard in w^* , which is denoted by n^* , is not greater than k . Next, we prove that $ed(w, w_i) \leq n^*$. We will prove the inequality with induction method. First, we prove it holds when $n^* = 1$. There are nine cases should be considered: If w^* is derived from the

operation of deletion from both w_i and w , then, $ed(w_i, w) = 1$

because the other characters are the same except the character at the same position. If the operation is deletion from w_i and substitution from w , we have $ed(w_i, w) = 1$ because they will be the same after at most one substitution from w_i . The other cases can be analyzed in a similar way and are omitted. Now,

assuming that it holds when $n^* = \gamma$, we need to prove it also holds when $n^* = \gamma + 1$. If $w^* = a^*1 a^*2 \dots a^*n \in S_{w_i, d} \cap S_{w, k}$, where $a^*i = a_j$ or $a^*i = _$. For a wildcard at position t , cancel the underlying operations and revert

it to the original

characters in w_i and w at this position. Assume two new elements w_i^* and w^* are derived from them respectively. Then perform one operation at position t of w_i^* to make the character of w_i at this position be the same with w , which is denoted by w_{iJ} . After this operation, w_i^* will be changed to w^* , which has only k wildcards. Therefore, we have $ed(w_{iJ}, w) \leq \gamma$ from the assumption. We know that $ed(w_{iJ}, w) \leq \gamma$ and $ed(w_{iJ}, w_i) = 1$, based on which we know that $ed(w_i, w) \leq \gamma + 1$. Thus, we can get $ed(w, w_i) \leq n^*$. It renders the contradiction $ed(w, w_i) \leq k$ because $n^* > k$. Therefore, $S_{w_i, d} \cap S_{w, k}$ is empty if $ed(w, w_i) > k$. Theorem 2: The fuzzy keyword search scheme is secure

regarding the search privacy.

Proof: In the wildcard-based scheme, the computation of index and request of the same keyword is identical. Therefore, we only need to prove the index privacy by using reduction. Suppose the searchable encryption scheme fails to achieve the index privacy against the indistinguishability under the chosen keyword attack, which means there exists an algorithm who can get the underlying information of keyword from the index.

Then, we build an algorithm A_J that utilizes A to determine whether some function $f_J(\bullet)$ is a pseudo-random function such that $f_J(\bullet)$ is equal to $f(sk, \bullet)$ or a random function. A_J has an access to an oracle $fr(\cdot)$ that takes as input secret value x and returns $f_J(x)$. Upon receiving any request of the index computation, A_J answers it with request to the oracle $fr(\bullet)$. After making these trapdoor queries, the adversary outputs two challenge keywords w_0^* and w_1^* with the same length and edit distance, which can be relaxed by adding some redundant trapdoors. A_J picks one random $b \in \{0, 1\}$ and sends w_b^* to the challenger. Then, A_J is given a challenge value y , which

is either computed from a pseudo-random function $f(sk, \cdot)$ or a random function. A_J sends y back to \mathcal{C} , who answers with $b_J \in \{0, 1\}$. Suppose \mathcal{C} guesses b correctly with non-negligible probability, which indicates that the value y is not randomly computed. Then, A_J makes a decision that $f_J(\cdot)$ is a pseudo-random function. As a result, based on the assumption

of the indistinguishability of the pseudo-random function from some real random function, at most guesses b correctly with approximate probability $1/2$. Thus, the search privacy is obtained.

VII.CONCLUSION

In this paper, for the first time we formalize and solve the problem of supporting efficient yet privacy-preserving fuzzy search for achieving effective utilization of remotely stored encrypted data in Cloud Computing. We design an advanced technique (i.e., wildcard-based technique) to construct the storage-efficient fuzzy keyword sets by exploiting a significant observation on the similarity metric of edit distance. Based on the constructed fuzzy keyword sets, we further propose an efficient fuzzy keyword search scheme. Through rigorous security analysis, we show that our proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search.

As our ongoing work, we will continue to research on security mechanisms that support: 1) search semantics that takes into consideration conjunction of keywords, sequence of keywords, and even the complex natural language semantics to produce highly relevant search results; and 2) search ranking that sorts the searching results according to the relevance criteria.

REFERENCES

- [1] Google, "Britney spears spelling correction," Referenced online at <http://www.google.com/jobs/britney.html>, June 2009.
- [2] M. Bellare, A. Boldyreva, and A. O'Neill, "Deterministic and efficiently searchable encryption," in Proceedings of Crypto 2007, volume 4622 of LNCS. Springer-Verlag, 2007.

- [3] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of IEEE Symposium on Security and Privacy'00, 2000.
- [4] E.-J. Goh, "Secure indexes," Cryptology ePrint Archive, Report 2003/216, 2003, <http://eprint.iacr.org/>.
- [5] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. of EUROCRYPT'04, 2004.
- [6] B. Waters, D. Balfanz, G. Durfee, and D. Smetters, "Building an encrypted and searchable audit log," in Proc. of 11th Annual Network and Distributed System, 2004.
- [7] Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in Proc. of ACNS'05, 2005.
- [8] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. of ACM CCS'06, 2006.
- [9] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in Proc. of TCC'07, 2007, pp. 535–554.
- [10] F. Bao, R. Deng, X. Ding, and Y. Yang, "Private query on encrypted data in multi-user settings," in Proc. of ISPEC'08, 2008.
- [11] C. Li, J. Lu, and Y. Lu, "Efficient merging and filtering algorithms for approximate string searches," in Proc. of ICDE'08, 2008.
- [12] A. Behm, S. Ji, C. Li, , and J. Lu, "Space-constrained gram-based indexing for efficient approximate string search," in Proc. of ICDE'09.
- [13] S. Ji, G. Li, C. Li, and J. Feng, "Efficient interactive fuzzy keyword search," in Proc. of WWW'09, 2009.
- [14] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. Strauss, and R. N. Wright, "Secure multiparty computation of approximations," in Proc. of ICALP'01.

[15] R. Ostrovsky, "Software protection and simulations on oblivious RAMs," Ph.D dissertation, Massachusetts Institute of Technology, 1992.

[16] V. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones," Problems of Information Transmission, vol. 1, no. 1, pp. 8-17, 1965.

Why Predictive Analytics is key for each one? And this tool is the perfect Use Case for Cloud Computing

K Madan Mohan, Assistant Professor, Dept. of CSE, MLRITM, JNTU Hyderabad. (madan.keturu@gmail.com)

Prof. P. PremChand, Dept. of CSE, Osmania University, Hyderabad. (prof.premchand@gmail.com)

1. Abstract

At the present time, it has become a challenge for many organizations to deal with enormous volume of data and to study the customer behavior, sales trend, and a lot of other factors to evaluate the market in order to operate in an effective way and to produce more income. To achieve the goals, organizations rely on different tools and techniques to get perfect data. Predictive Analytics is a tool which uses different techniques to predict future events to categorize risks and opportunities for organizations. Predictive analytics consists of advanced analytics and decision optimization.

2. Introduction

It provides the simple use of the tools used for analysis as they are easily easily reached by the business analysts. It provides a different move toward other than data mining, by providing earlier analysis, gives more significance to prediction rather than the explanation of data. It transforms the raw data to provide more information. Advanced analytics is studying data from past to project future actions related to specific issues of the organization. It uses statistical, mathematical and many other algorithms which are complex in nature and from this analysis the result is taken as approaching to conclude the actions to achieve optimal results. The actions derived along with the essential information are provided to the system or analysts for accomplishment. It improves decision making by measuring the suspicions which enable practical risk management. By using predictive analytics in operation systems, organizations are able to achieve cost reduction, improvement in process and an enlarge in income.

3. Definition

Predictive analytics is a shape of advanced analytics, which uses techniques like data mining(DM), machine learning(ML), and artificial intelligence(AI) to provide predictions for future events from the patterns found in historical and transactional

data. It incorporates the above techniques with modeling business process, management, and information technology (IT).

4. Understanding Predictive Analytics

Let us take an illustration of a certain organization which wants to know what will be its profit after a few years in the business given the present trends in sales, customer base in different locations, etc. Predictive Analytics[2] will use the variables given and using techniques such as data mining (DM), artificial intelligence (AI) would predict the future profit or any other aspect that the organization is fascinated in.

A. Predictive Analytics and Big Data

Predictive analytics is an enabler of big data[3]: Businesses gather huge amounts of real-time customer data and predictive analytics uses this historical data, combined with customer insight, to predict future events. Predictive analytics enable organizations to use big data (both stored and real-time) to move from a historical view to a forward-looking viewpoint of the customer. For example, stores that use data from trustworthiness programs can examine long-ago buying behavior to predict the coupons or promotions a customer is most to participate in or buy in the future. Predictive analytics could also be functional to customer website browsing behaviors to bring a personalized website experience for the customer.

B. Predictive Analytics Software Vendors for the Enterprise

Predictive analytics and data mining solutions for the enterprise are currently available from a number of companies, including SAS (Predictive Analytics Suite), IBM (IBM SPSS Statistics), and Microsoft (Microsoft Dynamics CRM Analytics Foundation).

5. How does Predictive Analytics make working so easy?

Predictive Analytics is these days used in the field of business analysis for optimizing confrontation in marketing, forecasting to improve operations which efficiently facilitate in dropping risks by using interactive and easy to use the software. It makes

the working of the organizations easier by providing them the forethought to calculate the risks and take decisions to avoid them.

6. Why Predictive Analytics is important?

Organizations are revolving to predictive analytics to help solve difficult problems and reveal new opportunities [4]. Common uses include:



Figure 01: The Important task in Organizations

- Detecting fraud:** Combining multiple [analytics](#) methods can improve pattern detection and prevent criminal behavior.
- Optimizing marketing campaigns:** Predictive analytics are used to determine customer responses or purchases, as well as promote cross-sell opportunities.
- Improving operations:** Many companies use predictive models to expect index and manage resources. Airlines use predictive analytics to set ticket prices.

Hotels try to predict the number of guests for any given night to take full advantage of possession and increase income.

- d. **Reducing Risk:** Credit scores are used to evaluate a buyer's possibility of defaulting for purchases and are a well-known example of predictive analytics.

7. PMML - Predictive Model Markup Language

A standard developed by the Data Mining Group (DMG) to represent predictive analytic models. *Predictive Model Markup Language (PMML)*[5] is supported by leading business intelligence and analytics vendors like IBM, SAS, Micro Strategy, Oracle and SAP. The XML-based PMML enables sharing predictive analytic models between different applications, making it possible, for example, to build a model in one system and then move it to another system to test its performance against a test data set

8. Predictive Analytics: The Perfect Use Case for Cloud Computing

The cloud computing provides the processing and big data support needed for predictive analytics. Predictive analytics matching current datasets against historical patterns to determine the probability of an event occurring in the future requires a lot of compute power and draws on a lot of data. In other words, a perfect use case[1] for cloud. Cloud computing is elevating the art and science of predictive analytics to a whole new level. 43% have already developed predictive analytics solutions within their companies, and 82% have predictive analytics in their plans going forward. "Separately, predictive analytics and cloud solutions are changing the way organizations do business,". They open up a wealth of opportunities. Other industry research confirms the growing level of interest in moving business intelligence to the cloud. While predictive analytics has a range of applications, from fraud detection to production system management.

9. Predictive analytics in the cloud

Decision Management Solutions (DMS) recently conducted research into predictive analytics[6] in the cloud. Sponsored by FICO, Lityx and SAP, the research has at its core

a survey of more than 350 respondents from a wide range of industries. Following up on a 2011 survey, the 2013 results make it clear that predictive analytics in the cloud is becoming increasingly mainstream, with broader and accelerating adoption.

The most striking result is that the number of companies reporting a positive impact from predictive analytics has risen dramatically since 2011. More than two-thirds of this year's respondents have seen a positive impact from using predictive analytics in their business. It is also conspicuous how much greater the reported impact is in 2013 relative to 2011. In 2013, many more companies reported transformative or significant impact than in 2011, while far fewer reported no usage or no plans as shown in Figure 2.

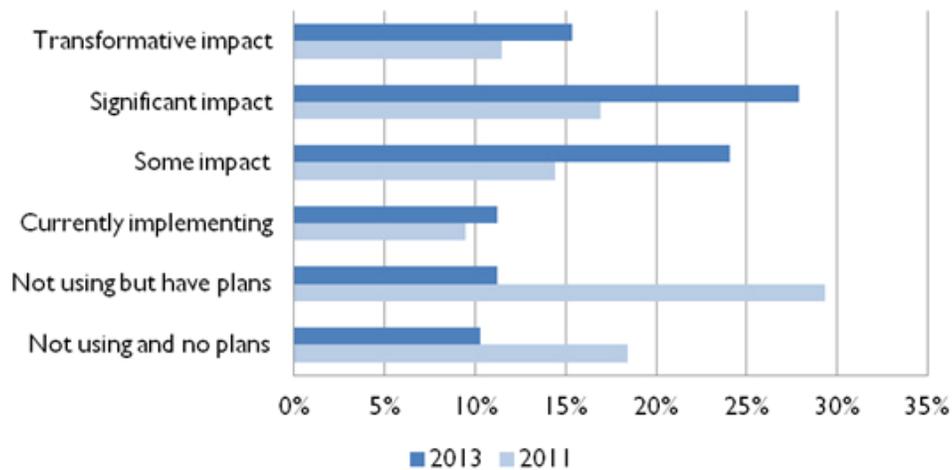


Figure 2: Increasing impact from predictive analytics.

Matching this rise in overall impact from predictive analytics is a similar rise in both current and planned deployment of predictive analytics in the cloud since 2011. The research divided predictive analytics in the cloud into three use cases:

1. Pre-packaged, cloud-based decision-making solutions that embed predictive analytics.

2. Cloud-based predictive modeling – building models in the cloud.

3. Cloud-based deployment of predictive analytics – scoring in the cloud.

These three scenarios leverage the scalability and pervasiveness of the cloud as well as the growing use of the cloud to deliver data. As Figure 3 shows, an amazing 90 percent

said it was likely they would have at least one class of solution widely deployed in the next few years. Predictive analytics in the cloud is going main stream and may, in fact, already be there.

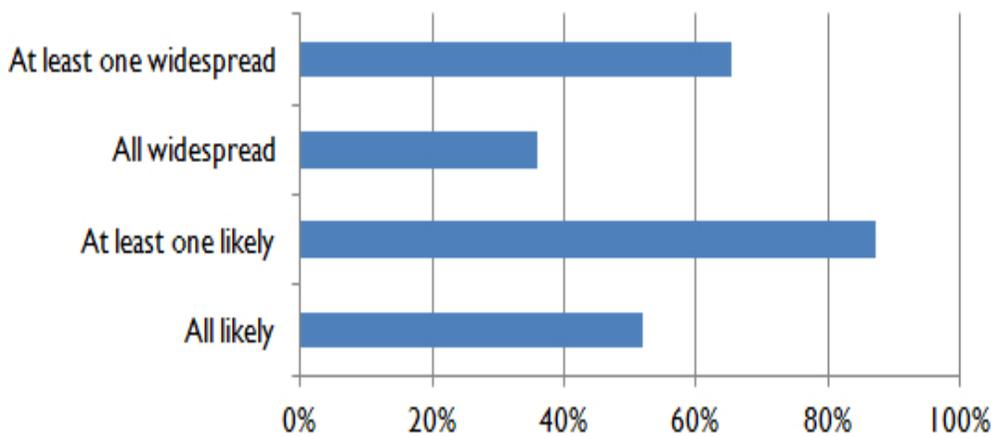


Figure 3: Broad adoption of predictive analytics in the cloud

Social media, sensor, weblog, audio and image data types are all rated as much more important in analytic models among those with successful analytic deployments as shown in Figure 4

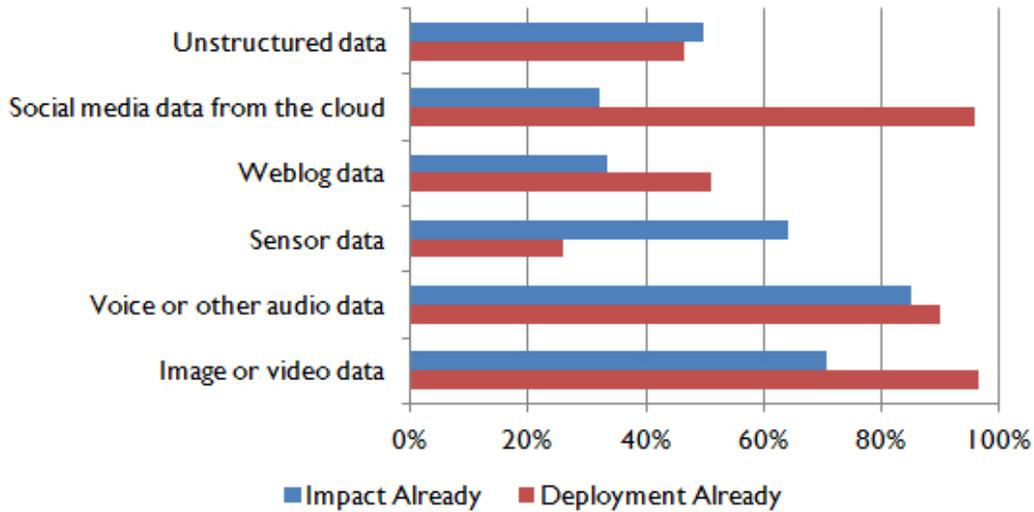


Figure 4: More experienced practitioners use more data types.

The velocity of data also matters. Predictive analytics is increasingly focused on near real-time, operational data. This kind of data grew the most in importance between 2011 and 2013.

Putting predictive analytics to work in operations is strongly correlated with the most impressive results as shown in Figure 5.

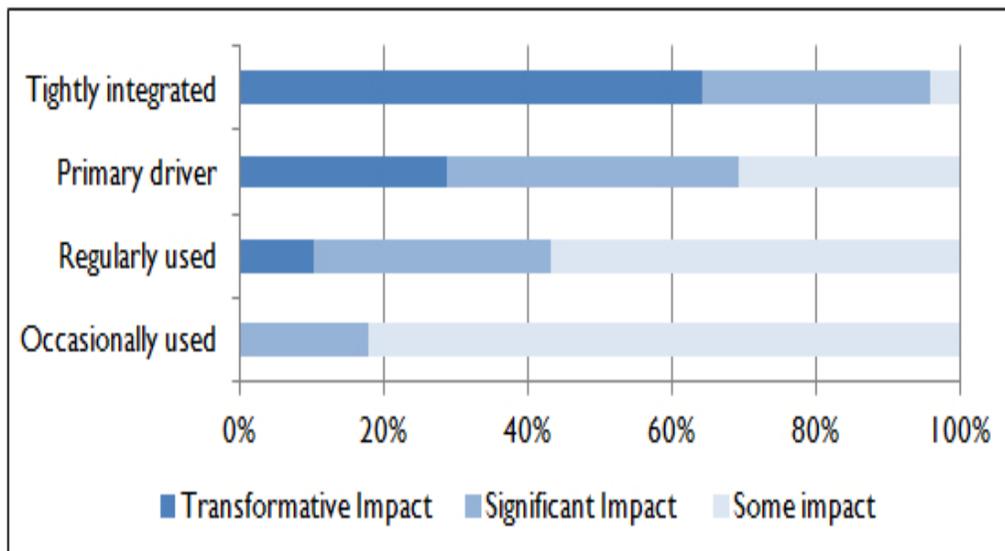


Figure 5: Decision Management transforms results.

While a similar result was found in 2011, the percentage reporting the use of this approach has risen significantly since 2011 as shown in Figure 6.

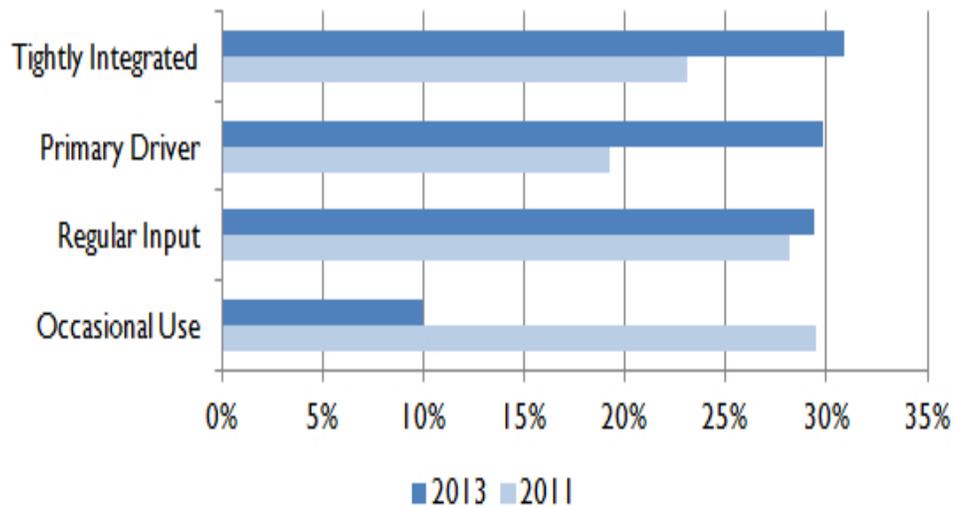


Figure 6: Decision Management on the rise.

10. Predictive Analytics Tools

A wide range of tools is used in predictive modeling and analytics. IBM, Microsoft, SAS Institute and many other software vendors offer predictive analytics tools[7] and related technologies supporting machine learning and deep learning applications.

Five leading customer analytics tools	
PLATFORM	FEATURES
Adobe Analytics	<ul style="list-style-type: none"> ■ Platform for standardized video and ad engagement ■ Predictive workbench through machine learning ■ Live stream with real-time events ■ Data workbench ■ Mobile app campaign tool
Google Analytics 360	<ul style="list-style-type: none"> ■ Data studio ■ Surveys ■ Value of marketing channels through attribution ■ Audience center to match customers with the right message ■ Analytics ■ Tag Manager allowing tag updates without editing code
IBM Watson Customer Experience Analytics	<ul style="list-style-type: none"> ■ Role-based dashboards ■ Journey analytics ■ Mindset analysis ■ Event alerts ■ Site optimization
SAP Hybris Marketing Cloud	<ul style="list-style-type: none"> ■ Gains customer insights ■ Categorizes customers based on score values ■ Builds customer target groups ■ Triggers successful campaigns
SAS Customer Intelligence 360	<ul style="list-style-type: none"> ■ Utilizes customer-centric data model ■ Records activities of all visitors ■ Contextualizes captured data ■ Better business goal creation ■ Digital asset management

©2017 TECHTARGET. ALL RIGHTS RESERVED. 

11. WHAT ARE THE BIGGEST CHALLENGES FACING PREDICTIVE ANALYTICS TODAY?

Predictive analytics is the use of data, statistical algorithms and machine learning[8] techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what *has* happened to provide the best assessment of what *will* happen in the future.

DATA

Every company has developed its method for managing data in their databases. Some of the common data challenges are:

1. Data preparation
2. Data cleansing, data wrangling
3. Ensuring the data format is correct
4. Identifying important variables
5. Recognizing correlations
6. Dealing with imbalanced data
7. Understanding how different algorithms work
8. Choosing the right algorithm for the right problem
9. Deciding the right configurations the algorithm
10. Understanding the output of the algorithm
11. Re-training the algorithm with new data
12. Deploying/re-deploying the model
13. Predicting in real time/batch
14. Integrating with your primary application to build data insights into the application and initiate user action.

12.CONCLUSION

Predictive Analytics presents daunting challenges to data scientists. The business benefits are enormous, and failure is not an option. New, emerging automated AI services can decrease some of the burden and enable a faster and easier time to market.

13. REFERENCES

- [1].<https://www.forbes.com/sites/joemckendrick/2012/06/18/predictive-analytics-the-perfect-use-case-for-cloud-computing/#6db15a372a53>
- [2]. https://www.sas.com/en_in/insights/analytics/predictive-analytics.html
- [3] <https://www.pcmag.com/article/345858/predictive-analytics-big-data-and-how-to-make-them-work-fo>
- [4] <https://www.edupristine.com/blog/importance-of-predictive-analytics>
- [5] <https://sourceforge.net/projects/pmml/>
- [6] <http://analytics-magazine.org/predictive-analytics-in-the-cloud/>
- [7] <https://www.educba.com/predictive-analytics-tool/>
- [8] <https://www.quora.com/What-are-the-biggest-problems-facing-predictive-analytics-today>

A Survey on Multimedia Information Retrieval Based on Annotation

H.R. Chennamma
JSS Science and Technology University,
Mysuru, INDIA
Email: anuamruthesh@gmail.com

Abstract— Multimedia information retrieval is an important research discipline of computer science that aims at extracting semantic information from multimedia data sources. Data sources include text, audio, image, video etc. There is a rapid growth in the amount of multimedia data from real world multimedia sharing websites, such as flicker and YouTube etc, mining of organized data which are scattered in gigantic unstructured information is a challenging task. In this paper we exhibit a survey of different techniques of annotation for information retrieval from multimedia databases include documents, images, audio files and videos. This high level survey addresses researchers that are new to the field and require a proper overview about the annotation and retrieval possibilities from multimedia databases.

Keywords: Document Annotation, Image Annotation, Video Annotation, Audio Annotation

I. Introduction

In this electronic world internet plays a major role. It takes your time and gives knowledge and information. Internet provides us knowledge and information in the form of multimedia content includes text, audio, image and videos. With every increasing minute the content over the internet is also increasing. Since its initiation, the World Wide Web has been overwhelmed by unstructured data. Retrieving an organized data from vast accumulation of unstructured data is a dreary employment [1]. In recent years, tagging is the act of adding keywords (or tags) to objects has become a popular means to annotate various web resources. The tags provide meaningful descriptors of the objects and allow the user to organize and index their content. This becomes even more important, when dealing with multimedia annotation. Fig. 1 shows the classifications of such multimedia annotation. Annotation can be conceptualized to assist in the field of information retrieval and decision making. Different annotations for the same file may be expected from different user (multiuser cannot make the same type of annotations on the same documents).

II. DOCUMENT ANNOTATION

There are numerous application spaces where clients make and offer data; for occasion, news blogs, scientific networks, interpersonal interaction gathering and so on. Such collection of textual data contains significant amount of structured information, which remains in the unstructured format. Ruiz, Hristidis and Ipeirotis introduced a framework in which annotation process utilizes the query but not content of the document.

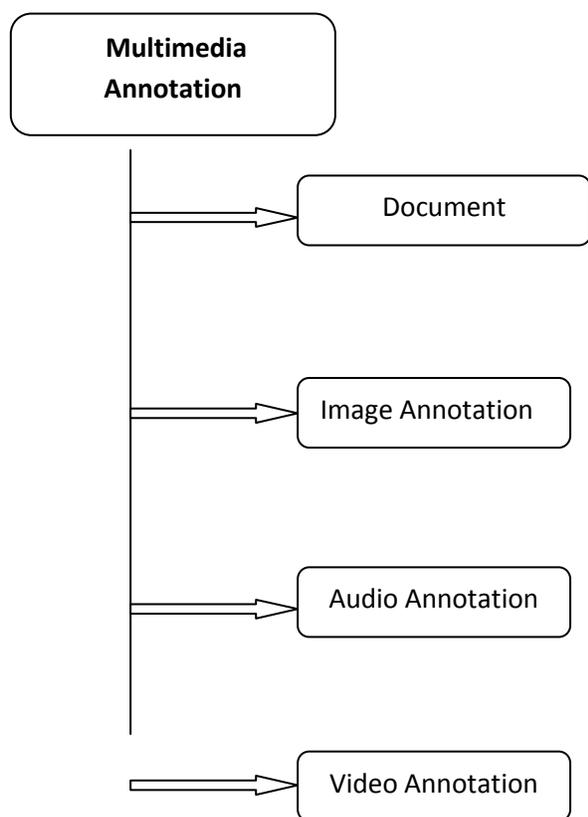


Fig.1 Classification of multimedia annotation

In other words, they are attempting to prioritize the annotation of documents towards generating attribute values that are regularly utilized by questioning users. In order to annotate a document, while trying to satisfy the user querying needs, the proposed solution is based on a probabilistic framework that considers the evidence in the document content and the query workload. The resultant two pieces of evidence are content value and querying value. Two models were used to combine these two pieces of evidences: a model that considers both components conditionally independent and a linear weighted model.

An alternate methodology for grouping documents in view of a semantic closeness and the viable representation of the content of the documents is introduced in [3]. The content of a document is annotated and the resulting annotation is represented by a labeled tree whose nodes and edges are represented by concepts lying within domain ontology. A reasoning procedure may be done on annotation trees, permitting the correlation of records between every other, for classification or

information retrieval purposes. Furthermore they proposed an approach for characterizing a semantic annotation of Web text documents based on the content of the documents. The core of the methodology depends on the notions of annotation tree and semantic similarity, allowing manipulation of documents with respect to their content, for, e.g. reasoning and information retrieval.

Noll and Meinel have studied the characteristics of social annotations and tagging with regard to their usefulness for web document classification by an analysis of large sets of real-world data. They also found out, which kinds of documents are annotated more by end users than others, how users tend to annotate these documents, and in particular how this user-generated folk-sonomy compares with a top-down taxonomy maintained by classification experts for the same set of documents.

Archana shukla had developed a tool, which tells the quality of document or its usefulness based on annotations. Annotation may include comments, notes, observation, highlights, underline, explanation, question or help etc. comments are used for evaluation purpose while others are used for summarization or for expansion also. Further these comments may be on another annotation. Such annotations are referred as Meta annotation. All annotation may not get equal weightage. This tool considered highlights, underline as well as comments to infer the collective sentiment of annotators. Collective sentiments of annotators are classified as positive, negative, objectivity. This tool computes collective sentiment of annotations in two manners. It counts all the annotation present on the documents as well as it also computes sentiment scores of all annotation which includes comments to obtain the collective sentiments about the document or to judge the quality of document. They have developed a tool named KMAD (Metadata Extractor for Annotated Document) to annotate a PDF document. And have designed an annotation schema using DTD (Document Type Definition) to capture the information of annotation. The relationship between the annotations is complex. The author had only considered those annotations or Meta annotation which is of type comment, highlights, and underline.

Since paper documents have high legibility and portability, they are widely used as media for communication. Furthermore, it is common for us to mark up or annotate paper documents. Therefore annotations on paper documents have important information such as users' interests or knowledge. Such valuable information can be obtained by extracting and analyzing annotations on paper documents. Annotations on paper documents include important information. One can exploit the information by extracting and analyzing annotations. A method of annotation extraction from paper documents is proposed in [6]. In this method, annotations are extracted by subtracting original document images from aligned annotated document images under the assumption that original electronic documents are available. The proposed method is characterized by the fast alignment process based on the point matching method and the flexible subtraction process. Experimental results have shown that color annotations can be extracted from color documents.

III. Image annotation

Due to the easy availability and low cost of high resolution digital cameras digital image archives increasing rapidly. For effective search of image, there is an urgent need of successful

indexing and searching image retrieval system. In the text based image retrieval, the images are annotated and the data base management system retrieves them in the same way as text documents. However, it is very difficult and time consuming task due to the need of human intervention. To address these drawbacks, content based image retrieval using low level image features such as colour, texture, shape were proposed. To search for images relevant to a query some image processing techniques are used for feature extraction. Unfortunately the naive user may not be familiar with low level visual features and it is difficult to specify their query concepts by using the low level visual features directly because users are familiar with natural language like queries such as text and typically query images by semantics. Online photo services such as Flickr and Zoomr allow users to share their photos with family, friends, and the online community at large. An important facet of these services is that users manually annotate their photos using so called tags, which describe the contents of the photo or provide additional contextual and semantic information.

Sigurbjornsson and Zwol [7] Investigated how to assist users in the tagging phase. They analyze a representative snapshot of Flickr and present the results by means of a tag characterization focusing on how users tag photos and what information is contained in the tagging. Based on this analysis, they present and evaluate tag recommendation strategies to support the user in the photo annotation task by recommending a set of tags that can be added to the photo. The results of the empirical evaluation show that one can effectively recommend relevant tags for a variety of photos with different levels of exhaustiveness of original tagging. This method of tagging is complementary to the approaches like content based methods [20, 21] or the spatial, temporal and social content of the users [22, 23]. The authors recommend that a combination of different complimentary methods is likely to give a more robust performance.

According to [8], a large collection of images with ground truth labels to be used for object detection and recognition. Such data is useful for supervised learning and quantitative evaluation. To achieve this, they developed a web-based tool that allows easy image annotation and instant sharing of such annotations. Using this annotation tool, they have collected a large dataset that spans many object categories, often containing multiple instances over a wide variety of images. They quantify the contents of the dataset and compare against existing state of the art datasets used for object recognition and detection. Also, they show how to extend the dataset to automatically enhance object labels with WordNet, discover object parts, recover a depth ordering of objects in a scene and increase the number of labels using minimal user supervision. Here they have described LabelMe, a database and an online annotation tool that allows the sharing of images and annotations. The online tool provides functionalities such as drawing polygons, querying images, and browsing the database. The goal of the annotation tool

is to provide a drawing interface that works on many platforms, is easy to use, and allows instant sharing of the collected data. The annotation tool design choice emphasizes simplicity and ease of use. Authors described a web-based image annotation tool that was used to label the identity of objects and where they occur in images. They collected a large number of high quality annotations, spanning many different object categories, for a large set of images, many of which are high resolution.

Automatic image annotation is a challenging problem in the field of image retrieval. It can be used to facilitate semantic search in large image databases. However, retrieval performance of the existing annotation schemes is far from the user's expectation. Image annotation task consists of assigning a set of semantic tags or labels to a novel image based on some models learned from certain training data. Automatically assigning keywords to images is of great interest as it allows one to index, retrieve, and understand large collections of image data. Many techniques have been proposed for image annotation in the last decade that gives reasonable performance on standard datasets. In [9] authors proposed algorithm to annotate image by comparing test image feature vector with feature matrix of training data sets and similar/dissimilar image pairs. Ranking technique is used for transferring the key word from similar image pair to the test image by counting local frequency of keywords. It include training and testing procedures training part selects low-level features (e.g., bins in the feature histogram) using saliency detection technique. Testing a part automatically annotates input images by transferring keywords from similar images. They have proposed a framework and algorithm for automatic image annotation problem. They have considered a holistic approach and saliency detection technique, and then compared the obtained results with other studies in the literature.

Cao and Luo addressed the problem of annotating photo collections instead of considering a single photo at a time or label photos individually [10]. They built a sizable collection of geo tagged personal photos and defined a compact ontology of events and scenes suitable for consumers. They constructed a Conditional Random Field [CRF] based models that accounts for two types of correlations: (i) Correlation by time and GPS (Global Positioning System) tags and (ii) Correlation between scene and event level labels. In this way multi level annotation hierarchy has been used to annotate consumer photo collections. In order to provide descriptive annotation for personal photos, two level annotations were introduced. Personal photos were taken in different places and at different times, describing different activities of different people. Indeed, these diverse factors make photo annotation a challenging task. In the upper level, photos were clustered into groups and assign an event label to each group to denote the main activity common to all the photos in that group. The event label can either be

one and only one of the pre-defined event classes or "NULL". In the lower level, each photo will be assigned by one or more scene class labels. To explore the correlation between scene labels, camera metadata is used. Generally camera metadata includes timestamp and GPS tags. Every digital photo file records the date and the time when the photo was taken. An advance camera can even record the location via GPS receiver. However due to the sensitivity limitation of the GPS receiver, GPS tags can be missing. This paper discusses how to make good use of such incomplete metadata information.

IV. Audio Annotation

Music is a form of communication that can represent human emotions, personal style, geographic origins, spiritual foundations, social conditions, and other aspects of humanity. Listeners naturally use words in an attempt to describe what they hear even though two listeners may use drastically different words when describing the same piece of music. However, words related to some aspects of the audio content, such as instrumentation and genre, may be largely agreed upon by a majority of listeners. Audio can create the illusion of reality and immersion and is fundamental component of the artistic creation. Sounds are also needed in other audiovisual productions, such as in computer games or web pages. Sometimes sounds are recorded for the occasion. Many times, sound engineers access already compiled sound effects libraries.

Tzanetakis and Cook focuses on the task of annotating audio data and especially music. An example would be structuring hours of archived radio broadcasts for audio information retrieval. Annotation of simple cases like musical instruments or music vs. speech can be performed automatically using current classification systems. Advances in digital storage technology and the wide use of digital audio compression standards like MPEG have made possible the creation of large archives of audio material. In order to work efficiently with these large archives much more structured than what is currently available is needed. One way to approach the problem is through the creation of text indices and the use of traditional information retrieval techniques [24] familiar from the popular Web search engines. They have described a prototype audio browsing tool that was used to perform user experiments in semi-automatic audio segmentation annotation.

A human computation tool that is capable of gathering perceptually meaningful descriptions for audio data that are agreed upon by multiple players is presented in [12]. Such data are useful for several purposes. Large audio databases have become invaluable resources for listeners, sound designers and composers. Current audio retrieval systems are primarily text-based, relying on accurate and comprehensive annotation of the data. However, keywords that

describe a particular audio file are often subjective, based on one person's opinion [25]. TagATune has the potential to produce better labels at lower cost because the labor is essentially free and the validity of each label is confirmed by many players. The cost of obtaining a comprehensive set of annotations manually is high. One way to lower the cost of labeling is to create games with a purpose that people will voluntarily play, producing useful metadata as a by-product. TagATune is an audio-based online game that aims to extract descriptions of sounds and music from human players. They present the rationale, design and preliminary results from a pilot study using a prototype of TagATune to label a subset of the Free Sound database. The cost of obtaining a comprehensive set of annotations manually is high. One way to lower the cost of labeling is to create games with a purpose that people will voluntarily play, producing useful metadata as a by-product. TagATune is an audio-based online game that aims to extract descriptions of sounds and music from human players. They present the rationale, design and preliminary results from a pilot study using a prototype of TagATune to label a subset of the Free Sound database.

A central goal of the music information retrieval community is to create systems that efficiently store and retrieve songs from large databases of musical content. The most common way to store and retrieve music is that using metadata such as the name of the composer or artist, the name of the song, or the release date of the album.

Turnbull, Barrington, Torres and Lanckriet presented a computer audition system that can both annotate novel audio tracks with semantically meaningful words and retrieve relevant tracks from a database of unlabeled audio content given a text-based query. They consider the related tasks of content-based audio annotation and retrieval as one supervised multiclass, multi label problem in which the joint probability of acoustic features and words were needed. Authors have collected a data set of 1700 human-generated annotations that describe 500 Western popular music tracks. For each word in a vocabulary, they use this data to train a Gaussian mixture model (GMM) over an audio feature space and estimate the parameters of the model using the weighted mixture hierarchy's expectation maximization algorithm. This algorithm is more scalable to large data sets and produces better density estimates than standard parameter estimation techniques. The quality of the music annotations produced by this system is comparable with the performance of humans on the same task. This "query-by-text" system can retrieve appropriate songs for a large number of musically relevant words are general by learning a model that can annotate and retrieve sound effects. Data collection process can be extended in two ways. The first involves vocabulary selection: if a word in the vocabulary is inconsistently used by human annotators, or the word is not clearly represented by the underlying acoustic representation then the word can be considered as noisy and should be

removed from the vocabulary to denoise the modeling process. The second extension involves collecting a much larger annotated data set of music using web-based human computation games. A web-based game called "Listen Game" was developed which allows multiple "annotators" to label music through real time competition. This helps to grow vocabulary by allowing users to suggest words that describe the music.

Sound engineers need to access vast collections of sound effects for their film and video productions. Sound effects providers rely on text-retrieval techniques to offer their collections. Currently, annotation of audio content is done manually, which is an arduous task. Automatic annotation methods, normally fine-tuned to reduced domains such as musical instruments or reduced sound effects taxonomies, are not mature enough for labeling with great detail any possible sound. A general sound recognition tool would require: first, a taxonomy that represents the world and, second thousands of classifiers, each specialized in distinguishing little details. [14] Presented experiments on an all-purpose sound recognition system based on nearest-neighbor classification rule. To tackle the taxonomy definition problem they use Word-Net, a semantic network that organizes real world knowledge. In order to overcome the need of a huge number of classifiers to distinguish many different sound classes, they use a nearest-neighbor classifier with a database of isolated sounds unambiguously linked to Word-Net concepts

An amount of online music grows, automatic music recommendation becomes an increasingly important tool for music listeners to find music that they would like. In the domain of music, Web sites such as Last.fm use social tags as a basis for recommending music to listeners. A method for predicting social tags using audio feature extraction and supervised learning is proposed in [15]. These automatically-generated tags (or "auto tags") can furnish information about music for which good, descriptive social tags are lacking. They propose a method for predicting these social tags directly from MP3 files. Using a set of boosted classifiers, audio features are mapped onto social tags collected from the Web. The resulting automatic tags (or auto tags) furnish information about music that is otherwise untagged or poorly tagged, allowing for insertion of previously unheard music into a social recommender. Auto tags can also be used to smooth the tag space from which similarities and recommendations are made by providing a set of comparable baseline tags for all tracks in a recommender system.

V. *Video annotation*

The classical view of video documents is characterized by two well-defined role models for video content generation and usage. On one hand, users consume, retrieve, and share specific videos. Metadata plays a central role in this context to shorten the bridge between users and

videos as it facilitates efficient retrieval on the other hand producers generate videos and annotate them with metadata using specific tools. Automatic annotation techniques are mainly focused on the generation of additional, richer metadata that is suitable for text-based video retrieval. Such techniques stem from several research directions and they operate on different modalities [16]

Lavrenka, Feng and Manmatha have applied a continuous relevance model (CRM) to the problem of directly retrieving the visual content of videos using text queries[17]. The CRM computes a joint probability model for image features and words using training set of annotated images. The model may then be used to annotate unseen test images. The probabilistic annotations are used for retrieval using text queries. The CRM is a statistical model for automatically assigning keywords to unlabeled images. The model relies on a training set of annotated images and operates as follows. First partitioning each training image into regions is done (either using an unsupervised segmentation algorithm or by partitioning the image into rectangular regions). Then computing a real-valued feature vector for each region is done. The features may include shape, color and texture of a region.

An interactive technique for visually annotating independently moving objects in a video stream is introduced in [18]. Features in the video are automatically tracked and grouped in an off-line preprocess that enables later interactive manipulation and annotation. Examples of such annotations include speech and thought balloons, video graffiti, hyperlinks, and path arrows. The system also employs a direct manipulation interface for random frame access using spatial constraints. This annotation interface can be employed in a variety of applications including surveillance, film and video editing, visual tagging, and authoring rich media such as hyperlinked video. A System is described that makes easy to author annotations that transform along with objects or regions in an arbitrary video. This system first analyzes the video in a fully automatic pre-processing step that tracks the motion of image points across the video and segments those tracks into coherently moving groups. These groups and the motion of the tracked points are then used to drive an interactive annotation interface. These annotations are known as “video object annotations” (VOAs) because they are associated with specific objects or regions of the video.

Khurana and Chandak introduced a novel approach for video annotation[19]. The key frames are extracted from the video and are analyzed. Instead of complete video frames, only the key frames are analyzed to identify the objects present. The object detectors are trained for identification of the object. The detected objects are then added in the annotation file. The annotation is based on ontology which eases the semantic retrieval of videos. Ontology based

video annotation greatly accelerates the performance of retrieval systems. The approach presented by Khurana and Chandak for the annotation of video which is based on ontology. This author system consists of extraction of frames that contain most significant visual information. Hence the video is reduced to less number of images known as key frames. The classifiers for different objects are trained using the features for the respective objects. They have used scale invariant feature transform (SIFT) as the features for training and testing the classifiers. When the features of key frames are given as input to different classifiers; they can detect whether the object is present in the frame or not. If the object is present then the frame names are given to the annotation module which will create an xml file corresponding to each key frame. The xml file contains the object / objects present in the image (video frame) along with object ontology. So this paper mainly focuses on the video key frame annotation which is performed using the SIFT features and SVM as the classifier. Furthermore, the annotation file created for all the frames are based on ontology which can greatly assist retrieval and browsing. The possible extension and application areas of the system and video annotation are boundless; with innovation in the related domain, video access time and relevance can be greatly enhanced.

VI. Conclusion

This paper gives a high level perspective of current multimedia information retrieval based on annotations. Our survey contributes to multimedia research with a proper overview of the field. An overview of different techniques of annotation for multimedia data sources (includes documents, images, audio files and videos) are presented. The current state of the art leaves a lot of space for further advances and automatic annotation in multimedia information retrieval.

REFERENCES

- [1] S. J. Kadam, S. Bajpai and P. M. Yelmar, "An Investigative Survey of Annotation Types and Systems" Proceedings of the International Conference on Advances in Engineering and Technology ICAET, 2014
- [2] Eduardo J. Ruiz , Vagelis Hristidis , Panagiotis G. Ipeirotis, "Facilitating Document Annotation using Content and Querying Value" IEEE Transactions on knowledge and data engineering vol.pp no.99 year 2013.
- [3] Emmanuel Nauer, Amedeo Napoli, "A proposal for annotation, semantic similarity and classification of textual documents" The 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications - AIMS 2006. Pp..201-212.
- [4] Michael G. Noll, Christoph Meinel, " Exploring Social Annotations for Web Documents Classification" The 23rd International ACM Symposium on Applied Computing Fortaleza, Ceara, Brazil, March 2008, pp.2315-2320.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- [5] Archana Shukla, "Sentiment Analysis of Document Based On Annotation", International Journal of Web & Semantic Technology; Oct2011, Vol. 2 Issue 4, pp. 91.
- [6] Tomohiro Nakai, Koichi Kise, Masakazu Iwamura, "A Method of Annotation Extraction from Paper Documents Using Alignment Based on Local Arrangements of Feature Points", 12th International Conference on Document Analysis and Recognition 2007, pp. 23-27.
- [7] B.Sigurbjornsson and R. van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 327-336, 2008.
- [8] B. Russel, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," Int'l J. Computer Vision, vol.77, pp. 157-173, 2008.
- [9] S.K.K. Kharkate and N. J. Janwe, "A Novel Approach For Automatic Image Annotation Using Color Saliency" IJIRCCE, vol. 1, Issue 5, July 2013
- [10] L. Cao and j. Luo "Image Annotation Within the Context of Personal Photo Collection Using Hierarchical Event and Scene Model," IEEE transaction on Multimedia vol. 11, No. 2, pp. 208-219, February 2009.
- [11]G. Tzanetakis and Perry Cook, "Experiments in computer-assisted annotation of audio" In Proc. Int. Conf. Auditory Display (ICAD), Atlanta, Georgia, 2000
- [12] [Edith L. M. Law](#), [Luis von Ahn](#), [Roger B. Dannenberg](#), and [Mike Crawford](#) "TagATune: A Game for Music and Sound Annotation" International Society for Music Information Retrieval (ISMIR), pp. 361-364, Austrian Computer Society, 2007.
- [13]D. Turnbull, L Barrington, D torres and G Lanckriet "Semantic Annotation and Retrieval of Music and Sound Effects," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, No. 2, pp.467-476, February 2008.
- [14]P. Cano and M. Koppenberger, "Automatic Sound Annotation," IEEE workshop on Machine Learning for Signal Processing (2004), pp. 391-400.
- [15]D. Eck, P. Lamere, T Bertin-Mahieux and S. Green, "Automatic Generation of Social Tags for Music Recommendation," In Advances in Neural Information Processing Systems, Vol. 20 (2007).
- [16]R. Sorschag, "A High-Level Survey Of Video Annotation and Retrieval Systems" International Journal of Multimedia Technology, vol. 2 Iss. 3, pp. 62-71, Sep. 2012.
- [17][Lavrenko, V](#) ,[Feng, S.L.](#) ; [Manmatha, R.](#)" " IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.1044-7 vol.3, May 2004.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- [18] D. B Goldman, B Curless, D. Salesin and S. M. Seitz, "Video object annotation, navigation, and composition," in Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology , Oct. 2008, pp. 3–12.
- [19]K. Khurana and M.B. Chandak " Video Annotation Methodology Based on Ontology for transportation Domain" IJARCSSE, vol 3, Issue 6, pp. 540-548, June 2013
- [20]K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. Journal of Machine Learning Research, 2003.
- [21]J. Li and J. Z. Wang. Real-time computerized annotation of pictures. In Proceedings of the ACM Multimedia Conference, pp. 911-920, 2006.
- [22]S. Ahern, S. King, M. Naaman, R. Nair, and J. H.-I. Yang. ZoneTag: Rich, community-supported context-aware media capture and annotation. In Mobile Spatial Interaction workshop (MSI) at the SIGCHI conference on Human Factors in computing systems (CHI 2007), 2007.
- [23]Zonetag. <http://zonetag.research.yahoo.com/>.
- [24]C. van Rijsbergen, Information retrieval, Butterworths, London, 2nd edition, 1979.
- [25]P. Cano and M. Koppenberger. Automatic sound annotation.IEEE Workshop on Machine Learning for Signal Processing, pages 391–400, 2004.

Malware Detection And Control In Decentralized Peer To Peer Network

S.Ganeshmoorthy¹, Abhijith. P², Ashar Henson.H³

1. Assistant Professor, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore.
2. Final year-BCA, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore
3. Final year-BCA, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore

Abstract

In this paper, we formulate an analytical model to characterize the spread of malware in decentralized, Gnutella type peer-to-peer (P2P) networks and study the dynamics associated with the spread of malware. Using a compartmental model, we derive the system parameters or network conditions under which the P2P network may reach a malware free equilibrium. In this paper, we present a cryptographic protocol for ensuring secure and timely availability of the reputation data of a peer to other peers at extremely low costs. The past behavior of the peer is encapsulated in its digital reputation, and is subsequently used to predict its future actions. As a result, a peer's reputation motivates it to cooperate and desist from malicious activities. The cryptographic protocol is coupled with self-certification and cryptographic mechanisms for identity management and countering Sybil attack. I illustrate the security and the efficiency of the system analytically and by means of simulations in a completely decentralized Gnutella-like P2P network. The model also evaluates the effect of control strategies like node quarantine on stifling the spread of malware. The model is then extended to consider the impact of P2P networks on the malware spread in networks of smart cell phones.

KEY TERMS-Peer to Peer network, Malware, Attack

I. INTRODUCTION

The use of peer-to-peer (P2P) networks as a vehicle to spread malware offers some important advantages over worms that spread by scanning for vulnerable hosts. This is primarily due to the methodology employed by the peers to search for content. For instance, in decentralized P2P architectures such as Gnutella where search is done by flooding the network, a peer forwards the query to its immediate neighbors and the process is repeated until a specified threshold time-to-live, TTL, is reached. Here TTL is the threshold representing the

number of overlay links that a search query travels. A relevant example here is the Mandragore worm that affected Gnutella users. Having infected a host in the network, the worm cloaks itself for other Gnutella users.

Every time a Gnutella user searches for media files in the infected computer, the virus always appears as an answer to the request, leading the user to believe that it is the file the user searched for. The design of the search technique has the following implications: first, the worms can spread much faster, since they do not have to probe for susceptible hosts and second, the rate of failed connections is less. Thus, rapid proliferation of malware can pose a serious security threat to the functioning of P2P networks. Understanding the factors affecting the malware spread can help facilitate network designs that are resilient to attacks, ensuring protection of the networking infrastructure.

This paper addresses this issue and develops an analytic framework for modeling the spread of malware in P2P networks while accounting for the architectural, topological, and user related factors. We also model the impact of malware control strategies like node quarantine. Though the initial thrust in P2P research was measurement oriented, subsequent works have proposed analytical models for the temporal evolution of information in the network. The focus of these works is on transfer of regular files and they do not apply to malware that spread actively. In addition, they are specialized to BitTorrent like networks and cannot be extended for P2P networks such as Gnutella or KaZaa.

The issue of worms in peer-to-peer networks is addressed in using a simulation study of P2P worms and possible mitigation mechanisms. Epidemiological models to study malware spread in P2P networks. These studies assume that a vulnerable peer can be infected by any of the infected peers in the network. This assumption is invalid since the candidates for infecting a peer are limited to those within TTL hops away from it and not the entire network. Another important omission is the incorporation of user behavior. Typically, users in a P2P network alternate between two states: the on state, where they are connected to other peers and partake in network activities and the off state wherein they are disconnected from the network. Peers going offline result in fewer candidates for infection thereby lowering the intensity of malware

spread. An empirical model for malware spreading in Bit Torrent is developed in while models for the number of infected nodes by dynamic hit list-based malware in Bit Torrent networks.

However, these models ignore node dynamics such as online-offline transitions and are applicable only to Bit Torrent networks. In the authors use hyper cubes as the graph model for P2P networks and derive a limiting condition on the spectral radius of the adjacency graph, for a virus/worm to be prevalent in the network. The models do not account for the fact that once a peer is infected, any susceptible peer within a TTL hop radius becomes a likely candidate for a virus attack. In the current work, we formulate a comprehensive model for malware spread in Gnutella type P2P networks that addresses the above shortcomings. We develop the model in two stages: first, we quantify the average number of peers within TTL hops from any given peer and in the second stage incorporate the neighborhood information into the final model for malware spread.

System Study

On the social status of the project participants must be assessed to ensure compatibility. It is common Knowledge the computer installations have something to do understandable that the introduction of a candidate system requires special effort to educate, sell and train the Staff on new ways of considering business.

- Manual system is easy
- Training required
- Ensures security and confidentially
- No packet loss

Information will be more accurate than performed manually

II. LITERATURE STUDY

Social networking and peer-to-peer sites, web applications and mobile platforms makes today's users highly vulnerable to entirely new generations of malware that exploit vulnerabilities in web applications and mobile platforms for new infections, while using the power-law connectivity for finding new victims.

The traditional epidemic models based on assumptions of homogeneity, averagedegree distributions, and perfect-mixing are inadequate to model this type of malware propagation.

THE use of peer-to-peer (P2P) networks as a vehicle to spread malware offers some important advantages over worms that spread by scanning for vulnerable hosts.

This is primarily due to the methodology employed by the peers to search for content. The design of the search technique has the following implications: first, the worms can spread much faster, since they do not have to probe for susceptible hosts and second, the rate of failed connections is less. Thus, rapid proliferation of malware can pose a serious security threat to the functioning of P2P networks

Drawbacks:

- A specialist network operating system is needed
- The server is expensive to purchase
- Specialist staff such as a network manager is needed
- If any part of the network fails a lot of disruption can occur

III. Development of malware detection and control in decentralized peer to peer network

In this paper addresses this issue and develops an analytic framework for modeling the spread of malware in P2P networks while accounting for the architectural, topological, and user related factors. We also model the impact of malware control strategies like node quarantine.

We have proposed analytical models for the temporal evolution of information in the network. The focus of these works is on transfer of regular files and they do not apply to malware that spread actively. In addition, they are specialized to BitTorrent like networks and cannot be extended for P2P networks such as Gnutella or KaZaa.

In the authors use hypercubes as the graph model for P2P networks and derive a limiting condition on the spectral radius of the adjacency graph, for a virus/worm to be prevalent in the network. The models do not account for the fact that once a peer is infected, any susceptible peer within a TTL hop radius becomes a likely candidate for a virus attack.

In the current work, we formulate a comprehensive model for malware spread in Gnutella type P2P networks that addresses the above shortcomings. We develop the model in two stages: first, we quantify the average number of peers within TTL hops from any given peer and in the second stage incorporate the neighborhood information into the final model for malware spread.

BENEFITS:

The main advantage of peer to peer network is that it is easier to set up

- The peer to peer network is less expensive.
- In peer-to-peer networks all nodes are act as server as well as client therefore no need of dedicated server.
- It is easier to set up and use this means that you can spend less time in the configuration and implementation of peer to peer network.
- It is not require for the peer to peer network to use the dedicated server computer. Any computer on the network can function as both a network server and a user workstation.
- No need for a network operating system
- No need for specialist staff such as network technicians because each user sets their own permissions as to which files they are willing to share.
- If one computer fails it will not disrupt any other part of the network. It just means that those files aren't available to other users at that time.

Modules:

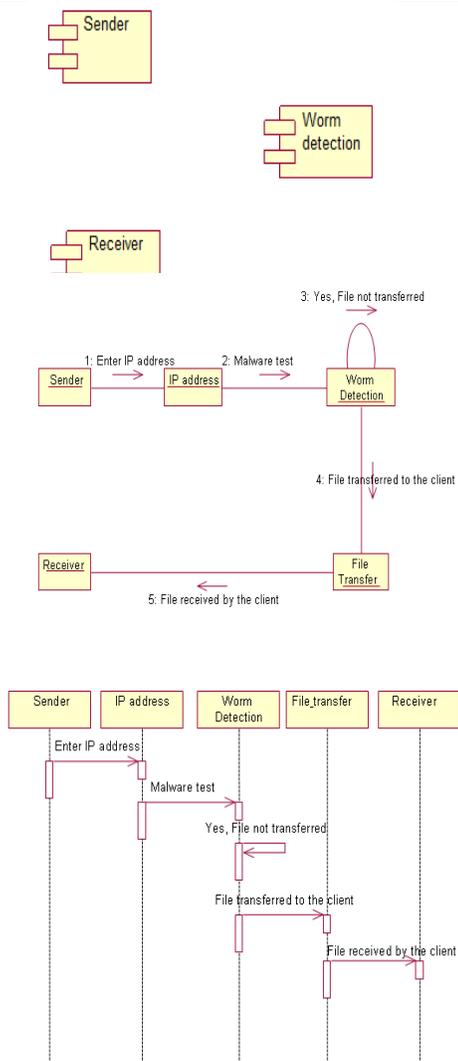
The System Consists of 3 Modules:

- P2P NETWORK MODULE
- MALWARE PROPAGATION
- INTERNET WORMS AND VIRUSES

IV. RESULT AND DISCUSSION

Software testing is a critical element if software quality assurance represents the ultimate reviews of specification, design and coding. Testing is vital of the system.

Errors can be injected at any stage during development. During testing, the program is executed with correctness.



. Unit Testing

In the unit testing the testing is performed on each module and this module is known as module testing. This testing was carried out during programming state itself. In this testing all the modules working satisfactorily as regard to the expected output from the module. Unit testing is a method by which individual units of source code are tested to determine if they are fit for use. A unit is the smallest testable part of an application. In procedural programming a unit may be an individual function or procedure. Unit tests are created by programmers or occasionally by white box testers.

Unit test cases embody characteristics that are critical to the success of the unit. These characteristics can indicate appropriate/inappropriate use of a unit as well as negative behaviors that are to be trapped by the unit. A unit test case, in and of itself, documents these critical characteristics, although many software development environments do not rely solely upon code to document the product in development. Unit testing provides a sort of living documentation of the system. Developers looking to learn what functionality is provided by a unit and how to use it can look at the unit tests to gain a basic understanding of the unit API.

Acceptance Testing

Acceptance testing is black-box testing performed on a system (e.g. software, lots of manufactured mechanical parts, or batches of chemical products) prior to its delivery. It is also known as functional testing, black-box testing, release acceptance, QA testing, application testing, confidence testing, final testing, validation testing, or factory acceptance testing.

Acceptance testing generally involves running a suite of tests on the completed system. Each individual test, known as a case, exercises a particular operating condition of the user's environment or feature of the system, and will result in a pass or fail, or Boolean, outcome. There is generally no degree of success or failure. The test environment is usually designed to be identical, or as close as possible, to the anticipated user's environment, including extremes of such. These test cases must each be accompanied by test case input data or a formal description of the operational activities (or both) to be performed—intended to thoroughly exercise the specific case—and a formal description of the expected results.

Types of Acceptance Testing

Typical types of acceptance testing include the following

A. User Acceptance Testing

This may include factory acceptance testing, i.e. the testing done by factory users before the factory is moved to its own site, after which site acceptance testing may be performed by the users at the site.

b. Operational Acceptance Testing

Also known as operational readiness testing, this refers to the checking done to a system to ensure that processes and procedures are in place to allow the system to be used and maintained.

c. Contract and Regulation Acceptance Testing

In contract acceptance testing, a system is tested against acceptance criteria as documented in a contract, before the system is accepted. In regulation acceptance testing, a system is tested to ensure it meets governmental, legal and safety standards.

Alpha And Beta Testing

Alpha testing takes place at developers' sites, and involves testing of the operational system by internal staff, before it is released to external customers. Beta testing takes place at customers' sites, and involves testing by a group of customers who use the system at their own locations and provide feedback, before the system is released to other customers. The latter is often called "field testing".

Integration Testing

One module can have adverse effect on another such functions when combined may not produce the desired results. Integration testing is a systematic technique for constructing the program structure and conducting test to uncover errors associated with interface. All the modules are combined in this testing step. The entire program is tested as the whole. The errors uncovered are corrected for the next testing step.

Black Box Testing

The black box approach is attesting method in which test data are delivered from the functional requirement without regard to the final program structure. Because only functionality of the software is concerned. In black box testing, only the functionality is determined by observing the outputs to the corresponding input. In this testing various input images are exercised and the output images are compared as required by the content retriever.

White Box Testing

White box testing are the software predicates on close examination of procedure details. It provides test cases that exercise specific test for conditions and loops. White box testing was carried out in the order to guarantee that

- All independent parts within a module exercised at least once.
- All logical decision on this true and false side was exercised

Computer input procedures are designed to detect errors in the data at the lower level of detail which is beyond the capability of the control procedures. The validation succeeds when the software functions in the manner that can be reasonably expected by the customer.

V CONCLUSION AND FUTURE DIRECTION

Peer-to-Peer (P2P) Networks continue to be popular means of trading content. Some existing studies have shown that Malware proliferation can pose significant threats to P2P Networks, defending against such an attack are largely an open problem. In this project we have explained the file sharing and problems in file sharing in P2P Networks and provide software implementations for P2P. Malware is highly pervasive in P2P file-sharing systems and is difficult to detect. Here in order to detect the Malware we have provided two basic analysis Code (Static) analysis, Behavioral (Dynamic) analysis and explained the goals for analyzing the Malware. These approaches reactive and proactive techniques are provided to prevent the malware.

REFERENCES

1. J. Munding, R. Weber, and G. Weiss, "Optimal Scheduling of Peer-to-Peer File Dissemination," *J. Scheduling*, vol. 11, pp. 105-120, 2007.
2. A. Bose and K. Shin, "On Capturing Malware Dynamics in Mobile Power-Law Networks," *Proc. ACM Int'l Conf. Security and Privacy in Comm. Networks (SecureComm)*, pp. 1-10, Sept. 2008.
3. R. Thommes and M. Coates, "Epidemiological Models of Peer-to-Peer Viruses and Pollution," *Proc. IEEE INFOCOM '06*, Apr. 2006.
4. J. Schafer and K. Malinka, "Security in Peer-to-Peer Networks: Empiric Model of File Diffusion in BitTorrent," *Proc. IEEE Int'l Conf. Internet Monitoring and Protection (ICIMP '09)*, pp. 39-44, May 2009.

Selective Jamming/Dropping Insider Attacks In Wireless Mesh Networks

S.Ganeshmoorthy¹, Ujwal U², Sujith V³

1. Assistant Professor, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore.
2. Final year-BCA, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore
3. Final year-BCA, Dept. of Computer Applications & IT, Sree Narayana Guru College, Coimbatore

Abstract

The latest Federal Communications Commission (FCC) ruling has enforced database-driven cognitive radio networks (CRNs), in which all secondary users (SUs) can query a database to obtain spectrum availability information (SAI). Database-driven CRNs are regarded as a promising approach for dynamic and highly efficient spectrum management paradigm for large-scale Internet of Things (IoT). However, as a typical location-based service (LBS), before providing services to the user, there is no verification of the queried location, which is very vulnerable to location spoofing attack. A malicious user can report a fake location to the database and access the channels that may not be available for its location. This will introduce serious interference to the primary users (PUs).

In this study, we identify a new kind of attack coined as location cheating attack, which allows an attacker to spoof other users to another location and make them query the database with wrong location, or allows a malicious user to forge location arbitrarily and query the database for services. To thwart this attack, we propose a novel infrastructure-based approach that relies on the existing WiFi or cellular network access points (or AP) to provide privacy-preserving location proof. With the proposed solution, the database can verify the locations without knowing the user's accurate location. We perform comprehensive experiments to evaluate the performance of the proposed approach. Experimental results show that our approach, besides providing location proofs effectively, can significantly improve the user's location privacy.

Key words: FCC,CRN, SU, SAI, LBS, IOT, AP, PU

I. INTRODUCTION

In this article we discuss various forms of sophisticated attacks in WMNs, in which an insider adversary intelligently exploits knowledge of leaked cryptographic secrets and protocol semantics to attack critical network functions such as channel access, routing, and end-to-end reliable data delivery.

We focus our attention on insider attacks that take the form of selective jamming and/or dropping of high-value packets in any given layer or combination of layers. Whereas selective jamming aims at preventing reception while the packet is in transmission, selective dropping is applied post reception.

Besides describing such attacks, we also highlight possible detection and mitigation mechanisms.

SYSTEM DESCRIPTION

Feasibility Study

Feasibility is the determination of whether a project is worth doing. The process followed in making this determination is called feasibility study. The feasibility of the project is analyzed in this phase and business proposal is out forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out.

Five keys considerations involved in the feasibility analysis are

- Technical feasibility
- Economic feasibility
- Operational feasibility

Technical Feasibility

This study is carried out to check the technical feasibility that is the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will leads to high demand on the resources that are placed on the client. The developed system must have a modest requirement, as only minimal or null chances are required for implementing this system. The main advantage of the project is to fine the multimedia files quickly.

Economic Feasibility

This involves the feasibility of the project to generate economic benefits. A benefit cost analysis and a breakeven analysis are important aspects of evaluating the economic feasibility of new industrial projects. It should demonstrate the net benefit of the proposed application in

light of the benefits and costs to the agency, other state agencies and the general public as a whole. The benefits and savings expected from the developed system out weigh the estimated cost. The developed system is economically feasible. This project is done with the available hardware and therefore it is economically feasible.

Operational Feasibility

Operational feasibility addresses the influences that a proposed project may have on the social system in the project environment. The ambient social structure may be such that certain categories of workers may be in short supply or nonexistent. The effect of the project on the social status of the project participants must be assessed to ensure compatibility. It is common Knowledge the computer installations have something to do understandable that the introduction of a candidate system requires special effort to educate, sell and train the Staff on new ways of considering business.

- Manual system is easy
- Training required
- Ensures security and confidentially
- No packet loss

Information will be more accurate than performed manually

II. LITERATURE STUDY

Most existing methods assume a continuously active adversary that systematically drops packets. These adversaries are detected by aggregate behavioral metrics such as per-packet reputation and credit. However, these metrics cannot detect attacks of selective nature, where only a small fraction of high-value packets is targeted. While all types of wireless networks are susceptible to insider attacks, WMNs are particularly vulnerable to them for a number of reasons.

Drawbacks:

- First, MPs and MAPs are relatively cheap devices with poor physical security, which makes them potential targets for node capture and compromise.
- Second, given their relatively advanced hardware (e.g., multiple transceivers per MP and MAP), WMNs often adopt a multichannel design, with one or more channels dedicated to control/broadcast purposes. Such static design makes it easier for an attacker to selectively target control/broadcast information.
- Third, the reliance on multi hop routes further accentuates the WMN vulnerability to compromised relays, which can drop control messages in order to enforce a certain routing behavior (e.g., force packets to follow long or inconsistent routes).

III. Development of selective jamming/dropping insider attacks in wireless mesh networks

- **First**, beacons are sent all the time and at a fast rate (typically, 10 to 100 frames per second) independent of any application.
- **Second**, the granularity of 802.11 TSF timers is 1 microsecond which is much higher than that of TCP time stamp clocks.
- **Third**, as the beacon time stamp is the actual time when an AP sends a frame (i.e., the time after the channel is sensed to be free) rather than the time when it is scheduled to send the frame, we do not need to consider any significant unpredictable delays incurred by the network as in the case of TCP time stamps.
- Therefore, our scheme estimates more accurate clock skews and much faster compared to the TCP/ICMP time stamp approach.
- We also improve upon the time taken for estimating the clock skew by using high-precision timers, at the fingerprinting node, that have resolutions in the order of microseconds to measure the arrival time of beacon frames.

Modules:

The System Consists of 3 Modules:

1. Reputation Systems
2. ACK-Based Systems
3. Credit-Based Systems

a. Reputation Systems:

Reputation systems identify misbehaving nodes based on per-node reputation metrics, computed based on interactions of each node with its peers.

These systems typically incorporate two critical operations:

The collection of accurate observations of nodes' behavior and the computation of the reputation metric. Behavioral information is collected based on first-hand observations provided by neighboring nodes and second-hand information provided by other interacting peers.

Firsthand observations are collected by monitoring nodes that operate in promiscuous mode in order to verify the correct forwarding of transmitted packets. Overhearing becomes problematic in the case of multichannel WMNs, because MPs and MAPs are scheduled to communicate in parallel over orthogonal frequency bands, and hence might not be available to monitor the behavior of other nodes.

Several schemes have been proposed for managing second-hand information. A node may flood warnings to the entire network if it detects a misbehaving node. Alternatively, information can be provided on demand after a request from a particular node has been received. In the latter scenario flooding of the request is necessary to discover nodes that possess second-hand information.

Both methods consume considerable bandwidth resources due to the underlying flooding operations for the dissemination and collection of second-hand information. Robust computation of reputation metrics is equally important for the identification of packet droppers. Simple aggregate metrics have been shown to be vulnerable to false accusations from colluding malicious nodes and suddenly changing behavioral patterns.

For instance, a misbehaving node can exhibit a long history of good behavior in order to build a high reputation metric before it starts to misbehave.

Such instances are dealt by assigning larger weights to recent behavioral observations and/or adopting additive increase multiplicative decrease type algorithms for updating the reputation metrics. A critical challenge for any metric computation algorithm is the selective nature of packet droppers. When a very small fraction of packets is dropped, metrics that do not take into account the packet type are bound to have high rates of misdetection.

Dropping selectivity can be detected with the use of storage-efficient reports (e.g., based on Bloom filters) of the per-packet behavior of nodes. Based on these reports, it is possible to conduct multiple tests to identify malicious selective dropping patterns. These patterns are

likely to have some deterministic structure compared to packet losses due to congestion or poor channel quality.

b. ACK-Based Systems:

Acknowledgment (ACK)-based schemes differ from overhearing techniques in the method of collecting first-hand behavioral observations. Downstream nodes (more than a single hop away) are responsible for acknowledging the reception of messages to nodes several hops upstream [10]. These systems are suitable for monitoring the faithful relay of unicast traffic, at the expense of communication overhead for relaying an additional set of ACKs.

However, ACK-based schemes cannot be used to identify insiders that selectively drop broadcast packets. Such packets remain, in general, unacknowledged in wireless networks to avoid an ACK implosion situation. Moreover, a small set of colluding nodes can still provide authentic ACKs to upstream nodes while dropping packets.

c. Credit-Based Systems:

Credit-based systems alleviate selfish behavior by motivating nodes to forward packets. Nodes that relay traffic receive credit in return, which can be spent later to forward their own traffic. However, in the context of WMNs, MPs do not generate any traffic of their own, but act as dedicated relays. Hence, compromised MPs have no incentive for collecting credit.

Moreover, in the case of selective dropping attacks, misbehaving nodes can still collect sufficient credit by forwarding packets of low importance while dropping a few packets of high value. In addition, the credit collected by a particular node depends on the topology of the network. A highly connected node is expected to collect more credit due to the increased volumes of traffic routed through it.

An adversary compromising such a node is likely able to implement a selective dropping strategy without running out of credit. Finally, credit-based systems lack a mechanism for identifying the misbehaving node(s), allowing them to remain within the network indefinitely.

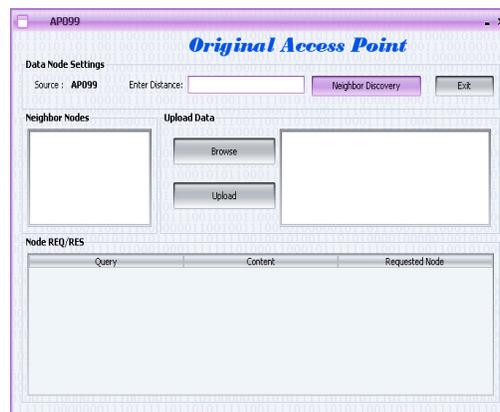
IV. RESULT AND DISCUSSION

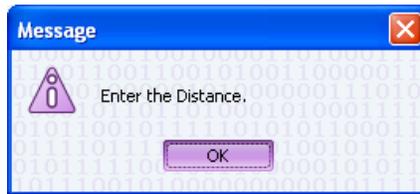
Software testing is a critical element if software quality assurance represents the ultimate reviews of specification, design and coding. Testing is vital of the system.

Errors can be injected at any stage during development. During testing, the program is executed with correctness. A series of testing are performed for the proposed systems before the system is delivered to the user.

Unit Testing

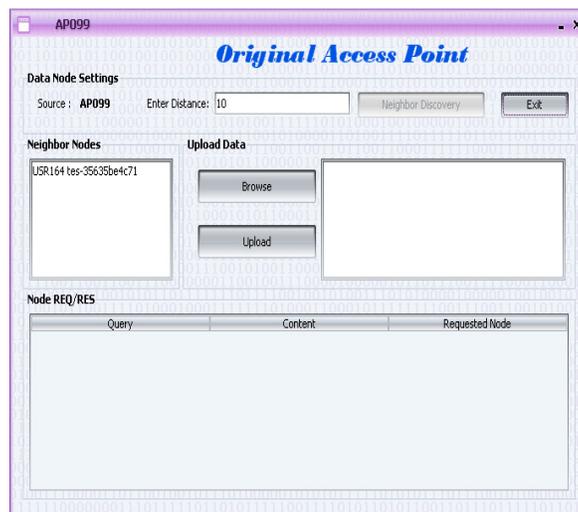
This is the first level of testing. The different modules are tested against the specifications produced during the integration. The input received and output generated is also tested to see whether it falls in the expected range of values. Unit testing is performed from the bottom up, starting with the smallest and lowest modules and proceeding one at a time. The units in a system are the modules and routines that are assembled and integrated to perform a specific function. Each of the modules was tested and errors are rectified. In the unit testing the entire project module is tested individually.





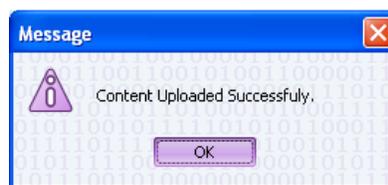
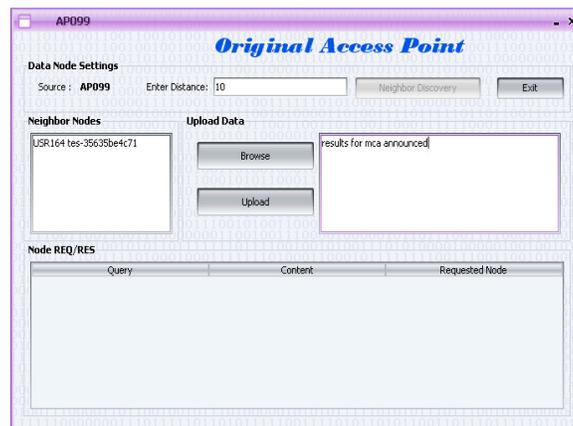
Validation Testing

Computer input procedures are designed to detect errors in the data at the lower level of detail which is beyond the capability of the control procedures. The validation succeeds when the software functions in the manner that can be reasonably expected by the customer.



Acceptance Testing

Acceptance testing is black-box testing performed on a system (e.g. software, lots of manufactured mechanical parts, or batches of chemical products) prior to its delivery. It is also known as functional testing, black-box testing, release acceptance, QA testing, application testing, confidence testing, final testing, validation testing, or factory acceptance testing.



SYSTEM IMPLEMENTATION

Implementation is the stage of the project where the theoretical design is turned into a working system. It can be considered to be the most crucial stage in achieving successful new system gaining the users confidence that the new system will work and will be effective and accurate. It is primarily concerned with user training and documentation. Conversion usually

takes place about the same time the user is being trained or later. Implementation simply means convening a new system design into operation, which is the process of converting a new revised system design into an operational one.

V CONCLUSION AND FUTURE DIRECTION

In this paper, we identify a new kind of attack coined as location cheating attack in database-driven CRNs, in which users can cheat their locations to gain more advantages, and this can cause interference to PUs. To thwart this attack, we propose a novel infrastructure-based approach that relies on the existing Wi-Fi AP network or cellular network to provide secure and privacy location proof.

On the one hand, we use a grid reference system with different granularities to represent locations; on the other hand, we adopt the private proximity testing technology to further improve the user's location privacy. We conduct the program to find the optimal strategy to maximum the user's location privacy. Simulations well demonstrate the effectiveness and efficiency of the proposed approach.

Further challenges include efficient behavioral monitoring mechanisms that do not rely on continuous overhearing, and efficient maintenance and dissemination of reputation metrics. Our future work includes other security issues in database-driven CRNs.

REFERENCES

- [1] Chen V, Das S, Zhu L, et al. Protocol to Access White-Space (PAWS) Databases. draft-ietf-paws-protocol-10 (work in progress), 2014.
- [2] Band, Broadcast. "FEDERAL COMMUNICATIONS COMMISSION 47 CFR Part 15."
- [3] Mills D, Martin J, Burbank J, et al. Network time protocol version 4: Protocol and algorithms specification. IETF RFC5905, June, 2010.
- [4] Narayanan, Arvind, et al. "Location Privacy via Private Proximity Testing." NDSS. 2011.

- [5] Zhang L, Fang C, Li Y, et al. "Optimal Strategies for Defending Location Inference Attack in Database-driven CRNs," International Conference on Communications (ICC), 2015 IEEE Conference on. IEEE, 2015.
- [6] Capkun, Srdjan, Levente Buttyan, and Jean-Pierre Hubaux. "SECTOR: secure tracking of node encounters in multi-hop wireless networks." Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks. ACM, 2003.
- [7] "TV Fool," March, 2012. [Online]. Available: <http://www.tvfool.com/> [8] Zhu, Zhichao, and Guohong Cao. "Applaus: A privacy-preserving location proof updating system for location-based services." INFOCOM, 2011 Proceedings IEEE. IEEE, 2011.
- [9] Zheng, Yao, et al. "Sharp: Private proximity test and secure handshake with cheat-proof location tags." Computer Security-ESORICS 2012. Springer Berlin Heidelberg, 2012. 361-378.
- [10] Li, Muyuan, et al. "All your location are belong to us: Breaking mobile social networks for automated user location tracking." ACM MobiHoc. ACM, 2014.

Advanced Techniques for Outlier Detection in High Dimensional Data

N Jayanthi¹, B Vijaya Babu ² and N Sambasiva Rao ³

Department of CSE, K L University, Guntur, Andhra Pradesh, India. email:jneelampalli.phd@gmail.com ¹

Department of CSE, K L University, Guntur, Andhra Pradesh, India. email: vijay_gemini@kluniversity.in²

Department of CSE, Vardhaman College of Engineering, Hyderabad, Telangana, India.

email: snandam@gmail.com.³

Abstract

Today's real-world applications of data mining or machine learning have a task for analyzing massive data. The huge amount of data with a large collection of dimensions or attributes is known as high dimensional data(HDD). Where ever data is huge some abnormal behavior or patterns are raised due to different activities. Such inconsistent patterns are known as outliers. Detection of outliers helps in good decision making. The challenging part is detecting outliers from high dimensional data. We shed light on various classification and clustering techniques for detecting outliers from high dimensional data. The aim of this paper is to provide a widespread and ordered summary of techniques used for detection of outlier in high dimensional data and propose a new technique for outlier detection.

1.Introduction

Outliers are generated due to malfunction hardware or malicious activities. There are various names for outliers in literature. Outliers behave differently or they deviate completely when compared with all other data points. Detecting outliers is a tough task in many research fields. Some of the fields where it is utilized are security purpose, medical sciences, detection of unauthorized access, crime and terrorist detection, liability finding, signal study, computer vision, abnormal detection of weather, uncharacteristic crowd behavior, surveillance.

Outlier detection is very much important as outliers hide interesting information. Ignoring the outliers may end up in gathering valuable information. In this direction, we would focus on basics and techniques for detecting outliers. Today data is generated from various sources in a huge amount resulting in high dimensional data. A data when arranged in the form of a table, which has a greater count of columns or attributes can be said as high dimensional data [1]. The accuracy of detecting outliers in high dimensional data will be degraded by traditional techniques as the count of columns increase.

Where existence of tiny amount of data is expected to contain outliers there would be a greater chance of outlier presences in high dimensional data Outliers from high dimensional data can be found either by classification or by clustering [2]. Figure 1 illustrates high dimensional data.

Cust	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
Cust 1	1	0	0	0	0	0	0	0	0	0
Cust 2	0	0	0	0	0	0	0	0	0	1
Cust 3	1	0	0	0	1	0	0	0	0	1

Fig 1: Example of High Dimensional Data.

In [3] authors illustrates data distribution in two dimensions shown in Fig 2. The data distribution has been done using a toy example. They explained that as dimensions increases it would become difficult to identify the clusters as well as outliers.

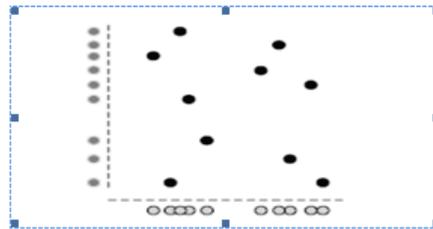


Fig 2: Data Distribution in 2 Dimension (taken from)

The basic ways for Outlier detection are classification or clustering or combination of both. There are several methods or algorithms derived, each having their own significance and each applicable to the different type of data. In classification labeled data is used. In clustering, unlabeled data is used. The third category is combination of these two where both types of data are used. Fig 3 shows first two methods of outlier detection.

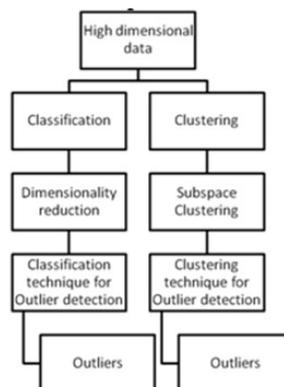


Fig 3: Different ways of detecting outliers.

The organization of the paper is Section 2 gives outlier detection techniques in high dimensional data using classification, the importance of clustering is given in Section 3, Subspace clustering details are presented in Section 4, Section 5 describes various outlier detection techniques in subspaces, the conclusion of the paper and future direction in section 6.

2. Classification and Dimensionality Reduction Methods for High Dimensionality Data

Classification is a critical task in data mining. Classification is used for classifying and prediction [4]. To do classification prior knowledge is required. From the outlier detection point of view, classification is considered as a binary problem where a point is labeled as an outlier or as a normal point.

In general, the high-dimensions contain relevant, irrelevant and duplicate dimensions. The unrelated dimensions and the redundant dimensions cannot be used in the training process as it degrades the performance. Such dimensions are handled by dimensionality reduction i.e., using a mapping function a low dimensional space is obtained from the high dimensional space. Principle component analysis, extreme learning machine, and linear discriminate analysis, low-rank matrix are some techniques for reducing dimensionality.

Dimensionality reduction is categories into three types: feature extraction, feature selection and feature clustering. Feature extraction means converting high dimensional data to low dimensional data. Feature selection means selecting relevant attributes from full space of attributes. Feature clustering reduces noise. It performs clustering on the feature set and so it keeps best possible most relevant portion of information about all features.

3. Importance of Clustering

In the real world, always having labeled information is very expensive and difficult thus clustering is advantageous over-classification; also the cost of labeling is reduced. From literature there exists a huge clustering algorithms are found. For classification, characterization and feature selection clustering can be used as preprocessing step [5].

Clustering is used for understanding as well as for discovering the formation of clusters in the dataset and extract the meaning of dataset. It split the data set into different cluster groups, where objects in one group will be highly similar and when compared to objects in another group or clusters [7]. Different sizes, densities, and shapes of clusters are formed during clustering. Clustering techniques measure the similarity of objects by measuring distance in between them. Common distance measures in traditional clustering are Euclidean or squared, Manhattan, maximum likelihood distance, Mahalanobis distance, Hamming distance, the angle between two vectors for high dimensional data [6].

Clustering algorithms are chosen by considering type of data, kind of clusters to be mined, and the purpose of derived clusters [8].

Clustering high dimensional data poses many challenges

Curse of dimensionality

Increasing dimensions will increase the distance between the points which in turn increases the Sparsity this is known as the curse of dimensionality.

Feature extraction

High dimensional data contains many attributes or features that are relevant and irrelevant. Only relevant attributes are helpful for good analysis and choosing such relevant attributes and ignoring irrelevant attributes is challenging.

Scalability

Outlier detection can be done by using traditional methods which gives superior performance in low dimensional data, but as the count of dimensions increases their performance decrease. Developing techniques to withstand in detecting outliers as dimensionality increase is alarming.

Computational complexity

With the increase in dimensionality, the detection of outliers increases exponentially.

Outlierness

Checking whether a point to be outlier or not an outlier is easy in low dimensional data but it is a difficult task in data with high dimensionality.

4. Subspace Clustering

Distance-based methods are not useful in high dimensional data as all attributes are distributed with equal distance in the space giving less scope in finding outliers. Thus a subset of attributes are selected for easy identification of outliers, such an approach is known as subspace clustering. Subspace clustering anatomy contains i) Subspace clustering ii) Projected clustering iii) Hybrid clustering iv) Correlation clustering.

Subspace clustering: This advanced clustering technique will convert given high dimensional data space into subspaces and search for clusters. [9]. In this technique, every data point is assigned to multiple cluster.

Projected clustering: Defines an algorithm not only to find clusters but also to find dimensions for the matching clusters, it also split out outliers from the clusters [6]. On the whole, in this method, a data point is assigned to only single cluster.

Hybrid clustering: Combination of any two clustering algorithms is hybrid clustering [10].

Correlation clustering: Here association of the vector containing features and their correlations are considered.

There are two approaches in subspace clustering: top down and bottom up. A list of these approaches is tabulated in Table 1 and Table2. The top-down approach uses projected clustering. Initially, all dimensions or attributes are considered for clustering then by iteratively ignoring irrelevant attributes finally optimal quality clusters are generated [11].

Table 1: Subspace Top Down Techniques

Top-down	Sub Space Clustering Name	Year
	ORCLUS	2000
	FINDIT	2004
	COSA	2004
	PROCLUS	2009

Bottom-up approaches initially consider single dimension clusters then two-dimensional clusters and so on till no dimensions are left. This is based on the Apriori principle.

Table 2: Subspace Bottom Up Techniques

Bottom Up	Sub Space Clustering Name	Year
	CLIQUE	1998
	MAFIA	1999
	ENCLUS	1999
	DOC	2002
	SUBCLU	2004
	FIRES	2005
	INSCY	2008
	DENCOS	2010
	SUBSCALE	2015

Subspace clustering techniques produce either 2D or 3D clusters. In a 2D cluster, solution produces 2D clusters where objects are represented by first dimension and attributes by second. A 3D cluster solution produces 3D clusters where, objects are represented by a first dimension, attributes by second and time or location by a third. [8].

5. Discovering Outliers in HDD using subspace

Global information and size of the neighborhood are the parameters for distance-based methods. Parameters for the neighborhood are used in density-based methods [12, 13]. As per machine learning algorithms, the methods are supervised, semi-supervised, and unsupervised scenarios based on labeled, unlabeled or combination of labeled and unlabeled data. [12]. A detailed list of existing outlier methods is listed in Table 3[14].

Outliers in high dimensional data can be detected either by classification or by clustering [15, 2]. An outlier can be detected either by using an integrated subspace plus outlier factor or by choosing subspace and outlier detection method separately. Different categories of outlier detection algorithms are

statistics-based or modeled based, distance-based, density-based, clustering based methods and subspace based. The methods with their example methods are shown in table 3.

Table 3: Existing outlier techniques in high dimensional data

Outlier detection method	Statistical Methods	Distance Methods	Density Methods	Clustering Methods	Subspace Methods
Example	STAT PC	LDO F	DBSCAN	PROCLUS	SUBSCALE
Use for high dimensional data	No	No	No	Not for all	Yes

Further classification of outlier detection methods for high-dimensional data includes the neighbor ranking-based methods, the subspace learning-based methods and the ensemble learning-based methods [16]. Fig 6 illustrates this classification. Examples of neighbor ranking-based methods are Rank based detecting algorithm (RBDA), Modified ranks with distance(MRD), examples of subspace learning methods are subspace outlier detection, Hics, and LoOP, examples of ensemble methods are bagging and sub sampling.

6. Conclusion and Further Direction

In this article, we provided a list of outlier detection techniques both by using classification and clustering. We would extend this survey by identifying new techniques for outlier detection in high dimensional data by using innovative subspaces, new outlier ranking method, choosing a coupled version of subspace and outlier ranking technique and choosing a decoupled way of subspace and outlier detection technique.

References

- [1] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," IEEE Trans. Knowl. Data Eng., vol. 26, no. 9, pp. 2250–2267, Sep. 2014.
- [2] Jiawei Han, Micheline Kamber, and Jian Pei. "Data Mining: Concepts and Techniques", 3rd Edition, 2012.
- [3] Ira Assent' "Clustering high dimensional data", John Wiley & Sons, Inc. Volume 2, Jul y /August 2012.
- [4] N Jayanthi et al, "survey on clinical prediction models for diabetes prediction", Journal of Big Data, Aug 2017.
- [5] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

- [6] M. Pavithra and Dr. R.M.S.Parvathi, “ A Survey on Clustering High Dimensional Data Techniques”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 11 (2017) pp. 2893-2899.
- [7] P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press, 25-72, 2011.
- [8] G.N.V.G. Sirisha et al,” Subspace clustering for high dimensional datasets” International Journal of Advanced Computer Research, ACCENTS, Vol 6(26), 2016.
- [9] hanj.cs.illinois.edu/cs412/bk3_slides/11ClusAdvanced.ppt
- [10] K. Choy, “Outlier detection for stationary time series,” Journal of Statistical Planning and Inference, vol. 99, no. 2, pp. 111–127, 2001.
- [11] Amardeep Kaur and Amitava Datta, “A novel algorithm for fast and scalable subspace clustering of high-dimensional data”, Journal of Big Data (2015) 2:17.
- [12] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Comput. Survey, vol. 41, no. 3, pp. 1–58, 2009.
- [13] C. C. Aggarwal, Outlier Analysis. New York, NY, USA: Springer, 2013.
- [14] Supriya Garule, “Outliers Detection using Subspace Method: A Survey”, International Journal of Computer Applications (0975 – 8887) Volume 112 – No 16, February 2015.
- [15] S. Upadhyaya and K. Singh, “Classification-based outlier detection techniques,” International Journal of Computer Trends and Technology, vol. 3, no. 2, pp. 294–298, 2012.
- [16] Xiaodan Xu “A Comparison of Outlier Detection Techniques for High-Dimensional Data” International Journal of Computational Intelligence Systems, Vol. 11 (2018) 652–662.

Market Dynamics of Bitcoin Currency with Indian Stock Market

Dinesh K,
Asst Professor at Department of MBA,
Ballari Institute of Technology & Management, Ballari.

ABSTRACT

Our paper examines the behavior of returns of emerging crypto currency namely Bitcoin traded in India with the stock Index returns of NSE NIFTY and India Volatility Index. We explored this study by taking monthly logarithmic rates of return for the period of March 2013 to December 2018. The data have been analyzed by making use of cross correlation between the Bitcoin India returns with NSE NIFTY Index returns and Bitcoin India returns with India Volatility Index (VIX). We have analyzed the significance of the effect by using the t-statistics from the calculation of Abnormal returns (AR), Cumulative Abnormal returns (CAR). From the study we observe the similarity of positive correlation behavior among of the Bitcoins, NSE NIFTY and NSE India VIX for the months April, September and November and has similarity of negative correlation for the month of February, March and June among the Bitcoins, NSE NIFTY and NSE India VIX for Monthly returns. The results in general evidence the partial existence of seasonality in market behavior among crypto currency and Indian Stock Market and in particular finds significant relationship between bitcoin with NSE NIFTY index and India VIX.

JEL classification: G02, G14, G18

Keywords: BSE SENSEX, Seasonal anomalies, Monthly effect, January effect.

1) Introduction

Today crypto currencies have become a global phenomenon known to most people. It has given birth to an incredibly dynamic, fast-growing market for investors and speculators. These crypto currencies are more used for payment, its use as a means of speculation and a store of value dwarfs the payment aspects. While [Bitcoin](#) remains by far the most famous crypto currency and most other crypto currencies have zero non-speculative impact, investors and users should keep an eye on several crypto currencies. This crypto currencies emerged as a side product of another invention. Satoshi Nakamoto, the unknown inventor of Bitcoin in late 2008 developed A Peer-to-Peer Electronic Cash System. Bitcoin began operating in January 2009 and is the first decentralized crypto currency, with the second crypto currency, Namecoin, not emerging until more than two years later in April 2011. Today, there are hundreds of crypto currencies with market value that are being traded, and thousands of crypto currencies that have existed at some point.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

According to Blockforce Capital, a U.S.-based asset manager. It compared Bitcoin and the Standard & Poor's 500 Index (S&P 500) from January 2015 through Oct. 11, 2018. and Found that , traditional Investment assets such as equity stocks, bonds, debenture, derivatives ,bullion ,silver, agricultural commodities and more novel assets, such as crypto currencies, may have sometimes perform similarly during short periods of time, however, In the long-term, it was found no relevant relationship between traditional investment assets such as marketable securities and crypto currencies Similar to gold and other types of precious metals, the value of the crypto market solely depends on the supply and demand within the market, unaffected by the performance of the global economy.

2) *Literature Review*

Moskowitz and Grinblatt (1999) documented a strong and prevalent momentum effect in industry components of stock returns which accounts for much of the individual stock momentum anomaly. Engelberg, and Gao (2011) measure of investor attention using search frequency in Google, In a sample of Russell 3000 stocks from 2004 to 2008 and found that ,its correlated with but different from existing proxies of investor attention .Moskowitz , Ooi, and Pedersen (2012) We document significant “time series momentum” in equity index, currency, commodity, and bond futures for each of the 58 liquid instruments and finds persistence in returns .Asness, Moskowitz, and Pedersen (2013) indicate the presence of common global risks that we characterize with a three-factor model. Global funding liquidity risk is a partial source of these patterns, which are identifiable only when examining value and momentum jointly across markets. Chen and Pandey (2014) examine the role of Bitcoin as a currency and its usefulness as an investment asset. They compare the correlation between Bitcoin and major world currencies. Yermack (2015) express that Bitcoin appears to behave more like a speculative investment than a currency. Huberman, Leshno, and Moallemi (2017) observes that Bitcoin payment system is a platform with two main constituencies: users and profit seeking miners who maintain the system's infrastructure. Chiu and Koepl (2017) estimates that the current Bitcoin scheme generates a large welfare loss of 1.4% of consumption, they also point out that cryptocurrencies can potentially challenge retail payment systems provided scaling limitations can be addressed.Routledge and Zetlin-Jones (2018) we show that blockchain distributed ledger technologies, such as those which support Bitcoin and Ethereum, can be adapted to eliminate self-fulfilling speculative attacks on a currency and they also shows that peg is immune to speculative attacks. Foley, Karlsen, and Putni (2018) find that cryptocurrencies are transforming the black markets by enabling “black e-commerce. They also estimate that around \$76 billion of illegal activity per year involves bitcoin (46% of bitcoin transactions), Cong Li and Wang (2018) provided a dynamic asset-pricing model of crypto currencies and highlight their roles on tokens facilitate transactions among decentralized users and allows them to capitalize future growth of promising platforms endogenous user adoption. Biais et al. (2018), Blockchains are distributed ledgers, operated within peer-to-peer networks. . Gilbert and Loi (2018) Bitcoin does not appear to carry much systematic risk.

3) *Significance of the Study*

There are various factors such as its price behavior, volatility of current market price and its correlation to other asset classes are predominantly the most preferred variables. Investor sentiment can play an important role in situations where both more traditional asset class (equity stocks) and more novel assets (crypto currencies) can demonstrate significant correlations The correlation analysis between the various assets is becoming more popular for investors using crypto assets

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

to diversify their overall portfolio risk and return. So, it's important to take a step back and look at what the currently know about crypto correlations with various other economic variables.

The following independent study made by [GeoLinkCrypto](#) and the analysis was done for a period of three years **2016–2018**. The visualization of the correlation in 2018 shows something interesting. The majority of pairs started to show dependency and increasing correlation. This crypto cross correlation is observed below in the table

Table 1: Crypto currency correlation Matrix 2018

Cryptocurrency Correlation Matrix,365 –days

	<i>BTC</i>	<i>ETH</i>	<i>XRP</i>	<i>BCH</i>	<i>XLM</i>	<i>LTC</i>	<i>XMR</i>	<i>DASH</i>	<i>ETC</i>	<i>XEM</i>	<i>ZEC</i>	<i>REP</i>	<i>SPX</i>	<i>VIX</i>	<i>GLD</i>	<i>TNX</i>
<i>BTC</i>	1	0.69	0.48	0.46	0.63	0.7	0.72	0.64	0.58	0.59	0.63	0.5	0.03	-0.12	-0.05	0.03
<i>ETH</i>	0.69	1	0.64	0.63	0.61	0.79	0.76	0.76	0.74	0.7	0.77	0.66	0.05	-0.12	0.04	0.05
<i>XRP</i>	0.48	0.64	1	0.39	0.66	0.55	0.52	0.52	0.52	0.49	0.61	0.4	0.08	-0.1	0.03	-0.03
<i>BCH</i>	0.46	0.63	0.39	1	0.37	0.53	0.64	0.7	0.62	0.47	0.6	0.5	-0.04	-0.05	-0.02	0.06
<i>XLM</i>	0.63	0.61	0.66	0.37	1	0.58	0.64	0.57	0.56	0.63	0.61	0.43	0.07	-0.09	0.05	-0.03
<i>LTC</i>	0.7	0.79	0.55	0.53	0.58	1	0.7	0.71	0.65	0.68	0.68	0.53	0.07	-0.11	0.02	0.03
<i>XMR</i>	0.72	0.76	0.52	0.64	0.64	0.7	1	0.78	0.65	0.63	0.77	0.6	-0.01	-0.08	-0.04	0.05
<i>DASH</i>	0.64	0.76	0.52	0.7	0.57	0.71	0.78	1	0.63	0.64	0.76	0.59	0.03	-0.08	0.01	0.09
<i>ETC</i>	0.58	0.74	0.52	0.62	0.56	0.65	0.65	0.63	1	0.56	0.66	0.54	0.04	-0.11	-0.02	0.02
<i>XEM</i>	0.59	0.7	0.49	0.47	0.63	0.68	0.63	0.64	0.56	1	0.63	0.51	0.08	-0.12	0	0.05
<i>ZEC</i>	0.63	0.77	0.61	0.6	0.61	0.68	0.77	0.76	0.66	0.63	1	0.57	-0.02	-0.05	-0.01	0.03
<i>REP</i>	0.5	0.66	0.4	0.5	0.43	0.53	0.6	0.59	0.54	0.51	0.57	1	-0.03	0	0.03	0.12
<i>SPX</i>	0.03	0.05	0.08	-0.04	0.07	0.07	-0.01	0.03	0.04	0.08	-0.02	-0.03	1	-0.82	0	0.24
<i>VIX</i>	-0.12	-0.12	-0.1	-0.05	-0.09	-0.11	-0.08	-0.08	-0.11	-0.12	-0.05	0	-0.82	1	-0.06	-0.25
<i>GLD</i>	-0.05	0.04	0.03	-0.02	0.05	0.02	-0.04	0.01	-0.02	0	-0.01	0.03	0	-0.06	1	-0.19
<i>TNX</i>	0.03	0.05	-0.03	0.06	-0.03	0.03	0.05	0.09	0.02	0.05	0.03	0.12	0.24	-0.25	-0.19	1

source: *Sifr Data*

The above picture depicts the correlation graph showing that the relationship between Bitcoin and S&P 500 is at a weak positive relationship. But, the correlation between VIX and Bitcoin -0.12 and Gold -0.05 making it a weak negative relationship. Volatility Index indicates the level of risk present in the market at any time.

According to the above graph, it demonstrated the inverse correlation between VIX and Bitcoin price in and not with the S & P 500 stock market Index. Understanding how the bitcoin in India moves relative to India's most popular and active stock market index and also to measure its movement with India Volatility Index is certainly a new research to look into by conducting correlation analysis in this given study.

4) *Objective, Hypothesis, Data and Methodology*

Objective

Objective of this paper is to test the relationship between returns of Bitcoin USD with NSE NIFTY index and India VIX

Hypothesis

Based on the available evidence from the independent study conducted by Geo Link Crypto analyzed for NYSE data for the period of 2016 to 2018 depicted the relationship between Bitcoin and S & P 500 is at weak positive relationship and between VIX and bitcoin weak negative relationship. Hence an attempt is made to study the similar sample to identify its correlation in Indian context. Therefore the hypothesis for the study stated as,

H_0 : There is no relationship between the returns of Bitcoin with NSE NIFTY Index and India VIX.

Data and Sample

This study observes data of Bitcoins, NSE NIFTY and NSE India VIX from March 2013 to December 2018. Data is obtained from two sources: yahoo finance website and nseindia website. We have used the monthly closing price of Bitcoin in USD, NSE Nifty and India VIX data for the study are obtained, the data for Bitcoin in USD in yahoo finance is available from the March 2013. Therefore the base period for the study is considered as the March 2013, but the data for the month of October of all the years is missing from the website source, therefore we have matched the data similarly for NSE NIFTY and India VIX for the study.

Methodology

We have used sample data from January 2014 to December 2018 to find the correlation among the variables. We have applied Cross Correlation as a standard method of estimating the degree to which Bitcoin with NSE NIFTY and India VIX series are correlated.

To understand the Abnormal return relationship between Bitcoin and NSE NIFTY and also to know the relationship between Bitcoin with India VIX. We consider five years' period (January 2014-December 2018) which will be further divided into two sub samples with an equal number of periods April 2013 to December 2015 and January 2016 to December 2018. The following methodology has been used to analyze the effect of bitcoin over NSE NIFTY and India VIX.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

The CAPM is used to evaluate the rate of return of an asset correlated to excess market return. The CAPM suggests that a return of an asset equals to Intercept represented by alpha plus product of systematic risk with market return. The mathematical expression of the CAPM is as follows $R_b = \alpha + \beta \times R_m$

Here R_b is bitcoin excess return, R_i is the actual returns of bitcoin, b is the price of bitcoin in USD. The abnormal return is the actual ex-post return of the security over sample period. The normal return is defined as the expected return estimated by market model. Abnormal return is the difference between actual log return (LN) of the security and the expected returns calculated from the OLS regression equation. The abnormal return is calculated are as follows;

The following formula are: $AR_{it} = R_{it} - \epsilon \frac{rit}{xt}$. AR_{it} : is the abnormal returns for the sample and R_{it} is actual returns for the given sample .The cumulative Abnormal Returns for each sample are aggregated to get the cumulative abnormal returns. The cumulative abnormal returns are calculated as follows: $CAR(t1, t2) = \sum AR$ returns of the security are aggregated to get cumulative abnormal returns and to test the hypothesis t-statistics at 5 % level of significance is applied.

5) Data Analysis & Discussion

Table 2: Showing the Cross correlation between Bitcoin and NSE Nifty log return from 2014 to 2018

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEPT	NOV	DEC
JAN (BITCOIN)	1										
JAN (NSE NIFTY)	1										
FEB (BITCOIN)	-0.005	1									
FEB (NSE NIFTY)	-0.142	1									
MAR (BITCOIN)	-0.143	0.879	1								
MAR (NSE NIFTY)	-0.451	-0.01	1								
APR (BITCOIN)	0.805	0.394	0.383	1							
APR (NSE NIFTY)	0.482	0.647	-0.522	1							
MAY (BITCOIN)	0.674	0.302	-0.092	0.375	1						
MAY (NSE NIFTY)	0.107	0.601	-0.08	0.813	1						
JUN (BITCOIN)	-0.932	-0.042	0.255	-0.638	-0.865	1					
JUN (NSE NIFTY)	-0.332	-0.114	0.783	-0.724	-0.6	1					

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

JUL (BITCOIN)	0.387	0.69	0.662	0.831	0.141	-0.279	1				
JUL (NSE NIFTY)	-0.33	0.498	0.745	0.144	0.589	0.27	1				
AUG (BITCOIN)	-0.512	0.002	0.118	-0.632	-0.108	0.468	-0.653	1			
AUG (NSE NIFTY)	0.59	0.355	-0.9	0.798	0.334	-0.744	-0.468	1			
SEPT (BITCOIN)	0.097	0.836	0.967	0.582	0.015	0.062	0.737	0.015	1		
SEPT (NSE NIFTY)	0.762	0.448	-0.604	0.756	0.264	-0.387	-0.259	0.862	1		
NOV (BITCOIN)	0.762	0.524	0.5	0.986	0.423	-0.612	0.847	-0.528	0.684	1	
NOV (NSE NIFTY)	0.319	-0.644	0.55	-0.476	-0.306	0.482	0.15	-0.572	-0.283	1	
DEC (BITCOIN)	0.446	0.573	0.698	0.654	0.288	-0.263	0.484	0.164	0.833	0.744	1
DEC (NSE NIFTY)	0.108	0.842	0.318	0.564	0.602	0.128	0.727	0.12	0.437	-0.134	1

Source: computed from excel

This table 2 analysis shows the cross correlation for the Bitcoin prices and NSE Nifty Index monthly returns from the 2014 to 2018 time period. We can observe the positive correlation for both the Bitcoin and NSE Nifty returns for the month of April(Bitcoin (0.805 & Nifty(0.482) , May(Bitcoin (0.674 & Nifty(0.107), September (Bitcoin (0.097 & Nifty(0.762) ,November (Bitcoin (0.762 & Nifty(0.319), and December(Bitcoin (0.446 & Nifty(0.108). While it has the negative correlation for the month of February (Bitcoin (-.005 & Nifty(-0.142) ,March (Bitcoin (-0.142 & Nifty(-0.143), and June(Bitcoin (-0.932 & Nifty(-0.332), . Where the correlation was not the same for the month of July (Bitcoin (0.387 & Nifty(-0.33), and August(Bitcoin (-0.512 & Nifty(0.59). While for the September month has the least positive value with bitcoin(0.097) and greater positive with Nifty(0.762) was observed in the study. The higher negative correlation for Bitcoin was observed in January and June (-0.932) and for Nifty with the month of March and August (-0.9). The higher positive correlation for Bitcoin was observed in April and November (0.986) and for Nifty with the month of September and August (0.862).

Table 3 :Showing the Cross correlation between Bitcoin and India VIX log return from 2014 to 2018

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEPT	NOV	DEC
Jan(BITCOIN)	1										
Jan (INDIA VIX)	1										
Feb (BITCOIN)	-0.005	1.000									
Feb (INDIA VIX)	-0.323	1.000									

DST Sponsored National Conference on Recent Advancements on Computer
Science (CONRACS 2019)- 26 to 28 July 2019

Mar (BITCOIN)	-0.143	0.879	1.000								
Mar (INDIA VIX)	-0.477	-0.671	1.000								
Apr (BITCOIN)	0.805	0.394	0.383	1.000							
Apr (INDIA VIX)	0.434	0.288	-0.587	1.000							
May (BITCOIN)	0.674	0.302	-0.092	0.375	1.000						
May (INDIA VIX)	-0.605	0.334	0.206	0.370	1.000						
Jun (BITCOIN)	-0.932	-0.042	0.255	-0.638	-0.865	1.000					
Jun (INDIA VIX)	-0.123	0.057	-0.018	-0.751	-0.699	1.000					
Jul (BITCOIN)	0.387	0.690	0.662	0.831	0.141	-0.279	1.000				
Jul (INDIA VIX)	-0.448	0.322	0.102	0.520	0.983	-0.809	1.000				
Aug (BITCOIN)	-0.512	0.002	0.118	-0.632	-0.108	0.468	-0.653	1.000			
Aug (INDIA VIX)	-0.203	-0.117	0.242	-0.922	-0.615	0.909	-0.735	1.000			
Sept (BITCOIN)	0.097	0.836	0.967	0.582	0.015	0.062	0.737	0.015	1.000		
Sept (INDIA VIX)	0.969	-0.474	-0.302	0.247	-0.662	-0.064	-0.521	-0.051	1.000		
Nov (BITCOIN)	0.762	0.524	0.500	0.986	0.423	-0.612	0.847	-0.528	0.684	1.000	
Nov (INDIA VIX)	0.081	0.348	-0.347	-0.386	-0.410	0.499	-0.423	0.625	0.172	1.000	
Dec (BITCOIN)	0.446	0.573	0.698	0.654	0.288	-0.263	0.484	0.164	0.833	0.744	1.000
Dec (INDIA VIX)	-0.678	0.316	0.160	-0.182	0.277	0.299	0.152	0.068	-0.781	-0.436	1.000

Source: computed from excel

This table 3 shows the analysis of cross correlation for the Bitcoin price and NSE India VIX monthly returns from the 2014 to 2018 time period. We can observe the positive correlation for both the Bitcoin and India VIX returns for the month of April (Bitcoin (0.805 & India VIX (0.432) , September (Bitcoin (0.097 & India VIX (0.969) and November (Bitcoin (0.762 & India VIX (0.081) , the negative correlation for the month of February (Bitcoin (-0.005 & India VIX (-0.323), March (Bitcoin (-0.143 & India VIX (-0.451), June (Bitcoin (-0.932 & India VIX (-0.123) and for August Bitcoin (-0.512 & India VIX (-0.203) .The correlation value was not the same for the month of May (Bitcoin (0.674 & India VIX (-0.605), July (Bitcoin (0.387 & India VIX (-0.448) and December (Bitcoin (0.446 & India VIX (-0.678). The higher negative correlation for Bitcoin was observed in January - June (-0.932) and for India VIX with the month of August-April (-0.922).

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

The higher positive correlation for Bitcoin was observed in April-November (0.986) and for India VIX with the month of July-May (0.983).

Table 4 :Showing the results of Average Abnormal returns between Bitcoin and NSE NIFTY ,India VIX for the data April 2013 to December 2015

RESULTS OF BITCOIN WITH NSE NIFTY INDEX(N=29)					RESULTS OF BITCOIN WITH NSE INDIA VIX (N=29)			
Date	AR	t-test	Sig	CAR	AR	t-test	Sig	CAR
Apr-13	-19.961	-0.568	NO	-0.568	-16.288	-0.469	NO	-16.288
May-13	-16.809	-0.478	NO	-1.046	-21.015	-0.605	NO	-37.303
Jun-13	10.777	0.307	NO	-0.739	2.123	0.061	NO	-35.180
Jul-13	30.665	0.872	NO	0.133	27.334	0.786	NO	-7.846
Aug-13	-10.429	-0.297	NO	-0.164	-5.930	-0.171	NO	-13.777
Sep-13	19.292	0.549	NO	0.385	26.867	0.773	NO	13.090
Nov-13	187.753	5.340	YES	5.725	186.359	5.361	YES	199.449
Dec-13	-54.856	-1.560	NO	4.165	-55.062	-1.584	NO	144.387
Jan-14	0.190	0.005	NO	4.171	-2.842	-0.082	NO	141.545
Feb-14	-48.753	-1.387	NO	2.784	-41.309	-1.188	NO	100.236
Mar-14	-32.171	-0.915	NO	1.869	-34.583	-0.995	NO	65.653
Apr-14	17.352	0.494	NO	2.362	23.090	0.664	NO	88.743
May-14	-6.299	-0.179	NO	2.183	-0.374	-0.011	NO	88.369
Jun-14	-18.309	-0.521	NO	1.662	-9.875	-0.284	NO	78.494
Jul-14	-22.222	-0.632	NO	1.030	-26.672	-0.767	NO	51.822
Aug-14	-23.552	-0.670	NO	0.360	-18.651	-0.537	NO	33.171
Sep-14	-24.365	-0.693	NO	-0.333	-22.937	-0.660	NO	10.234
Nov-14	0.820	0.023	NO	-0.309	6.245	0.180	NO	16.479

DST Sponsored National Conference on Recent Advancements on Computer
Science (CONRACS 2019)– 26 to 28 July 2019

Dec-14	-21.082	-0.600	NO	-0.909	-12.208	-0.351	NO	4.270
Jan-15	-26.573	-0.756	NO	-1.665	-36.984	-1.064	NO	-32.714
Feb-15	-10.324	-0.294	NO	-1.958	-14.680	-0.422	NO	-47.394
Mar-15	-0.077	-0.002	NO	-1.961	-7.807	-0.225	NO	-55.201
Apr-15	8.814	0.251	NO	-1.710	10.458	0.301	NO	-44.742
May-15	-18.106	-0.515	NO	-2.225	-14.965	-0.431	NO	-59.707
Jun-15	9.127	0.260	NO	-1.965	1.778	0.051	NO	-57.929
Jul-15	-12.107	-0.344	NO	-2.310	-5.696	-0.164	NO	-63.625
Aug-15	-2.818	-0.080	NO	-2.390	-7.456	-0.214	NO	-71.081
Sep-15	24.373	0.693	NO	-1.697	14.727	0.424	NO	-56.354
Nov-15	-5.160	-0.147	NO	-1.843	-6.234	-0.179	NO	-62.588
Dec-15	34.446	0.980	NO	-0.864	28.215	0.812	NO	-34.373
* Significant at 5% level								

Source: computed from excel

Table 5: Showing the results of Average Abnormal returns between Bitcoin and NSE NIFTY, India VIX for the data April 2016 to December 2018

RESULTS OF BITCOIN WITH NSE NIFTY INDEX(N=33)					RESULTS OF BITCOIN WITH NSE INDIA VIX (N=33)			
Date	AR	t-test	Sign	CAR	AR	t-test	Sig	CAR
Jan-16	-6.863	-0.256	-0.007	NO	-12.452	-0.358	NO	-46.825
Feb-16	18.361	-4.578	-0.130	NO	0.005	0.000	NO	-46.820
Mar-16	6.148	-10.209	-0.290	NO	-7.286	-0.210	NO	-54.106
Apr-16	9.551	12.363	0.352	NO	14.111	0.406	NO	-39.996
May-16	6.319	16.084	0.458	NO	18.728	0.539	NO	-21.267
Jun-16	9.924	-18.042	-0.513	NO	-19.555	-0.563	NO	-40.823

DST Sponsored National Conference on Recent Advancements on Computer
Science (CONRACS 2019)- 26 to 28 July 2019

Jul-16	6.518	-10.277	-0.292	NO	-6.890	-0.198	NO	-47.712
Aug-16	1.362	-4.766	-0.136	NO	-9.299	-0.268	NO	-57.011
Sep-16	4.593	18.341	0.522	NO	22.455	0.646	NO	-34.556
Nov-16	-3.316	11.232	0.320	NO	3.307	0.095	NO	-31.249
Dec-16	10.396	11.822	0.336	NO	13.231	0.381	NO	-18.017
Jan-17	9.236	-7.825	-0.223	NO	-2.721	-0.078	NO	-20.738
Feb-17	8.692	11.376	0.324	NO	13.467	0.387	NO	-7.271
Mar-17	6.121	3.354	0.095	NO	0.989	0.028	NO	-6.282
Apr-17	8.822	60.052	1.708	NO	62.375	1.794	NO	56.093
May-17	2.703	-2.795	-0.080	NO	-4.295	-0.124	NO	51.798
Jun-17	12.054	-15.097	-0.429	NO	-10.527	-0.303	NO	41.271
Jul-17	1.946	67.056	1.907	NO	63.971	1.840	NO	105.243
Aug-17	2.336	-25.412	-0.723	NO	-31.013	-0.892	NO	74.230
Sep-17	11.717	41.363	1.177	NO	48.809	1.404	NO	123.039
Nov-17	6.765	50.684	1.442	NO	51.308	1.476	NO	174.346
Dec-17	10.572	23.512	0.669	NO	32.654	0.939	NO	207.000
Jan-18	-2.752	-42.247	-1.202	NO	-42.705	-1.229	NO	164.295
Feb-18	-0.957	-7.633	-0.217	NO	-14.156	-0.407	NO	150.139
Mar-18	12.506	-22.185	-0.631	NO	-20.089	-0.578	NO	130.050
Apr-18	4.119	-14.181	-0.403	NO	-16.420	-0.472	NO	113.630
May-18	3.877	-21.246	-0.604	NO	-22.186	-0.638	NO	91.444
Jun-18	12.254	9.763	0.278	NO	11.599	0.334	NO	103.043
Jul-18	8.072	-19.843	-0.564	NO	-17.414	-0.501	NO	85.628
Aug-18	-5.067	-4.919	-0.140	NO	-16.915	-0.487	NO	68.713

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

Sep-18	-2.936	5.851	0.166	NO	8.926	0.257	NO	77.638
Nov-18	10.390	-49.386	-1.405	NO	-37.564	-1.081	NO	40.075
Dec-18	4.160	-31.564	-0.898	NO	-40.534	-1.166	NO	-0.460
* Significant at 5% level								

Source: computed from excel

The result analysis from the above table 4 and table 5, here the sample is tested by calculating the abnormal returns and Cumulative Abnormal return .The t-statistics is used to test the significance of the returns in the analysis for the period of 2013 to 2018.The average abnormal returns (AR) continues to be significant for the each month except for the month of November 2013.However we can indicate that , the returns of the Bitcoin with NSE Nifty and India VIX has statistically significant for the entire sample period . Therefore, as stated in the null hypothesis that there is no relationship with the returns among the Bitcoin with the NSE NIFTY Index returns and India VIX can be rejected. The findings in this study seem to indicate that there is a significant relationship among the Bitcoin currency with the Nifty Index and India VIX.

6) Conclusion

Investing in crypto currencies is highly risky and speculative In the upcoming years, digital assets could become a viable alternative to gold and other traditional assets. The behavior of returns of emerging crypto currency namely Bitcoin traded in India with the stock Index returns of NSE NIFTY.and India Volatility Index will certainly a opportunity for hedgers, investors and speculators and more appealing to institutional investors to make better investment decisions .In this study we observe the similarity of positive correlation behavior among of the Bitcoins with NSE NIFTY and NSE India VIX for the months April, September and November and has similarity of negative correlation for the month of February, March and June among the Bitcoins, NSE NIFTY and NSE India VIX for Monthly returns. this study indicate that there is a significant relationship among the Bitcoin currency with the Nifty Index and India VIX returns.

7) Bibliography

- 1) Agrawal, Kishore Tandon. (1994). Anomalies or illusions? Evidence from stock markets in eighteen countries. *Journal of International Money and Finance*, 13 (1), 83-106.
- 2) Bruno Biais, Christophe Bisière, Mathieu Bouvard, & Catherine Casamatta. (2018). The block chain folk theorem. TSE Working paper.
- 3) Lin William Cong, Ye Li & Neng Wang. (2018, April). Tokenomics: Dynamic Adoption and Valuation. Columbia Business School Research, 18-46.
- 4) Vijeta Banwari (2017). CRYPTOCURRENCY-SCOPE IN INDIA. *International Research Journal of Management Sociology & Humanities*, 8 (12), 82-92.
- 5) David Yermack (2014, April). Is Bitcoin a Real Currency? An economic appraisal. NBER Working Paper No. 19747, 31-44.
- 6) Eugene F. Fama. (1965). Behavior of stock market prices. (38, Ed.) *Journal of Business*, 34-105.
- 7) Eugene F. Fama (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25, 383-417.
- 8) Eugene F. Fama (1991). Efficient Capital Markets: II. *Journal of Finance*, 46 (5), 1575–1617.
- 9) Gur Huberman, Jacob Leshno ,Ciamac C. Moallemi .(2017, October). Monopoly without a Monopolist: An Economic Analysis of the Bitcoin Payment System. Working Paper.
- 10) Kewei Hou, Chen Xue, & Lu Zhang. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28 (3), 650–705.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

- 11) Tobias Moskowitz., Yao Huaooi & Lasse Heje Pedersen. (2012). Time series momentum. Journal of financial economics 1, 104 (2), 228–250.
- 12) Jonathan Chiu, Thorsten V. Koepp (2017, September). The Economics of Crypto currencies – Bit coin and beyond. SSRN.
- 13) Kapil Choudhary & Sakshi Choudhary. (2008). Day-of-the-Week Effect: Further Empirical Evidence. Asia-Pacific Business Review, 4 (3), 67-74.
- 14) Keim, D. (1983, 12). Size related anomalies and stock return seasonality: Further empirical evidence. (1, Ed.) Journal of Financial Economics, 13-32.
- 15) Moskowitz, T. J., & M. G. (1999). Do Industries Explain Momentum? The Journal of Finance, 54 (4), 1249-1290.
- 16) Routledge, B., & Jones, A. Z. (2018). Currency Stability Using Blockchain Technology. Society for Economic Dynamics.
- 17) S. F., Karlsen, J. R., & Putniņš, T. J. (2018, Janury). Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed Through Cryptocurrencies? Review of Financial Studies, 65.
- 18) [Scott Gilbert](#), & Choudhary, S. (2008). Day-of-the-Week Effect: Further Empirical Evidence. Asia-Pacific Business Review [Hio Loi](#). (2018). Digital Currency Risk. International Journal of Economics and Finance, 10 (2), 108-123.
- 19) Taufiq Choudhry (2000). Day of the Week Effect in Emerging Asian Stock Markets: Evidence from the GARCH Model. Applied Financial Economics, 10, 235 – 242.
- 20) VK Gimba. (2011). testing the Weak-form Efficiency Market Hypothesis: Evidence from Nigerian Stock Market. CBN Journal of Applied Statistics, 3 (1), 117-136.
- 21) Valentina Zakirova, G. C. (2017). Calendar anomalies in the Russian stock market . Russian Journal of Economics, 3 (1), 101-108.
- 22) Valentina, T. E. (2015). Intra monthly Anomalies on the Bucharest Stock Exchange . Procedia Economics and Finance, 32, 271-277.

8) Website

- 1) <https://cointelegraph.com/news/so-is-there-a-correlation-between-bitcoin-and-stock-market-yes-but-no>
- 2) <https://www.forbes.com/sites/cbovaird/2018/10/16/is-the-correlation-between-stocks-and-bitcoin-real/#3ffd1ae6146e>
- 3) <https://www.investopedia.com/news/are-bitcoin-price-and-equity-markets-returns-correlated/>
- 4) <https://www.newsbtc.com/2018/10/16/higher-correlation-between-stocks-and-bitcoin-may-hurt-crypto-market-analyst-says/>
- 5) <https://www.ccn.com/crypto-surges-23-billion-amid-us-stock-market-recovery-direct-correlation/>
- 6) <https://www.ccn.com/stock-market-plunges-but-has-no-impact-on-crypto-no-correlation/>
- 7) <https://bravenewcoin.com/insights/a-scary-month-for-stocks-but-will-bitcoin-benefit>
- 8) <https://hackernoon.com/about-those-crypto-correlations-e68720707efe>
- 9) <https://dailyhodl.com/2018/11/12/correlation-between-bitcoin-and-stocks-to-damage-crypto-industry/>
- 10) <https://economics.yale.edu/sites/default/files/files/Faculty/Tsyvinski/cryptoreturns%208-7-2018.pdf>

A SURVEY ON DIAGNOSIS OF DENTAL CAVITIES DATA USING BIG DATA TECHNOLOGIES

P.T.S.S Roopesh, Dr.Asadi Srinivasulu and Dr.P.Venugopal Prudhvi

tharakroopeh@gmail.com, srinu.asadi@gmail.com, venugopal.prudhvi@gmail.com

Sree Vidyanikethan Engineering Collage, Tirupathi AP

ABSTRACT

Dental cavity is that the unwellness within the human mouth that is caused by completely different microorganism activities. Cavities create associate everlasting injury within the tooth and it leads to holes within tooth. Dealing properly with dental cavities associated taking an pressing treatment is usually counseled to avoid a lot of injury. Tooth doctor acknowledges the cavity in patients' teeth by trying directly with eyes and typically with facilitate of x-ray (radiograph) of teeth. The machine-controlled system would facilitate the tooth doctor to spot the cavity in teeth by creating use of x-ray. This paper proposes a model to notice the cavities exploitation x-ray pictures by creating use of varied image process techniques, involving RGB to grey conversion, generation of binary image, finding the region of interest, removing background, distinguishing regions and dividing image into multiple blocks and at last distinguishing the cavities gift in x-rayimage.

Keywords

Big data, Dental caries, dental cavity, cavity detection, image processing, caries detection, x-ray images, region detection.

1. INTRODUCTION

A human tooth is a structure made up of dentin, pulp and enamel. Mouth normally consists of various types of bacteria, they causes infection in human teeth. These infections generally termed as dental caries. Caries further damage the teeth permanently and results in tooth cavity. It is a very common disease found in world, about 60 to 90% of school children and nearly all adults have dental cavities [1]. Dental cavity affects the daily task of teeth by weakening the biting capacity, increased sensitivity, tooth ache, etc. Commonly when after meals, if the mouth is not washed properly, the food stays in corners of teeth, this deposited food generates acid. Such type of acid and sugar on teeth cover produces bacteria and it leads to breakdown of the tooth enamel (hard tissue of the teeth). Which causes caries in teeth, but if ignored at initial stage, caries can harm more neighboring teeth and go deeper inside the teeth till the pulp inside the teeth which can cause severe toothache.

Dentists normally treat their patients with caries by removing the radiographs (x-ray) of their tooth jaw and spot the cavities by observing the x-ray with naked eyes. This could lead to miss few cavities which are at early stage or which are not properly recognizable. Hence, there is a need of automation due to many reasons such as, lack of dental experts, different levels of expertise in each dentist and a second opinion is always helpful for confirm decision making.

This paper propose a novel technique to use combination of various image processing techniques, like RGB to Gray conversion, generation of binary image, finding the region of interest, removing background, identifying regions and dividing image into various blocks to extract the exact cavities from the x-rays captured from patients tooth jaw and also speed up the operation. To diagnose the caries in teeth, best way of caries scanning is using an x-ray; Dental x-ray clearly shows entire details of mouth which cannot be recognized while examining visually like hidden dental structures and bone loss [2]. The cavities in teeth can be identified uniquely from x-ray images because; discontinuity in the brightness of an x-ray corresponds to discontinuity in depth, surface orientation, material property and variation in scene illumination. Cavity properties are different than properties of a healthy tooth [3]. The x-rays need to undergo required image processing techniques. The captured x-ray is in jpeg format thus, it is in RGB format of color code. The image is then converted to gray color code where the pixel values of image are not changed. After converting to gray color, feature enhancement is applied to highlight the affected part (cavity in teeth) followed by background removal and then region and block analysis and finally cariesdetection.

The entire paper is organized in stepwise format first part is introduction about the cavities and how to overcome it using Image Processing. Section 2 covers, the analysis performed on cavities handling. Sections 3 elaborate more on the working of the proposed work. Section 4 concludes the proposed concept and suggests the future work possible.

2. ANALYSIS

Dental caries is a disease for which the dentist needs a final confirmation on existence of cavity in his patients' teeth. Scanning of x-ray is the important part on which the final decision is executed. X-ray scanning needs to use various image processing techniques. Image processing consists of a collection of huge methods and algorithms.

One of the treatments used for this cause is proposed by, Villette, A., et al. In their study, they describe a new device for the rapid and painless injection of a calcium hydroxide paste under controlled pressure in the tooth pulp cavity with a controlled volume of the paste. They observed that the reduction of the volume of the pulp cavity and that the extent of the tooth devitalization depend on the amount of the paste injected [4].

There is also a technique that is making use of ultrasound concept. The researchers have done experimental evaluation of human teeth using noninvasive ultrasound. It is verified that these experimental measurements confirm predictions reported in earlier finite element and transmission line studies and suggest that ultrasound is a tool is helpful and can result in better improvement in current dental imaging systems [5].

Other than cavity detection, leaf disease detection also depends on captured leaf images. Thus, it makes use of image processing. Francis, Jobin, and B. K. Anoop have proposed an algorithm for detection and identification of leaf disease using image processing. The pepper plant leaves are used here, these leaves undergo image segmentation process in two phases, masking process and threshold based segmentation.

This algorithm produces better results in terms of differentiating healthy and unhealthy leaves [6].

Dandawate Yogesh and Radha Kokare have Used support vector machine to classify soybean leaves as healthy or diseased [7].

Thus, image processing plays an important role in performing cavity detection along with the supporting image processing techniques like image conversion, preprocessing,

Segmentation, region identification, block division.

3. WORKING

The proposed model for cavity detection using Image Processing is explained below with the following block diagram. Fig. 1 illustrates the working of the proposed model. Each block from this diagram is elaborated further with proper explanation.

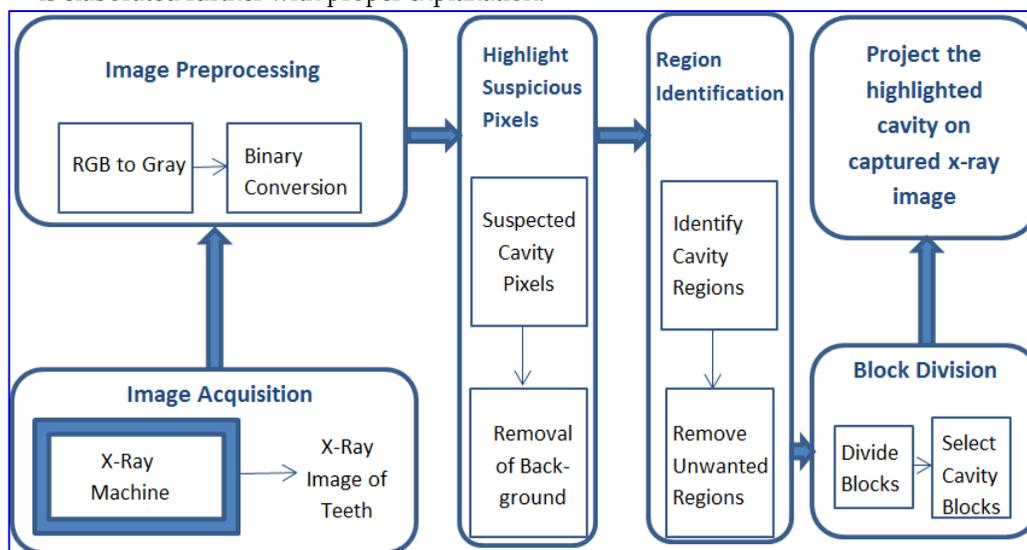


Figure 1: Block diagram of proposed work

3.1 Image Acquisition

The dentists normally don't use any tools for finding the cavities in teeth. Caries are not seen easily with eyes; It can be recognized effectively only with the use of x-ray images of teeth. In x-ray, a healthy tooth is normally seen in white color and other area such as soft tissue, jaw and mouth are seen in dark shades. The cavities and caries are also seen with specific shades and structure. X-ray images when captured, they might have lots of different shades from bright to dark formed due to light at x-ray center or the position of teeth from x-ray plate. Digital images can be affected by various types of noise [8]. Various types of noise found are Gaussian noise, salt and pepper noise, white noise, etc. [9]. For better quality of images that is avoiding noise; x-ray should be captured with proper care and attention. Below Fig. 2 is the original x-ray image used as an input for the model.



Figure 2: Original x-ray image of human teeth in RGB format

3.2 Image Preprocessing

Image Preprocessing is a collection of operations performed for enhancing the image data and removing unwanted noise and background data. Following subsections explain image pre-processing steps in detail.

3.2.1 RGB to Gray Conversion

The captured x-ray images are saved in jpeg format. These images are very similar to gray images; that is, almost all pixels in entire image have all three colors Red, Green and Blue hold same intensity value. The x-ray images are thus converted to gray images for saving storage space and processing speed.

Above figure 2 after converting from RGB to Gray is shown as below in Fig. 3.



Figure 3: Gray scale image of teeth

3.2.2 Binary Conversion

Gray color x-ray images are converted to binary images. A binary image has basically two colors only, black and white. This binary image is then compared with gray image of same sample and the black portion of image is trimmed down from original gray color x-ray image. Following Fig. 4 shows the binary image of above image shown in Fig. 3.

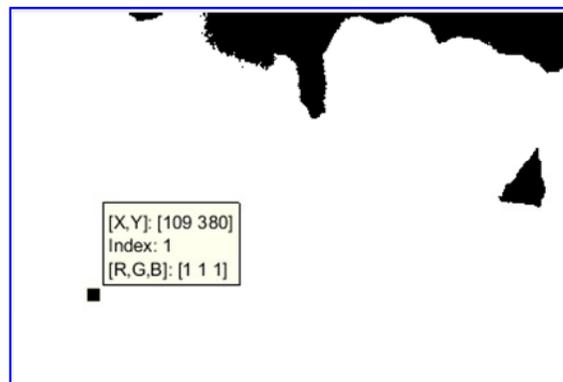


Figure 4: Binary image of teeth

3.3 Highlight Suspicious Cavity Pixels

Image preprocessing step is followed by the classification step; in this step, the pixel values of cavities are accepted from dataset of dental x-ray images.

3.3.1 Suspected Cavity Pixels

The pixel's values are classified in two classes one is named 'class 1' (indicates class of cavity pixels) and remaining pixels are classified under 'class 2' (indicates class of non- cavity pixels). The image after highlighting the suspected cavities is given as follows in Fig.5.



Figure 5: Suspected Cavities identified in range

3.3.2 Removal of Background

The cavity pixel color and background colors falls in nearly same range. Removal of background part from image, indirectly ease the process of identifying the cavities present in x-ray. The human teeth are selected as region of interest (ROI) very first. The process of extracting ROI from given x- ray is followed by generation of binary image. The above image in figure- is displaying cavities, but here some part of the background is wrongly shown as cavity. The background from the above image in Fig. 5 is removed with the use of binary image shown in Fig. 4; the resultant image is displayed below in Fig.6.

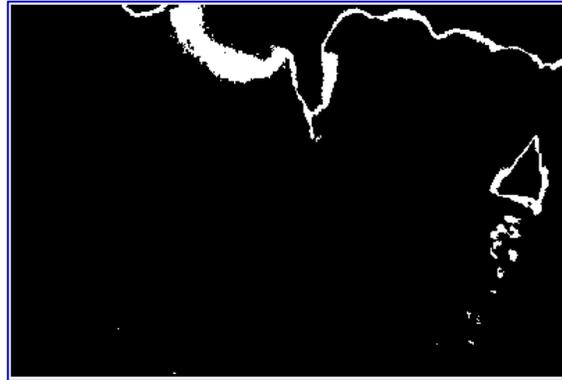


Figure 6: Suspected Cavities after removing background

3.4 Region Identification

3.4.1 Identify Cavity Regions

The cavity pixels highlighted in the image shown in Fig. 7 are represented in regions with proper outline. These regions are further examined and the regions centroid and circular area is observed.

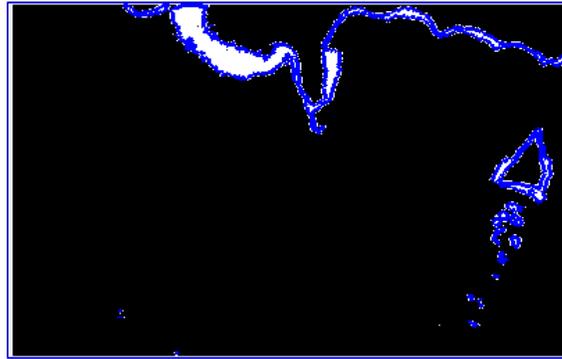


Figure 7: Cavities represented with regions

3.4.2 Remove Unwanted Regions

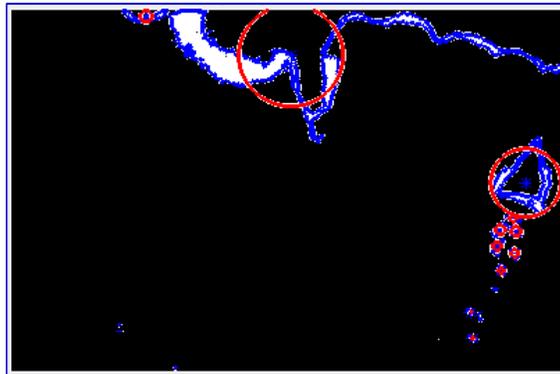


Figure 8: Centroids of cavity regions after removing unwanted regions

All the identified regions are further examined and then by making use of Major Axis Length and Minor Axis Length, some of the regions are removed which doesn't fall in category of true cavity regions.

Following Fig. 8 displays the selected regions with their centroids.

Highlighted color. The output image generated by this model is very useful for dentists to easily identify the cavity existing in the x-ray of teeth.

Below image in Fig. 10 is the original x-ray image which is used as an input to the model.



Figure 10: Original X-ray image

Following image in Fig. 11 is the final resultant image which shows cavity portion in highlighted area that is helpful for dentists to easily identify the cavity from an x-ray image.

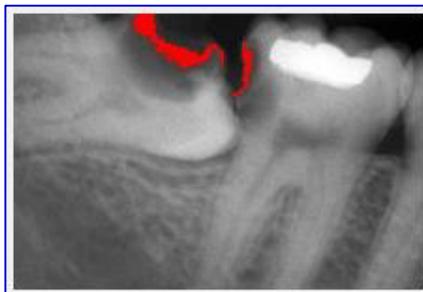


Figure 11: Final result showing cavities on original X-ray image

3.5 Block Division

3.5.1 Divide image into Blocks

The cavity regions identified till now in Fig. 8 are long enough, which are not so easy to evaluate to identify the true cavity. The image is further divided into blocks of similar sizes (50 X 50pixels).

3.5.2 Select Cavity Blocks

The blocks are now cross checked to find the number of cavity pixels in each of them. Here the blocks which are showing false cavity that is edge of teeth; are eliminated and only true cavity in which number of cavity pixels are large in numbers are selected.



Figure 9: Selected blocks after removing unwanted blocks

Above Fig. 9 displays the blocks that are selected as the block with true cavity region. In these blocks, the red region is representing cavity area and other portion is representing healthy area.

3.6 Output

The selected blocks with true cavity are now used to generate final output image, in which the cavity pixels are shown with

4. CONCLUSION

Dental cavity could be a widespread malady that affects many folks worldwide. Because of human dependency the tooth decay will generally be unnoticed. This paper proposes a model to simply determine the cavities in teeth by availing the potential of Image process. several of the Image process functions used are RGB to grey conversion, generation of binary image, finding the region of interest, removing background, distinctive regions and dividing image into same sized blocks and eventually turning out with distinctive the cavities gift in x-ray image with correct position. In future, cavity detection work will be any jury-rigged by victimization the ultimate known blocks as input to ANN model with Backpropagation rule. Supervised learning of ANN helps to come up with patterns by itself from coaching dataset. The trained ANN will be later tested and valid. Finally the model will be used for brand spanking new unseen samples of dental x-ray to observe the cavities.

5. REFERENCES

- [1] <http://www.who.int/mediacentre/factsheets/fs318/en/> World Health Organization.
- [2] A. K. Jain and H. Chen, "Matching of dental x-ray images for human identification", *Pattern Recognition*, 37:1519–1532,2004.
- [3] Wang, X.Q. Ye, W.K. Gu, "Training a Neural Network for Moment Based Image Edge Detection" *Journal of Zhejiang University SCIENCE*(ISSN 1009-3095, Monthly), Vol.1, No.4, pp. 398-401 CHINA,2000.
- [4] Villette, A., et al. "Pulp tissue response to partial filling of the pulp cavity, under compression, by calcium hydroxide, using a new device." *Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE*. Vol. 3. IEEE,1992.
- [5] Ghorayeb, Sleiman R., and Teresa Valle. "Experimental evaluation of human teeth using noninvasive ultrasound: echodentography." *ieee transactions on ultrasonics, ferroelectrics, and frequency control* 49.10 (2002): 1437- 1443.
- [6] Francis, Jobin, and B. K. Anoop. "Identification of leaf diseases in pepper plants using soft computing techniques." *Emerging Devices and Smart Systems (ICEDSS), Conference on. IEEE,2016*.
- [7] Dandawate, Yogesh, and Radha Kokare. "An automated approach for classification of plant diseases towards development of futuristic Decision Support System in Indian perspective." *Advances in Computing*,

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 2015.

- [8] K. M. Rao, Deputy Director, NRSA Hyderabad, "Overview of image processing", Readings in Image processing.
- [9] Gonzalviz R. C., Woods R. E., "Digital Image Processing", Pearson Publications, 2009.
- [10] Abdolvahab Ehsani Rad, Mohd Shafry Mohd Rahim, Amjad Rehman, and Tanzila Saba. Digital dental x-ray database for caries screening. *3D Research*, 7(2):1–5, 2016.
- [11] Rad, A. E., Mohd Rahim, M. S., Rehman, A., Altameem, A., & Saba, T. (2013). Evaluation of current dental radiographs segmentation approaches in computer-aided applications. *IETE Technical Review*, 30(3), 210-222.
- [12] Abdolvahab Ehsani Rad, Mohd Shafry Mohd Rahim, Hoshang Kolivand, and Ismail Bin Mat Amin. Morphological region-based initial contour algorithm for level set methods in image segmentation. *Multimedia Tools and Applications*, pages 1–17, 2016.
- [13] Rad, A. E., Amin, I. B. M., Rahim, M. S. M., & Kolivand, H. (2015). Computer-Aided Dental Caries Detection System from X-Ray Images. In *Computational Intelligence in Information Systems* (pp. 233-243). Springer International Publishing.

Observational Analysis for Quality of Soil Organic Matter Using Agricultural Data Science

Gulledmath Sangayya

School of Computer Science & Application
Research Scholar-Reva University Bengaluru
gsswamy@gmail.com

Dr.Arul Kumar V

School of Computer Science & Application
Asistant Professor-Reva University Bengaluru
arul Kumar.v@reva.edu.in

Abstract—Agriculture is the very basic, needed and ever demanding functional process to achieve food requirement all over the world; it is a backbone of our economy specially racing towards developed countries like India in comparison with other countries. The application driven and usage of Data science techniques in agriculture concern on soils can improve the situation of need-based decision for cultivation and supports the yields or growth of agricultural products in a better way. The critical analysis of soils and soil organic matter [SOM] plays an important role for decision making on several issues related to agriculture domain and other relevant applications. This brief experimental research presents about the role of data science which plays progressive thought process of soil analysis in the field of agriculture and also solicitation about several data science techniques and their related work by recent investigation by several authors in context to soil analysis domain. The data science under data mining techniques are of very updated phenomenon in the area of soil analysis considering in the field of finding organic matters of soil. Soil organic matter which is generally consists of decomposing plant and animal residues. In addition to these materials, soil organic matter contains living along the line of dead microbial cells, microbial synthetically driven compounds and a number of substitute derivatives produced as a result of microbial activity in the soil. It is the store house of all essential plant nutrients. Let's emphasize the importance of Data Mining Techniques under categorical process of data science to find the organic matter of soils.

Keywords—Agricultural, Data Science, Data Mining, Soil, Classification, Mining techniques and Algorithms

I. INTRODUCTION

In technology driven society, data mining is used as a major areas of application which is fast trending and many open source data mining tools, algorithms, techniques and utility based procedures are freely available as open source to access and implement in customised manner, which influence the developers and end users to adopt and deploy for their requirement without much efforts, but data science or data mining in agricultural and use of soil data tables generally referred as datasets is a comparatively a recent research field like new born baby. In future proof or current scenario data mining concept and techniques used to solve the many agriculture problems as benchmark solution. In this brief research work of observational analysis it has been generally discussed about how data mining techniques are used and applied comprehensively in agriculture field to know the facts of organic matters of soils. In comparison of other needs, the whole world the requirement or essential of need of food is become important entity;

hence the agricultural data scientists, farmers of local yields, government, and researchers of domain are attentive and putting their extra efforts and uses numerous techniques or procedures in agriculture for improvement in production of good yields. With this initiative, the data generated in various structured and unstructured in the field of agricultural science as we know data increasing and storing as repository with local server which is regular affair. As a result the data escalates continuously, it requires various ways for these data or datasets to be mined and analysed when needed with adopting systematic and scientific applications. The current scenario, a few farmers who are technically trained by local government initiative schemes for empowerment are really using the new methods, algorithms, tools and techniques in agriculture for better production and achieving good yields. Data science or generally mining can be used for visualising the future needs or trends of agricultural processes [1]. Data science or Data mining in totality is a guiding principle in the direction of using technology that results in the developing new algorithms and applications using large data sets or source of data. The main inherited property of the data mining process is to extract knowledge from an existing data sets either structured or unstructured and culminate it into a human understanding nature form for advance use to achieve the effective results. Data mining or broader sense we call as data science which is the common approach of analyzing data from different viewpoints and makes it abstract or summarization of useful information in any format. Data mining can analyse into readable data there is absolutely no restriction or objection on the type of data [2]. Data mining generally classified or categorised into two broad areas such that one is descriptive nature and another one is predictive nature. Descriptive data mining means considering the existing features of data which is regular appeal of data, that is raw data and then make it summarization report for knowledge expression. The descriptive mining leads the representation of various characteristics of the past events and which further mapped to influence the future of needs. The foundation of predictive nature mining depends on probabilities and some of know statistical approaches, it is used to predict future based on the values considered from known results or pre-processed data. Forecasting generally uses certain parameters which involves using the variables or field in the database to estimate unknown results [3].

Meaning and origin of soil matter

Meaning of Soil Organic Matter

Generally, it preserves the soils against various erosion or termed as preservative of

erosion and helps to form good nature or quality of soil structure. It provides natural aeration and better water movement or flow by loosening the soils or scattered granules. For achieving more benefit to the soil, organic matter or content must be decomposed and continuously refurbishing with adding of fresh organic materials to the soil. After active decomposition of soil of organic residues are collectively known as humus which in another turn referred as organic soil colloid. Most effective and dominant micro organisms are involved in the decomposition of organic matter namely bacteria actinomycetes and fungi. Other soil enzymes mean protein substances produced by these micro-organisms, are directly responsible for the decomposition by reducing the activation energy necessary to break the bonds of different materials.

Origin of Soil Matter

The soil organic matter termed as SOM usually originates from the plant tissue. The base for generation is leaves and roots of trees, palatial shrubs grasses and other plants. It usually supplies large quantity of organic materials or compounded organic flavour to the soil. These plant in another way orient and constituents form for the primary material or initial source both for the food of the soil organisms and for the producing the organic matter. In general animals are considered secondary sources of organic matter of elements. As a matter of fact, they originally break down the plant tissues or generic tissues, they contribute enormous waste products and leave their pedestal own bodies after death. If we consider other forms of animal life of longevity, particularly earthworms, centipedes, insects and ants, also plays an important and pivotal role in the turnover of plant residues.



Fig. A. Preparation of soil for testing



Fig B. Optional Layering before testing.

Nature and Composition of Soil organic matter

The soil organic matter consists of whole series or residual of products which begins from undecomposed plant or leaves and animal tissues to fairly amorphous brown or tickly brown to black material which identifies no traces of anatomical or noted structure of the material in other words its been referred as "Soil Humus". The native part of the soil organic matter is made up of heterogeneous mixture of polymerized aromatic molecules, polysaccharides, bound amino acids, uranic acid polymers and various organic phosphorus compounds. Organic residues consist of both organic and inorganic materials of the fractions.

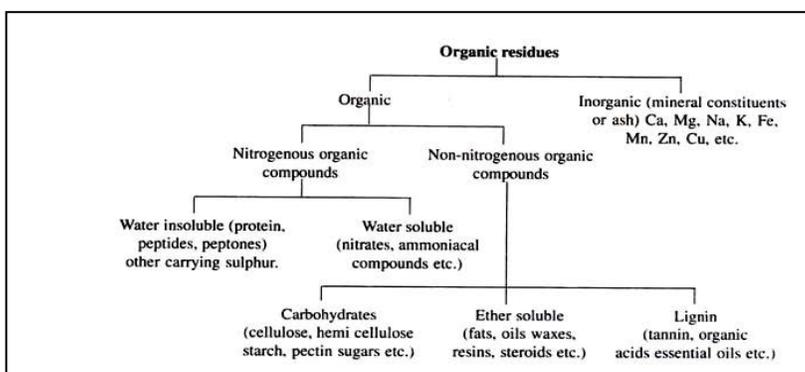


Figure 1.1 Simple outline for the presence of different compounds in organic residues [Image source: www.agrosciencejournal.com]

Karnataka forms southwestern part of Deccan peninsular India. It lies between 11° 30' and 18° 30' N latitude and 74° 55' and 78° 30' E longitudes with a geographical area of 1.91 lakh square kilometres. It experiences or foot holds a wide variety or nature of geological, climate, vegetation or general form of vegetation and physiographic conditions. As a result of deeds, the soils of Karnataka are highly diverse and variable depending upon these pivotal conditions. The northern part of the state is occupied by shallow to deep black soils. Southern part of the state is occupied by red loamy soils, whereas the heavy rainfall regions comprising Western Ghats and coastal districts are occupied by coastal alluvial and lateritic soils. With these varied agricultural climatic conditions and diverse nature of soil types the state is suitable or matching for cultivation of large variety of crops.



Fig. C. Loamy and Sandy soil for testing



Fig D. Mixed Soil for finding SOM.

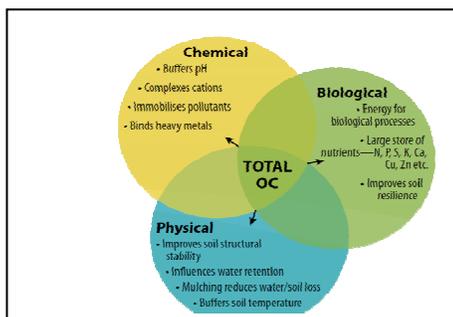


Figure 1.2 General Representation of Chemical, Biological and Physical properties of Soil [Image Source:krishikosh.egranth.ac.in]

Bangalore Urban district is located in the southeastern part of Karnataka. It is having an extent of 2174 sq.km and is located between the north latitude 12°39' 32": 13°14' 13" and East longitude 77°19'44": 77°50'13". The district is bounded in all the directions by Bangalore rural district except in southeast, where the district is bounded by Dharmapuri district of Tamil Nadu state. The soils of the districts can be broadly grouped into red loamy soil and lateritic soil. Red loamy soils generally occur on hilly to undulating land slope on granite and gneissic terrain. It is mainly seen in the eastern and southern parts of Bangalore north and south taluks. Laterite soils occur on undulating terrain forming plain to gently sloping topography of peninsular gneissic region. It is mainly covered in Anekal taluk and western parts of Bangalore North and south taluks.

review of literature

Various articles and journals and many people contributed in finding the facts about soil organic matter are reviewed here under the following headings.

3.1 collective opinion about soil chemical properties in the selected districts of Karnataka.

3.2 Statistical assessment and influence of soil properties on the nutrient status in the selected districts of Karnataka.

3.1 collective opinion about soil chemical properties in the selected districts of Karnataka.

Nayak *et al.* (2000) showed that, the available iron, copper, manganese content in alluvial soils of Arunchal Pradesh ranged from 7.0 to 73.6, 0.50 to 3.5, 2.7 to 53.3 mg kg⁻¹ respectively in surface soils.

Gowda *et al.* (2001) showed that the Mineral (calcium, phosphorus, magnesium, copper, zinc and iron) status in soils of coastal zone of Karnataka. Extractable Ca (0.15%) and Mg (0.02%) level in soil was slightly higher than the critical level but the levels of P, Cu, Zn and Fe in soil was much higher than the critical levels.

Korikanthimath *et al.* (2002) were collected samples from coffee and cardamom growing areas in Kodagu district, Karnataka to assess the nutrient status. The results revealed that, the soils were acidic in reaction (pH 4.2 to 6.9) with high organic carbon content (0.19 to 3.3%). The available phosphorus was found to be low due to acidic pH that results in formation of aluminium and iron phosphate which are not likely to be readily available to plants. The soils were rich in potassium (84 to 966 kg/ha) and adequate in micronutrients due to the increased solubility in acidic environment and greater recycling by vegetation.

The micronutrient content, *viz.*, Calcium (Ca), Phosphorus (P), Magnesium (Mg), Copper (Cu), Zinc (Zn) and Iron (Fe) content in soil, in northern dry and northern transition zones of Karnataka were analyzed by Ramana *et al.* (2002). The phosphorus content in soil in Northern dry zone (30.30 ppm) and Northern transition zone (34.33 ppm) were found to be well below the critical level (45 to 130 ppm). The calcium (0.47%), copper (4.17 ppm) and zinc (11.27ppm) content in soil in northern transition zone were found to be higher than in the northern dry zone and were well above the critical levels in both the zones.

Mustapha *et al.* (2007) studied that, soil fertility status of Bauchi State, Nigeria. The result indicated that, means of B and Zn were 0.41 and 1.46 mg kg⁻¹ respectively. The B and Zn varied widely (CV >30%) between locations in the State, irrespective of the parent material.

Ravikumar *et al.* (2007) reported that the micro nutrients status of Malaprabha Right Bank Command of Karnataka for site specific recommendations. The mapping of available micro nutrient status indicated that the majority of the area was deficient in iron (0.18 to 3.51 mg/kg) and zinc (0.01 to 0.37 mg/kg), whereas manganese (0.014 to 11.38 mg/kg) and copper (0.016 to 6.78 mg/kg) status was low to sufficient.

Sarwar *et al.* (2007) reported that the micronutrients status of soils of district Palandria, Azad Kashmir and to correlate the micronutrients with Physicochemical characteristics of soil as well as to categorize the soils as high, medium and low in Cu, Zn, Fe, Mn and HWS B. The samples were collected from thirty wheat fields and thirty apple orchard soils during 2003. None of the soil sample was low in Cu, Mn and Fe contents. Zn was low in 20 per cent wheat fields and 17 per cent apple orchard soils while HWS B was low in 53 and 40 per cent in wheat and orchard soils, respectively.

Bassey *et al.* (2008) showed that, the Iron (Fe), manganese (Mn), zinc (Zn), copper (Cu) content had a mean value of 142.62, 7.43, 8.50, 4.30 mg/kg for inland depression soils and 213.16, 3.19, 2.67, 2.20 mg/kg flood plain soils, respectively. All the above values were higher than the critical values of 4.2, 0.2, 0.5 and 1.0 mg/kg for Fe, Cu, Zn and Mn respectively.

3.2 Statistical assessment and influence of soil properties on the nutrient status in the selected districts of Karnataka.

Correlation coefficient analysis helps to determine the nature and degree of relationship between any two measurable characters.

Nayak *et al.* (2000) revealed that, Fe was negatively and significantly correlated with pH and sand. Positive and significant correlation was also observed with organic carbon ($r = 0.48$), silt ($r = 0.64$) and CEC ($r = 0.54$). The copper was significantly and positively correlated with pH ($r = 0.51$) and silt ($r = 0.49$) but it showed negative and non-significant correlation with sand and clay. Correlation studied revealed that, significant and positive correlation was found between available Mn and organic carbon, silt and CEC but showed negative correlation with pH and sand.

Patiram *et al.* (2000) reported that, Zinc was positive significant correlated with organic matter. The positive correlation may be due to the formation of organic complexes between organic matter and Zinc that protect it from leaching.

Singh *et al.* (2008) reported that, available manganese content in Entisols of Uttar Pradesh ranged from 0.4 to 0.8 mg kg⁻¹ of soil. It significantly and positively correlated with organic carbon (0.15**) and negatively correlated with pH (-0.20). The Zn, Cu, Fe, Mn and B showed positive correlations with silt plus clay and organic carbon, and negative correlations with pH and calcium carbonate content (Sharma *et al.*, 2003).

Nazif *et al.* (2006) showed that, extractable iron and manganese are negative significant correlation with soil pH and lime content, Iron was positively and significantly correlated with silt. Copper, Zinc and hot water-soluble Boron were positively significantly correlated with organic matter. Both Iron and hot water-soluble Boron gave negative significant correlation with sand. Other physio-chemical properties of soil showed either negative or positive non-significant correlation with micronutrient during the study.

Nazif *et al.* (2006) studied the micronutrient status of soils of district Bhimber (Azad Jammu and Kashmir) and reported that the Iron, Copper, Zinc and Manganese ranged from 5.37-23.36, 0.59-4.38, 0.74-2.08 and 4.59-21.08 mg kg⁻¹ respectively. Iron, Copper and Manganese was found high in all sites while Zinc was low in 26.66 per cent, medium in 70 per cent and high in 3.34 per cent sites. Boron ranged from 0.02-0.84 mg kg⁻¹ and Boron was found low in 80 per cent and medium in 20 per cent sites.

Kelin *et al.* (2007) studied that, spatio-temporal variability of soil organic matter (SOM) in the urban–rural transition zone of Beijing. SOM content in agricultural soils were measured in 1980, 1990 and 2000 in Daxing County of Beijing in-situ and data of 1980 and 1990 were obtained from the National Soil Survey (NSS). Descriptive statistics and geostatistics were used to analyze the data. The results showed that, mean SOM was 9.95 g kg⁻¹ in 1980, 12.76 g kg⁻¹ in 1990 and 12.89 g kg⁻¹ in 2000. SOM was spatially correlated at a larger distance of 32.0 km in the E–W direction for the three years, and at a shorter distance of 24.6, 23.3 and 19.0 km in the N–S direction in 1980, 1990 and 2000 respectively, which showed that there was more variability in SOM in the N–S areas across the period of 20 years. The SOM slightly increased from low to high levels from 1980 to 2000. The main factors affecting SOM levels were the soil texture, land use and farming practices. The increasing trend might be attributed to the widespread practices of mulching and organic manure applications.

collection of data

Nature and Source of Secondary Data

The taluk wise secondary data on the soil parameters *viz.*, pH, EC, OC, P, K, Zn, Fe, Mn and B content of Bangalore Rural, Bangalore Urban and Ramanagaram districts for a period of six years from 2007-2013 totally hundred samples from each taluk per year (Random selection) and District wise, per cent deficiency (Zn, Cu, Fe, Mn and B) for a period sixteen years from 1998-2013 were collected from Agricultural Department. Sample of Field attribute is indicated as follows.

Field (Attribute)	Description
N	Nitrogen
P	Phosphorous
K	Potassium
Fe	Iron content in the soil
Mn	Manganese content in the soil
Zn	Zinc content in the soil
Cu	Copper content in the soil

data mining techniques applied

The main techniques for data mining include Classification, Clustering, Association rules and Regression. The different data mining techniques used for solving different agricultural problem discussed by Mucherino, A., Papajorgji, P., & Pardalos, P. (2009). The graphical representation of different data mining techniques is shown in Figure 5.1. [4]

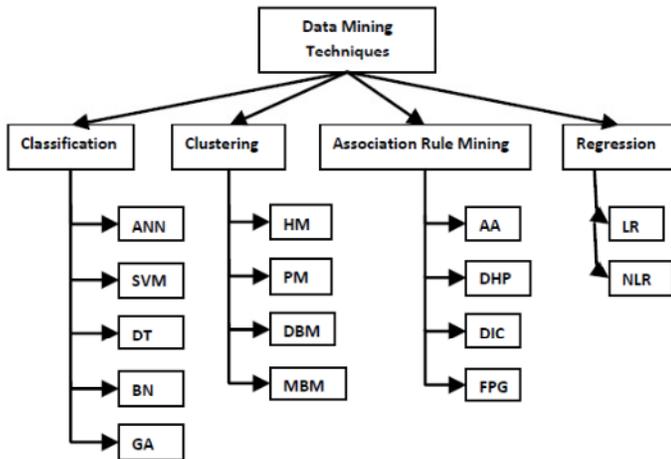


Figure 5.1: Different data mining techniques.

Classification: Classification and Prediction are of two major forms of data analysis or pattern of analysis that can be used or illustrated to extract models describing important data classes or models to predict future data trends or nature of future predictive data. It is a process or phenomenon in which a model learns or behaves to predict a class label from a set of training data which can then be used to predict discrete class labels on new data samples. To increase the marginal predictive accuracy obtained by the different classification model when classifying model in the test set unseen during training is one of the major satisfied condition of classification algorithm. Data mining classification algorithms can follow three different learning approaches: supervised learning, unsupervised learning, or semi-supervised learning. As we know various classification techniques for discovering knowledge for representation are Rule Based Classifiers, Bayesian Networks (BN), Nearest Neighbour (NN), Decision Tree (DT), Artificial Neural Network (ANN), Rough Sets, Support Vector Machine (SVM), Fuzzy Logic, Genetic Algorithms. [5] and many more in experimental domain.

Clustering: In clustering, the focus is on finding a partition of data records into clusters such that the points within each cluster are close to one another. Clustering groups the data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. Since we know the goal of clustering is to discover always new set of categories, the new groups are of phenomenal interest in themselves, and their assessment is intrinsic. [6] There is no prior knowledge about data. The different clustering methods are Hierarchical Methods (HM), Density-based Methods (DBM), Partitioning Methods (PM), Model-based Clustering Methods (MBCM), Grid-based Methods and Soft-computing Methods such as [fuzzy, neural network based], Squared Error–Based Clustering (Vector Quantization), Clustering graph and network data etc.. [1][7][8]

Association Rule Mining: The technique of discovering association rules was originated by Agrawal, Imielinski, & Swami in 1993.[10] Association rule mining technique is one of the most efficient techniques of data mining to find unseen or expected pattern among the large amount of data. In this technical approach, the focus is on finding relationships among the different items in a transactional database. Association rules generally deals to find out elements that co-occur repeatedly or recursively within a dataset consisting of many independent selections of elements (such as purchasing transactions or buying nature), and to discover rules to frame pattern. The simple problem statement to solicit the curiosity is: Given or selected set of transactions, where each transaction or record is a set of literals (called items or in weka we refer as labels), an association rule is an expression of the form X tending to Y as $X \Rightarrow Y$, where X and Y are sets of items or labels. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y . [9] An application of the association rules mining is the market basket analysis, customer segmentation, catalog design, store layout and telecommunication alarm prediction.[11] The different association rule mining algorithm are Apriori Algorithm(AA), Partition, Dynamic Hashing and Pruning(DHP), Dynamic Itemset Counting(DIC), FP Growth(FPG), SEAR, Spear, Eclat & Declat, MaxEclat.[11]

Regression: Regression fundamentally is learning a function that maps a data item or labels to a real-valued prediction variable. The different applications or usability of regression are predicting the quantity of biomass present in a forest, find the probability of patient will survive or not on the labelled sets of his diagnostic tests in series, predicting consumer demand for a new product. [3] Here the model is trained to predict a continuous target. Regression tasks are often treated as classification tasks with quantitative class labels. The methods for prediction are Linear Regression (LR) and Nonlinear Regression (NLR). There are different are conducted which have been carried out on the application of data mining techniques or using data science for agricultural data sets. Naive Bayes Data Mining Technique or approach is used to classify soils that analyze large soil profile experimental datasets. [12] Decision tree algorithm in data mining is used for predicting soil fertility. [13] By studying and analyzing clustering techniques (Based on Partitioning Algorithms or Hierarchical Algorithms) we examine the current usage and details of agriculture land vanished in the past seven years or may be decade. The overall aim of the research was to determine the land utilization for agriculture and non-agriculture areas for the past ten years. [14] D Ramesh [15] used k-means approach to estimate the crop yield

analysis. Some data mining methodology which are used in agricultural domain are reviewed by author Vamanan, R, & Ramar, K [16]

result and discussion

J48 is an openly available free source Java implementation of the C4.5 algorithm in the Weka data mining tool for experimental prediction. C4.5 is a program that creates a decision tree based on a set of labelled input data. This decision tree can then be tested against unseen labelled test data to quantify how well it generalizes. This algorithm was developed by Ross Quinlan. It is absolutely an extension of Quinlan's earlier ID3 algorithm. C4.5 uses ID3 algorithm that accounts for various comparison of data sets.

Classifier	Naïve Bayes	JRip	J48
Correctly Classified Instances	855	1998	2065
Incorrectly Classified Instances	1345	202	135
Accuracy	38.86%	90.81%	93.86%
Mean Absolute Error	0.324	0.0313	0.0283

Using regression algorithms like Linear Regression, Least Median Square, Simple Regression different attributes were predicted. According to these

results the values of Phosphorous attribute was found to be most accurately predicted and it depends on least number of attributes. When all attributes are numeric, linear regression is a natural and simple technique to consider for numeric prediction, but it suffers from disadvantage of linearity. If data exhibits non-linear dependency, it may not give good results. In this case, least median square technique is used. Median regression

techniques incur high computational cost which often makes them infeasible for practical problems [8]. Several regression tests were carried out using WEKA data mining tool to predict untested numeric attributes.



Fig. E. Compost soil for SOM measurements

measurements

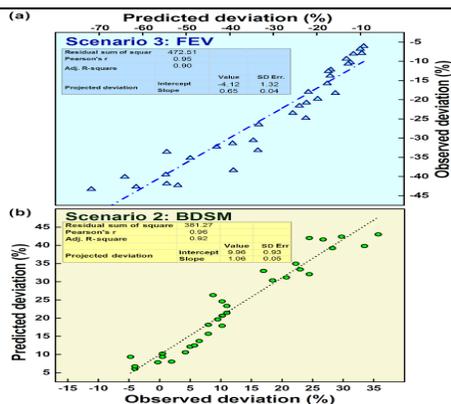


Fig F. Scaled findings of predicted deviation.

Algorithm	Linear Regression	Least Median Square Regression
Time taken to build the model	0.18 s	10.84 s
Relative Absolute Error	10.63%	10.08%
Correlation Coefficient	0.9710	0.9905

Table 3. Comparisons of Regressions Algorithms

Linear-Regression test for predicting phosphor gave the best and accurate results. These predictions can be used to find out phosphor content without taking traditional chemical tests in soil testing labs, and this will eventually save a lot of time. Statistical results of these tests are given in Table3. There were very limited variations amongst the predicted values of phosphor attribute. Though the Least Median of Squares algorithms is known to produce better results, we noticed that the accuracy of linear regression was relatively equivalent to that of least median of squares algorithm.

Table-4: Predictions on test data

Actual Value Using Soil Testing	Predicted Value Using Linear Regression	Error
10.3	10.661	0.361
7.7	7.431	-0.269
4.6	4.653	0.053
9.5	8.478	-1.022
2.9	3.035	0.135
5.1	4.915	-0.185
15.3	15.667	0.367
7	7.402	0.402
18.4	18.743	0.343
4.4	4.388	-0.012
13.5	13.438	-0.062

Here the Relative Absolute Error is nearly same for both the prediction algorithms. Even though Least Median Square regression gives better numeric predictions but the time taken to build the model is 67 times that of Linear Regression, hence computational cost used by Linear Regression is much lower than that of least median square technique.

Table 4.2.1: Inter-relationship between soil properties and nutrients of selected areas

Nutrients	Soil properties		
	pH	EC	OC
P	0.536	0.675*	0.817**
K	0.570	0.863**	0.878**
Zn	0.340	0.710*	0.918**
Cu	0.488	0.669*	0.814**
Mn	0.463	0.847**	0.890**
Fe	0.084	0.279 ^{NS}	0.661*
B	0.381	0.565 ^{NS}	0.690*

** Significant at 1 % level, * Significant at 5 % level
NS: Non -Significant

conclusion

In this paper, we have proposed an analysis of the soil data using different algorithms and prediction technique. In spite the fact that the least median squares regression is known to produce better results than the classical linear regression technique, from the given set of attributes, the most accurately predicted attribute was “N” (Nitrogen content of the soil) and which was determined using the Linear Regression technique in lesser time as compared to Least

Median Squares Regression. We have demonstrated a comparative study of various classification algorithms i.e. Naïve Bayes, J48 (C4.5), JRip with the help of data mining tool WEKA. J48 is very simple classifier to decide tree, but it gave the best result in the experiment. In future, we can plan to build Fertilizer Recommendation System which can be utilized effectively by the Soil Testing Laboratories. This System will recommend appropriate fertilizer for the given soil sample and cropping pattern.

References

- [1] Han, J, Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.
- [2] <http://www.publishyourarticles.net/knowledge-hub/essay/essay-on-the-importance-of-agriculture-in-the-indian-economy.html>
- [3] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- [4] Mucherino, A., Papajorgji, P., & Pardalos, P. (2009). Data mining in agriculture (Vol. 34). Springer.
- [5] Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. *International Journal of Engineering Research & Technology (IJERT)*, 1(6).
- [6] Lior Rokach, Oded Maimon. Clustering Methods. Chap-15
- [7] Xu, R & Wunsch, D (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3), 645-678.
- [8] Periklis Andritsos Data Clustering Techniques. University of Toronto, Department of Computer Science. <ftp://ftp.cs.toronto.edu/csrg-technical-reports/443/depth.pdf>
- [9] Srikant, R V Q & Agrawal, R (1997, August). Mining Association Rules with Item Constraints. In *KDD (Vol. 97, pp. 67-73)*.
- [10] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216)*. ACM.
- [11] Zaki, M J (1999). Parallel and distributed association mining: A survey. *IEEE concurrency*, 7(4), 14-25.

- [12] Bhargavi, P, & Jyothi, S. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. *International journal of computer science and network security*, 9(8), 117-122.
- [13] Jay Gholap. (2012). Performance tuning of j48 algorithm for prediction of soil fertility. *Asian Journal of Computer Science And Information Technology* 2: 8 (2012) 251– 252.
- [14] Megala, S., & Hemalatha, M. (2011). A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu. *International Journal of Computer Applications*, 23.
- [15] D Ramesh, B Vishnu Vardhan, (2013). Data Mining Techniques and Applications to Agricultural Yield Data. *International Journal of Advanced Research in Computer and Communication Engineering* 2(9).
- [16] V. Ramesh and K. Ramar, 2011. Classification of Agricultural Land Soils: A Data Mining Approach. *Agricultural Journal*, 6: 82-86.
- [17] Verheyen, K., Adriaens, D., Hermy, M., & Deckers, S. (2001). High-resolution continuous soil classification using morphological soil profile descriptions. *Geoderma*, 101(3), 31-48.
- [18] Meyer, G. E., Camargo Neto, J., Jones, D. D., & Hindman, T. W. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. *Computers and electronics in agriculture*, 42(3), 161-180.
- [19] Leemans, V., & Destain, M. F. (2004). A real-time grading method of apples based on features extracted from defects. *Journal of Food Engineering*, 61(1), 83-89.
- [20] K.A. Klise and S.A. McKenna.(2006). Water Quality Change Detection: Multivariate Algorithms. *Proceedings of SPIE 6203, Optics and Photonics in Global Homeland Security II*, T.T. Saito,D. Lehrfeld (Eds.)
- [21] Tellaeché, A., BurgosArtizzu, X. P., Pajares, G., & Ribeiro, A. (2007). A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms. In *Innovations in Hybrid Intelligent Systems* (pp. 72-79). Springer Berlin Heidelberg.
- [22] Urtubia, A., Pérez-Correa, J. R., Soto, A., & Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. *Food Control*,18(12), 1512-1517.
- [23] Rajagopalan, B., & Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *WATER RESOURCES RESEARCH*,35(10), 3089-3101.
- [24] Elizondo, D. A., McClendon, R. W., & Hoogenboom, G. (1994). Neural network models for predicting flowering and physiological maturity of soybean. *Transactions of the ASAE (USA)*.

- [25] Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental modelling & software*, 15(1), 101-124.
- [26] Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J. D., & Moreno, J. (2003). Support vector machines for crop classification using hyperspectral data. In *Pattern recognition and image analysis*(pp. 134-141). Springer Berlin Heidelberg.
- [27] Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, 330(3), 621-640.
- [28] "Naïve Bayes", Wikipedia, February 2012
- [29] "C4.5 (J48)", Wikipedia, February 2012
- [30] W. Cohen, (1995)," Fast Effective Rule Induction, in Twelfth International Conference on Machine Learning.
- [31] I. Witten & F. Eibe, (2005), "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition, San Francisco: Morgan Kaufmann,

The study of data mining in Health care Sector

Dr.K.Rajeshwar Rao ¹, Mr.B.Mahendhar Reddy ², Mr. K.Venkateshwar Rao ³, Mr.S.BalaKrishna Reddy ⁴

¹ CSE DEPT, Siddhartha Institute of Engineering & Technology (SIEI), Hyderabad, India

² CSE DEPT, Siddhartha Institute of Engineering & Technology (SIEI), Hyderabad, India

³ CSE DEPT, Vardhaman Engineering college, Hyderabad, India

, ⁴Siddhartha Institute of Engineering & Technology (SIEI), Hyderabad, India

Abstract

Data mining has been used in so many organizations. In healthcare, now a day's data mining is becoming very popular. Data mining applications can very benefit all parties involved in the healthcare industry. The very huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform some quantity of data into useful information for decision making. This article explores data mining applications in healthcare

Keywords- Data Mining, Health Care, Classification, Clustering, Association

INTRODUCTION

Data mining is a collection of algorithmic techniques to extract instructive patterns from raw data. Now a day's healthcare industry produces huge amounts of various data about hospitals, resources, disease diagnosis, electronic patient records, etc. Data mining processes include a hypothesis, gathering data, performing pre-processing, estimating the model, and understanding the model.

Before studying the data mining algorithms and how applied on medical data, let us understand what are the algorithms exists in data mining and how they are functioning.

In the middle of 1990's data mining is a strong tool that extracts useful information from a bulk of data.. Data Mining where suitable Data Mining technique is applied to the transformed data for extracting valuable information and evaluation is the last stage as shown in Figure 1

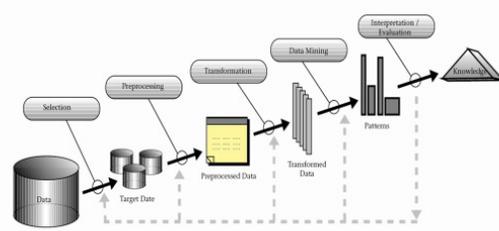


Figure 1. Stages of Knowledge Discovery Process

Knowledge Discovery Data in databases is the process of retrieving high-level knowledge from low-level data. It is an iterative process that comprises steps like Selection of Data, Pre-processing the selected data, Transformation of data into the appropriate form, Data mining to extract necessary information and Interpretation/Evaluation of data.

Selection step is to collect the heterogeneous data from varied sources for processing.

Pre-processing step is to perform basic operations of eliminating the noisy data and try to find the missing data or to develop a strategy for handling missing data, detect or remove outliers and resolve inconsistencies among the data.

Transformation step transforms the data into forms which is suitable for mining by performing any one. Data reduction task shrinks the data and represents the same data in less volume, but produces similar analytical outcomes. Data mining is the main component of the KDD process.

Interpretation/ Evaluation step includes presentation of mined patterns in understandable form. Various types of information need different type of representation, in this step the mined patterns are interpreted.

Knowledge discovery is the process of integrating the extracted knowledge into another system

KDD can be effective at working with huge data to define relevant pattern and to develop strategic results.

A health care organization can implement Knowledge Discovery in databases (KDD) by the help of experienced employee who has good understanding in health care domain.

Data mining algorithms are classified in two categories:

1. Descriptive model(or unsupervised learning)
2. Predictive model(or supervised learning).

Descriptive model is to discover patterns in the data and identifies the associations between attributes represented by the data.

In contrast, the purpose of Predictive mining model is largely to predict the future outcome than existing behavior.

2. DATA MINING TECHNIQUES

[1]Data mining techniques like association, classification and clustering are used by healthcare organization to increase their capability for building appropriate conclusions regarding patient health from raw facts and figures.

2.1. Classification

Classification can be categorized into two types –

- 1) Training
- 2) Testing.

Training data builds a classification model on the basis of training data collected for generating classification rules. The IF-THEN prediction rule is used in data mining; they signify facts at a high level of abstraction. The accuracy of classification model hinge on the degree to which classifying rules are true which is estimated by test data. In health care domain classification can be made useful as “if DiabeticFamilyHistory=yesANDhighSugerIntake=yes THEN DiabetesPossiblity=High”.

2.2. Clustering

Clustering is the process of generates a group of abstract objects into classes of similar objects.

A large database can be divided into number of small subgroups called clusters. It divides the data based on similarities it have. By using clustering algorithms to discovers collections of the data such that objects in the same cluster are more identical to each other than other groups.

2.3. Association

Association has great impact in the health care industry to discover the relationships between diseases, state of human health and the symptoms of disease. By using association in order to learn uncommon casual relationships in Electronic health databases. The integrated approach is useful for determining rules in the database and then by using these rules, an effective classifier is raised.

3. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

[2]Today's Healthcare industry creates huge amounts of complex data about patients like hospital resources, disease diagnosis, electronic patient records, medical devices etc. Data mining applications are used in healthcare can be grouped as the evaluation into broad categories, Treatment effectiveness Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves that the effective by comparing and contrasting causes, symptoms, and courses of treatments and better identify and tracking the chronic disease states and highly-risk patients, to design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare

management system.[2]Data mining applications are used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

Customer relationship management Customer relationship management is a core approach to managing interactions between commercial organizations typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings. Data mining applications are used to detect the fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

[3]Medical Device Industry Healthcare system's one important role is medical device.to best communication work this one is mostly used. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.Hospital Management System's including modern hospitals are capable of generating and collecting a large amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized [6]. Three layers of hospital management: Services for hospital management: Some of the Services for hospital management.

- Services for medical staff
- Services for patients
- System Biology

In biological databases contains a variety of data types, often with rich relational structure.Consequently multirelational data mining techniques are frequently applied to biological data.Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

[4]Most of the data mining tools are used to predict the successful results from the data recorded on healthcare problems. Different types of data mining tools are used to predict the accuracy level in different healthcare problems. In this study, the following list of medical problems has been analyzed and evaluated.[5]

- Heart Disease Cancer
- HIV/AIDS
- Blood
- Brain Cancer
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

In the most important healthcare problems specifically in disease side and research results have been illustrated. The diseases are the most critical problems in human beings. To analyze the data mining applications to effectiveness for diagnosing the disease, the traditional methods of mathematical / statistical applications are also given and compared.

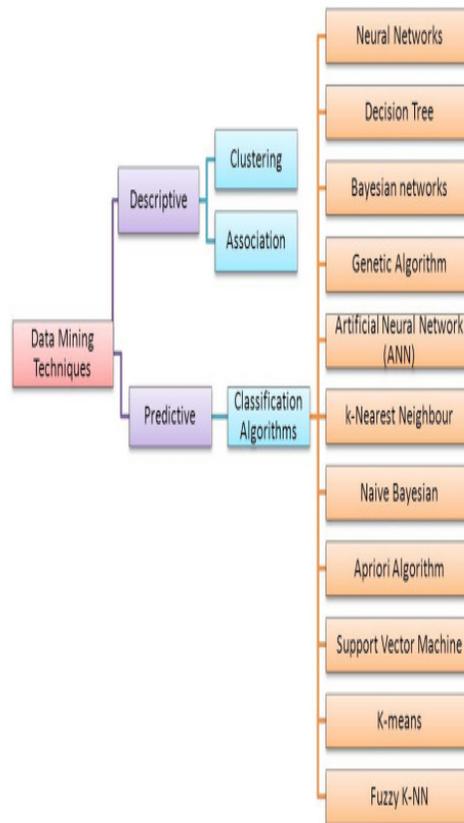
[6]TABLE 1. DATA MINING APPLICATIONS IN HEALTHCARE

DST Sponsored National Conference on Recent Advancements on Computer Science
(CONRACS 2019)– 26 to 28 July 2019

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naive	Probability	60
2	Cancer	WEKA	Classification	Rules, Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFLA		85
6	Tuberculosis	WEKA	Naive Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

APPLICATION OF DATA MINING TECHNIQUES IN HEALTH CARE

[7]The different classification algorithms mentioned below in figure 1 are used to predict or to analyse various diseases.



CONCLUSION

In this paper we observe that some data mining techniques that has been employed for medical data. As there is a huge of record in this industry and because of this, it has become requisite to use data mining techniques to help in decision support and prediction in the field of Healthcare system to identify the kind of disease. The medical data mining produces business intelligence which is useful for diagnosing of the disease. This paper throws that data mining techniques that is used for medical data for various diseases which are identified and diagnosed for human health.

REFERENCES

- 1) Data Mining Techniques by Michael J.A Berry, Gordon S. Linoff
- 2) Dhanya P Varghese & Tintu P B, "A SURVEY ON HEALTH DATA USING DATA MINING TECHNIQUES", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, Oct-2015
- 3) K.Sharmila & Dr.S.A.Vethamanickam, "SURVEY ON DATA MINING ALGORITHM AND ITS APPLICATION IN HEALTHCARE SECTOR USING HADOOP PLATFORM", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume: 05, Issue: 01, January-2015.
- 4) K.Sharmila & Dr.S.A.Vethamanickam, "SURVEY ON DATA MINING ALGORITHM AND ITS APPLICATION IN HEALTHCARE SECTOR USING HADOOP PLATFORM", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume: 05, Issue: 01, January-2015.
- 5) Pradnya P. Sondwale, "OVERVIEW OF PREDICTIVE AND DESCRIPTIVE DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer Science and Software Engineering, Volume: 05 Issue: 04, April-2015.
- 6) Sheetal L. Patil, "SURVEY OF DATA MINING TECHNIQUES IN HEALTHCARE", International Research Journal of Innovative Engineering, Volume: 01 Issue: 09, September-2015.

7) C. Hattice & K. Metin, “A DIAGNOSTIC SOFTWARE TOOL FOR SKIN DISEASES WITH BASIC AND WEIGHTED K-NN”, Innovations in Intelligent Systems and Applications (INISTA), 2012.

TRENDS ENABLED IN DATA SCIENCE

KADARI SRINIVASA RAO

Faculty in Computer Science,
BJR Govt Degree College,
Narayanguda, Hyderabad. Telangana.

kadarisrinu@gmail.com.

ABSTRACT:

The advancement in the analytical eco-space has reached new heights in the recent past. The emergence of new tools and techniques has certainly made life easier for an analytics professional to play around with the data. Moreover, the massive amounts of data that's getting generated from diverse sources need huge computational power and storage system for analysis. Three of the most commonly used terms in analytics are Data mining, Machine Learning, and Data Science which is a combination of both.

KEY WORDS: Data mining, Machine Learning, Data Science, Big Data,
Artificial Intelligence

Data Mining:

By term 'mining' we refer to extracting some object by digging. Similarly, that analogy could be applied to data where information could be extracted by digging into it. Data mining is one of the most used terms these days. Unlike previously, our life is circulated entirely by big data and we have the tools and techniques to handle such voluminous diverse meaningful data.

In the data, there are a lot of patterns which people could discover once the data has been gathered from relevant sources. The hidden patterns could be extracted to provide valuable insights by combining multiple sources of data even if it is junk. This entire process is known as Data mining.

Now the data used for mining could be enterprise data which are restricted and secured and has privacy issues. It could also be an integration of multiple sources which includes financial data, third-party data, etc. The more the data available to us, the better it is as we need to find patterns and insights in sequential and non-sequential data.

The steps involved in data mining are –

- **Data Collection** – This is one of the most important steps in Data mining as getting the correct data is always a challenge in any organization. To find patterns in the data, we need to ensure that the source of the data is accurate and as much as possible data is gathered.

- **Data Cleaning** – A lot of the times the data we get is not clean enough to draw insights from it. There could be missing values, outliers, NULL in the data which needs to be handled either by deletion or by imputation based on its significance to the business.
- **Data Analysis** – Once the data is gathered, and cleaned the next step is to analyze the data which in short known as Exploratory Data Analysis. Several techniques and methodologies are applied in this step to derive relevant insights from the data.
- **Data Interpretation** – Only analyzing the data is worthless unless it is interpreted through the form of graphs or charts to the stakeholders or the business who would make conclusions based on the analysis.

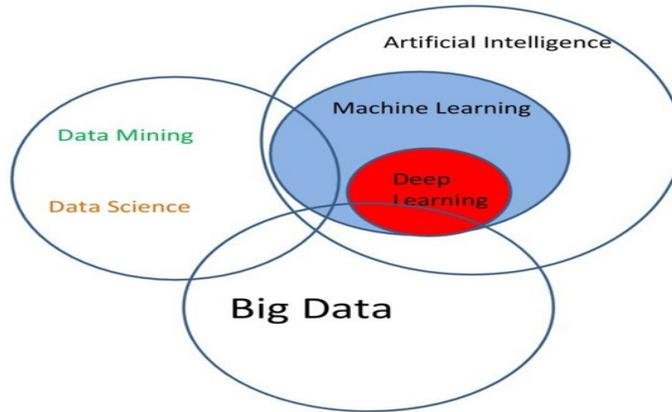
Data mining has several usages in the real world. For example, if we take the **logs data** for login in a web application, we would see that the data is messy containing information like timestamp, activities of the user, time spent on the website, etc. However, if we clean the data, and then analyze it, we would find some relevant information from it such as the user's regular habit, the peak time for most of the activities, and so on. All this information could help to increase the efficiency of the system.

Another example of data mining is in **crime prevention**. Though data mining has most usage in education and healthcare, it is also used by agencies in the crime department to spot patterns in the data. This data would consist of information about some of the criminal activities that have taken place. Hence, mining, and gathering information from the data would help the agencies to predict future crime events and prevent it from occurring. The agencies could mine the data and find out the place where the next crime could take place. They could also prevent cross-border calamity by understanding which vehicle to check, the age of the occupants, etc.

However, a few of the important points one should remember about Data Mining –

- Data mining should not be considered as the first solution to any analysis task if other accurate solutions are applicable. It should be used when such solutions fail to provide value.
- Sufficient amount of data should be present to draw insights from it.
- The problem should be understood to be a Regression or a Classification one

BLOCK DIAGRAM OF DATA



Machine Learning:

Previously, we learned about Data mining which is about gathering, cleaning, analyzing, and interpreting relevant insights from the data for the business to draw conclusions from it.

If Data mining is about describing a set of events, Machine Learning is about predicting the future events. It is the term coined to define a system which learns from past data to generalize and predict the future events from the unknown set of data.

Machine Learning could be divided into three categories –

- **Supervised Learning** – In supervised learning, the target is labeled i.e., for every corresponding row there is an output value.
- **Unsupervised Learning** – The data set is unlabelled in unsupervised learning i.e., one has to cluster the data into various groups based on the similarities in the pattern of the data points.
- **Reinforcement Learning** – It is a special category of Machine Learning which is mostly used in self-driving cars. In reinforcement learning, the learner is rewarded for every correct move, and penalized for any incorrect move.
- The field of Machine Learning is vast, and it requires a blend of statistics, programming, and most importantly data intuition to master it. Supervised and unsupervised learning are used to solve regression, classification, and clustering problems.

- In regression problems, the target is numeric i.e., continuous or discrete in nature. A continuous value could be an integer, float, or a decimal, whereas a discrete value is a number or an integer.
- In classification problems, the target is categorical i.e., binary, multinomial, or ordinal in nature.
- In clustering problems, the dataset is grouped into different clusters based on the similar properties among the data in a particular group.
- Machine Learning has a vast usage in various fields such as Banking, Insurance, Healthcare, Manufacturing, Oil and Gas, and so on. Professionals from various disciplines feel the need to predict future outcomes in order to work efficiently and prepare for the best by taking appropriate actions. Some of the real-life examples where Machine Learning has found its usage is –
- **Email Spam filtering** – This is the first application of Machine Learning where an email is classified as ‘Spam’ or ‘Not Spam’ based on certain keywords in the mail. It is a binary classification supervised learning problem where the system is initially trained with a set of sample emails to learn the patterns which would help in filtering out irrelevant emails. Once the system has generalized well, it is passed through a validation set to check for its efficiency, and then through a test set to find its accuracy.
- **Credit Risk Analytics** – Machine Learning has vast influence in the Banking, and Insurance domain with one of its usage being in predicting the delinquency of a loan by a borrower. Defaulting a credit loan is a prevalent issue in which the lender or the bank has lost millions by failing to identify the possibility of a borrower not repaying back the loans or meeting the contractual agreements. However, Machine Learning has been introduced by various banks which takes into several features of a borrower and builds a predictive model which helps in mitigating the risk involved in giving credit card loans to them.
- **Product Recommendations** – Flipkart, and Amazon are of the two biggest e-commerce industry in the world where millions of users shop every day the products of their choice. However, there is a recommendation algorithm that works behind the scenes which simplify the life of the customer by displaying them the products they make like based on their previous shopping or search patterns. This is an example of unsupervised learning where a customer is grouped based on their shopping patterns.

Data Science:

So far, we have learned about the two most common and important terms in Analytics i.e., Data mining and Machine Learning.

If Data mining deals with understanding and finding hidden insights in the data, then Machine Learning is about taking the cleaned data and predicting future outcomes. All of these together form the core of Data Science.

Data Science is a holistic study which involves both Descriptive and Predictive Analytics.

A Data Scientist needs to understand and perform exploratory analysis as well as employ tools, and techniques to make predictions from the data.

A Data Scientist role is a mixture of the work done by a Data Analyst, a Machine Learning Engineer, a Deep Learning Engineer, or an AI researcher. Apart from that, a Data Scientist might also be required to build data pipelines which is the work of a Data Engineer. The skill set of a Data Scientist consists of Mathematics, Statistics, Programming, Machine Learning, Big Data, and communication.

Some of the applications of Data Science in the modern world are –

- **Virtual assistant** – Amazon’s Alexa, and Apple’s Siri are two of the biggest achievements in the recent past where AI has been used to build human-like intelligent systems. A virtual assistant could perform most of the tasks that a human being could with proper instructions.
- **ChatBot** – Another common usage of Data Science is the ChatBot development which is now being integrated into almost every corporation. A technique called Natural Language Processing is in the core of ChatBot development.
- **Identifying cancer cells** – Deep Learning has made tremendous progress in the healthcare sector
- where it is used to identify the pattern in the cells to predict whether it is cancerous or not.
- Deep Learning uses neural networks which functions like the human brain.

Conclusion

Data mining, Machine Learning, and Data Science is a broad field and it would require quite a few things to learn to master all these skills. The advancement in the analytical eco-space has reached new heights in the recent past. The emergence of new tools and techniques has certainly made life easier for an analytics professional to play around with the data. Moreover, the massive amounts of data that’s getting generated from diverse sources need huge computational power and storage system for analysis. Three of the most

commonly used terms in analytics are Data mining, Machine Learning, and Data Science which is a combination of both.

REFERENCES

1. T. H. Davenport, J. G. Harris, *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, 2007.
2. "Big Data: The Next Frontier for Innovation Competition and Productivity" in McKinsey Global Institute, 2011.
3. F. Jones, S. Wuchty, B. Uzzi, "Multi-University Research Teams: Shifting Impact Geography and Stratification in Science", *Science* 322, pp. 1259-1262, 2008.
4. [online] Available: <https://catalyst.harvard.edu/spotlights/profiles.html>
5. N. Godbole, J. Lamb, "The Triple Challenge for the Healthcare Industry: Sustainability Privacy and Cloud-Centric Regulatory Compliance", *CEWIT*, 2013.
6. L. P. English. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, 1999.
7. W. H. Inmon, *Building the Data Warehouse*, 2005.
8. Amandeep Khurana, "Bringing Big Data Systems to the Cloud", *IEEE CLOUD COMPUTING PUBLISHED BY THE IEEE COMPUTER SOCIETY*, pp. 72-75, 2014.
9. A. Kabiri, D. Chiadmi, "A method for modelling and organazing ETL processes", *2012 Second International Conference on Innovative Computing Technology (INTECH)*, 18–20 Sept. 2012.
10. B. F. Jones, S. Wuchty, B. Uzzi, "Multi-University Research Teams: Shifting Impact Geography and Stratification in Science", *Science* 322, pp. 1259-1262, 2008.

An Insight on Data Science

C.V. KrishnaVeni,

Lecturer in Computer Science,
SKR&SKR GCW(A), Kadapa, Andhra
Pradesh, India.

Abstract: According to John Tukey’s quote: “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”. You may have 100 Gb and only 3 Kb are useful for answering the real question you care about. When you start with the question you often discover that you need to collect new data or design an experiment to confirm you are getting the right answer. It is easy to discover “structure” or “networks” in a data set. There will always be correlations for a thousand reasons if you collect enough data. Understanding whether these correlations matter for specific, interesting questions is much harder. Often the structure you found on the first pass is due to a phenomena (measurement error, artifacts, data processing) that doesn’t answer an interesting question. The answer for all the above questions is Data Science. This paper presents an insight on Data Science. Covers the topics Data Analytics vs Data Science, History of Data Science, Data Science in Detail, Lifecycle of Data Science, Data Science Strategy Competencies, Characteristics of Good Data Science, Role of Data Scientist under various sections in detail.

I. **Introduction :** The data are observations that are measured and communicated in such a way as to be intelligible to both the recorder and the reader. So, you as a

person are not data, but recorded observations about you are data. For example, your name when written down is data; or the digital recording you speaking your name is data; or a digital photograph of your face or video of you dancing are data[1]. The term "science" implies knowledge gained through systematic study[2]. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data[3].

II. *Data Analytics vs Data Science :*

Data analysis is the process of evaluating data using analytical and statistical tools to discover useful information and aid in business decision making. Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data[4].

A Data Analyst usually explains what is going on by processing history of the data. Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future[4].

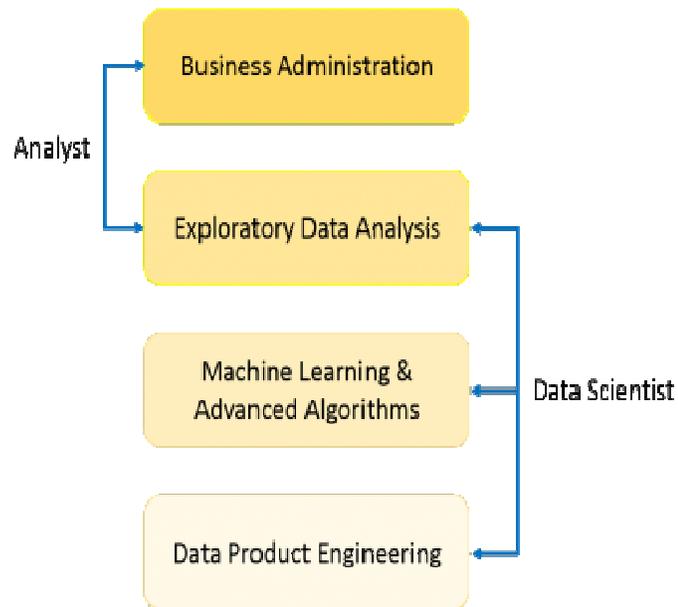


Fig: Difference between Data Analyst and Data Scientist, Source: Reference[4].

III. History of Data Science:

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage, it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy" [5]. In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

The modern definition of "data science" was first sketched during the second Japanese-French statistics symposium organized at the University of Montpellier II (France) in 1992 [6]. The attendees acknowledged the emergence of a new discipline

with a specific focus on data from various origins, dimensions, types and structures. They shaped the contour of this new science based on established concepts and principles of statistics and data analysis with the extensive use of the increasing power of computer tools. In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods") [7], after the term was introduced in a roundtable discussion by Chikio Hayashi.

IV. Data Science in Detail :

Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

- **Predictive causal analytics** – If you want a model which can predict the possibilities of a particular event in the future, you need to apply predictive causal analytics. Say, if you are providing money on credit, then the probability of customers making future credit payments on time is a matter of concern for you. Here, you can build a model which can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.
- **Prescriptive analytics:** If you want a model which has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters, you certainly need prescriptive analytics for it. This relatively new field is all about providing advice. In other terms, it not only predicts but suggests a range of prescribed actions and associated outcomes.

The best example for this is Google’s self-driving car which I had discussed earlier too. The data gathered by vehicles can be used to train self-driving cars. You can run algorithms on this data to bring intelligence to it. This will enable your car to take decisions like when to turn, which path to take, when to slow down or speed up.

- **Machine learning for making predictions** – If you have transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bet. This falls under the paradigm of supervised learning. It is called supervised because you already have the data based on which you can train your machines. For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

- **Machine learning for pattern discovery** – If you don’t have the parameters based on which you can make predictions, then you need to find out the hidden patterns within the dataset to be able to make meaningful predictions. This is nothing but the unsupervised model as you don’t have any predefined labels for grouping. The most common algorithm used for pattern discovery is Clustering.

Let’s say you are working in a telephone company and you need to establish a network by putting towers in a region. Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength[4].

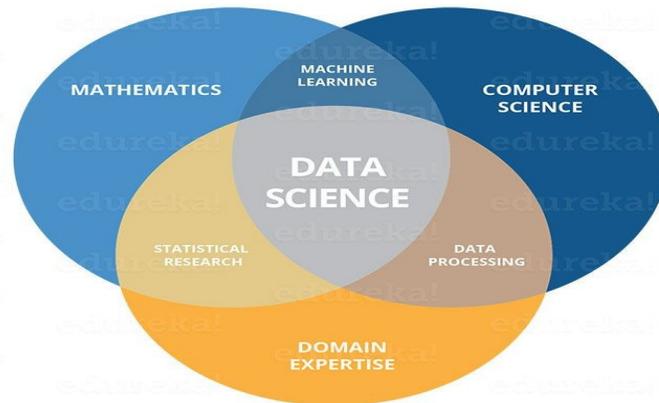
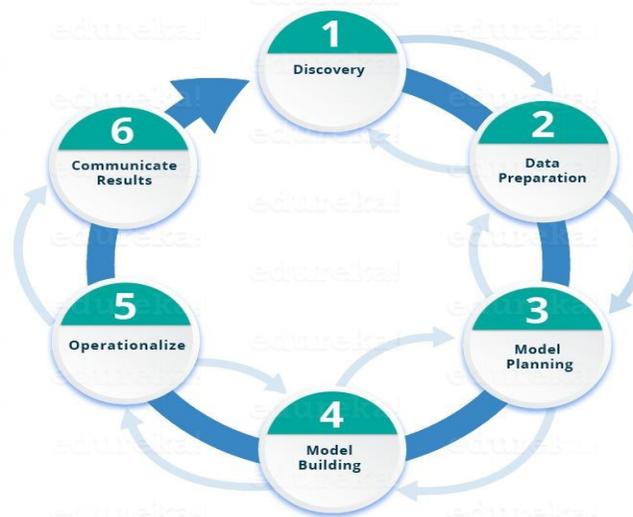


Fig: Data Science is a combination of Data processing, machine learning etc., Source: reference[4].

V. **Lifecycle of Data Science:** As per the reference cited [4], life cycle of data science is



described as follows:

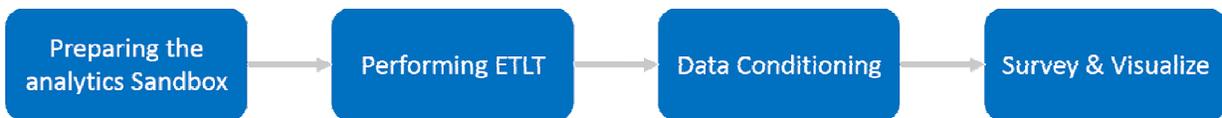


Phase 1—Discovery: Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the

project. In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.



Phase 2—Data preparation: In this phase, you require analytical sandbox in which you can perform analytics for the entire duration of the project. You need to explore, preprocess and condition data prior to modeling. Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox. Let's have a look at the Statistical Analysis flow below.



You can use R for data cleaning, transformation, and visualization. This will help you to spot the outliers and establish a relationship between the variables. Once you have cleaned and prepared the data, it's time to do exploratory analytics on it. Let's see how you can achieve that.



Phase 3—Model planning: Here, you will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which you will implement in the next phase. You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

VI. *Data Science Strategy Competencies*

The craft of data science combines three different competencies. Data scientist Drew Conway visualised the three core competencies of data science. Firstly and most importantly, data science requires *domain knowledge*. Any analysis needs to be grounded in the reality it seeks to improve. Subject-matter expertise is necessary to make sense of the investigation. Professional expertise in most areas uses *mathematics* to understand and improve outcomes. New mathematical tools expand the traditional approaches to develop a deeper understanding of the domain under consideration. *Computer science* is the competency that binds the available data with mathematics. Writing computer code to extract, transform and analyse data to create information and stimulate knowledge is an essential skill for any data scientist.

VII. *Characteristics of Good Data Science*

To create value with data, we need to know how to create or recognise good data science. The second chapter uses three principles originally introduced two thousand years ago by Roman architect and engineer Vitruvius. He wrote that buildings need to be *useful, sound* and *aesthetic*. These requirements are also ideally suited to define best-practice in data science.

- *Usefulness*

Useful data science meaningfully improves our reality through data. Data is a representation of either a social or physical reality. Any data source is ever only a

sample of the fullness and complexity of the real world. Information is data imbued with context. The raw data collected from reality needs to be summarised, visualised and analysed for managers to understand the reality of their business. This information increases knowledge about a business process, which is in turn used to improve the reality from which the data was collected. This feedback loop visualises the essence of analysing data in businesses. Data science is a seductive activity because it is reasonably straightforward to create impressive visualisations with sophisticated algorithms. If data products don't improve or enlighten the current situation, they are in essence useless.

- *Soundness*

Data science needs to be *sound* in that the outcomes are valid and reliable. The validity and reliability of data are where the science meets the traditional approaches to analysing data. Validity is the extent to which the data represents the reality it describes. The reliability of data relates to the accuracy of the measurement. These two concepts depend on the type of data under consideration. Measuring physical processes is less complicated than the social aspects of society. Validity and reliability are in essence a sophisticated way of expressing the well-known Garbage-In-Garbage-Out principle. Reliability and validity of data and analysis. The soundness of data science also relates to the reproducibility of the analysis to ensure that other professionals can review the outcomes. Reproducibility prevents that the data and the process by which it was transformed and analysed become a black-box where we have no reason to trust the results. Data science also needs to be sound concerning the governance of the workflow. All data sources need to be curated by relevant subject

matter experts to ensure their validity and reliability. Data experts provide that the data is available to those who need it.

- *Aesthetics*

Lastly, data science needs to be *aesthetic* to ensure that any visualisation or report is easy to understand by the consumer of the analysis. This requirement is not about beautification through infographics. Aesthetic data products minimise the risk of making wrong decisions because the information is presented without room for misinterpretation. Any visualisation needs to focus on telling a story with the data. This story can be a comparison, a prediction, a trend or whatever else is relevant to the problem.

One of the essential principles of aesthetic data science is the data-to-pixel ratio.

This principle means that we need to maximise the ratio between all the pixels on a screen and those pixels that present information. Good data visualisation practices austerity to ensure that the people that consume the information understand the story that needs to be told.

VIII. Role of Data Scientist :

The first data scientist in a business context was Frederick Taylor. He was a pioneer in using data in business to lower the influence of opinions and rules-of-thumb in favour of a scientific approach to management. Various roles of data scientist are discussed below:

1. Empowering management and officers to make better decisions

An experienced data scientist is likely to be as a trusted advisor and strategic partner to the organization's upper management by ensuring that the staff maximizes their analytics' capabilities. A data scientist communicates and demonstrates the value of the institution's data to facilitate improved decision-making processes across the entire organization, through measuring, tracking, and recording performance metrics and other information.

2. Directing actions based on trends—which in turn help to define goals

A data scientist examines and explores the organization's data, after which they recommend and prescribe certain actions that will help improve the institution's performance, better engage customers, and ultimately increase profitability.

3. Challenging the staff to adopt best practices and focus on issues that matter.

One of the responsibilities of a data scientist is to ensure that the staff is familiar and well-versed with the organization's analytics product. They prepare the staff for success with the demonstration of the effective use of the system to extract insights and drive action. Once the staff understands the product capabilities, their focus can shift to addressing key business challenges.

4. Identifying opportunities

During their interaction with the organization's current analytics system, data scientists question the existing processes and assumptions for the purpose of developing additional methods and analytical algorithms. Their job requires them to continuously and constantly improve the value that is derived from the organization's data.

5. Decision making with quantifiable, data-driven evidence.

With the arrival of data scientists, data gathering and analyzing from various channels has ruled out the need to take high stake risks. Data scientists create models using existing data that simulate a variety of potential actions—in this way, an organization can learn which path will bring the best business outcomes.

6. Testing these decisions

Half of the battle involves making certain decisions and implementing those changes. What about the other half? It is crucial to know how those decisions have affected the organization. This is where a data scientist comes in. It pays to have someone who can measure the key metrics that are related to important changes and quantify their success.

7. Identification and refining of target audiences

From Google Analytics to customer surveys, most companies will have at least one source of customer data that is being collected. But if it isn't used well—for instance, to identify demographics—the data isn't useful. The importance of data science is based on the ability to take existing data that is not necessarily useful on its own and combine it with other data points to generate insights an organization can use to learn more about its customers and audience.

8. Recruiting the right talent for the organization

Reading through resumes all day is a daily chore in a recruiter's life, but that is changing due to big data. With the amount of information available on talent—through social media, corporate databases, and job search websites—data science specialists can work their way through all these data points to find the candidates who best fit the organization's needs. By mining the vast amount of data that is already available, in-house processing for resumes and applications—and even sophisticated data-driven aptitude tests and games—data science can help your recruitment team make speedier and more accurate selections[8].

IX. Conclusion:

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science" has recently become a popular term among business executives, many critical academics. A data scientist can help with the identification of the key groups with

precision, via thorough analysis of disparate sources of data. With this in-depth knowledge, organizations can tailor services and products to customer groups, and help profit margins flourish. This paper dealt with data science and role of data scientist in detail.

References :

1. https://en.wikibooks.org/wiki/Data_Science:_An_Introduction/Definitions_of_Data
 2. <https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/abstract>
 3. https://en.wikipedia.org/wiki/Data_science
 4. <https://www.edureka.co/blog/what-is-data-science/>
 5. Naur, Peter (1 July 1966). "The science of datalogy". *Communications of the ACM*. **9** (7): 485.
 6. Escoufier et al., editors (1995). "Preface". *Data Science and its Application*. Tokyo: Academic Press.
 7. Press, Gil. "A very Short History of Data Science".
- <https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article>

Loading, Searching and Retrieving data from clustered data nodes on HDFS

¹M.Sreerama Murty, ²Dr N Nagamalleswara Rao

¹Research Scholar, Department of CSE, Achary Nagarjuna
University, Guntur

sreeramssit@gmail.com

²Professor, Department of CSE, RVR & JC College of Engineering, Guntur

nmrao@rvrjc.ac.in

Abstract: The storing, selecting and tracking of data from the heterogeneous data nodes on the different data nodes on hadoop distributed file system (HDFS) frame work in clustered data nodes. In particular, the capacity to store, search and query execution on distributed environment becoming a more complex. We present one method to handle the data from clustered nodes without delay of the executing a query on distributed environment. We define easy and understandable process of the way results were derived from the data and how particular bit of data were combined to answer a query. We use replica analysis for efficient query for data retrieve from data files in cluster node and results stored in cluster data nodes. We empirically examine the presented methods and display that the process of storing and searching data in clustered nodes is acceptable. We show that final results of a query with linked data nodes can also improve the performance of a query execution.

Index Terms----HDFS, replica, cluster, distributed, capacity.

1 Introduction

The Hadoop Distributed File System is the main data storage used by Hadoop application environment. It employed Namenode and Datanode architecture to execute a distributed file system that provides high performance access to data across highly scalable Hadoop clusters. HDFS support the rapid transfer of data between computer nodes. It was closely coupled with MapReduce framework for data processing.

The Hadoop Distributed File System is designed to be highly fault-tolerant. The file system replicates, or copies, each piece of data multiple times and distributes the copies to individual nodes, placing at least one copy on different server rack than the others. If the data nodes that crash can be found elsewhere within a cluster. This ensures that processing can continue while data is recovered.

HDFS uses Master/Slave skeleton. Each Hadoop cluster consisted of a single Namenode that managed file system operations and supporting DataNodes that data storage on individual compute node. The HDFS elements combine to support applications with large datasets. Hadoop programs written in java that allows distributed processing of large datasets across clusters of computers using programming models .the Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadop is designed to scale up from single server to thousands of machines, each offering local computation and storage. The following fig 1.1 is to describe the how write a program can executed in multiple cluster nodes.

1. 1 **The Hadoop architecture**

- i. MapReduce
- ii. Hadoop Distributed File System

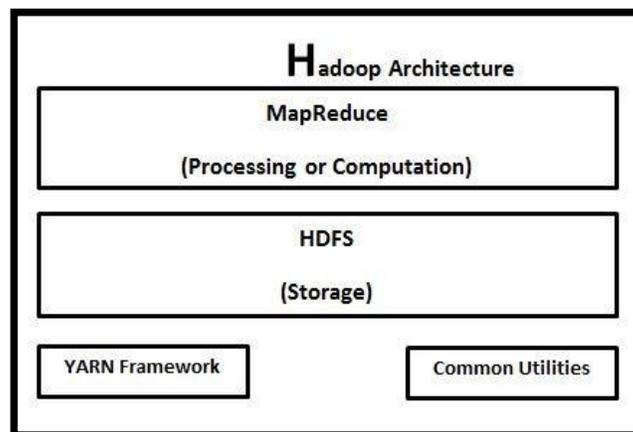


Fig 1.1 Hadoop Architecture

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

1.2 **Map Reduce Algorithm**

Algorithm: MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The

Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

1. During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
2. The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
3. Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
4. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

1.3 MapReduce Input and Output

The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.

The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework. Input and Output types of a MapReduce job: (Input) <k1, v1> -> map -> <k2, v2>-> reduce -> <k3, v3> (Output).

	INPUT	OUTPUT
Map	<k1,v1>	list(<k2,v2>)
Reduce	<k2,list(v2)>	list(<k3,v3>)

i. Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware.

Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:

- (i). **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules.
- (ii) **Hadoop YARN:** This is a framework for job scheduling and cluster resource management

1.2 Storage Model of Hadoop Distributed File System

Hadoop uses HDFS to store files efficiently in the cluster. When a file is placed in HDFS it is broken down into blocks, 64 MB block size by default. These blocks are then replicated across the different nodes (*DataNodes*) in the cluster. The default replication value is 3, i.e. there will be 3 copies of the same block in the cluster. We will see later on why we maintain replicas of the blocks in the cluster. A Hadoop cluster can comprise of a single node (single node cluster) or thousands of nodes.

The Hadoop file system model consists of five parts:

1.2.1 Namenode

1.2.2 Datanode

1.2.3 Secondary Name Node

1.2.4 Job Tracker

1.2.5 Task Tracker

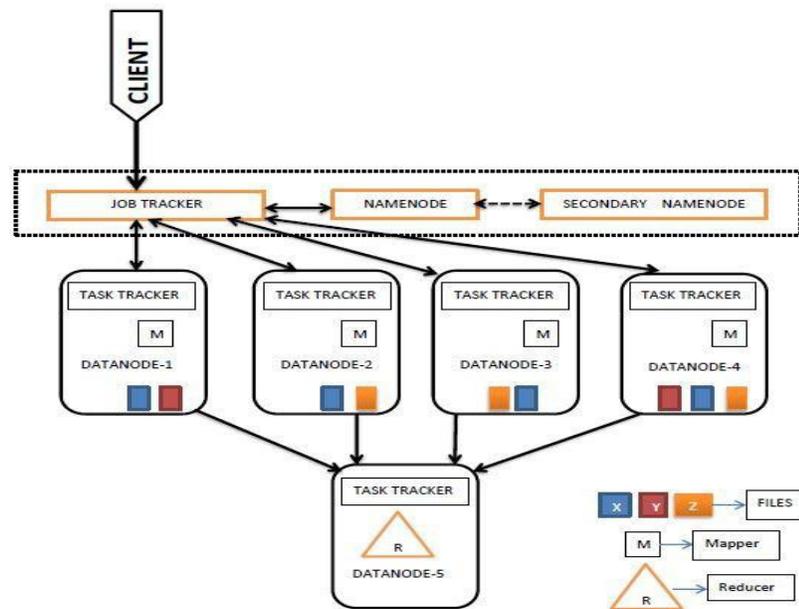


Figure 1.2.1 Typical Hadoop Cluster

In the figure 1.2.1 the Namenode, Secondary Namenode and JobTracker running on a single machine. A file in HDFS is split into blocks and replicated across Datanodes in a hadoop cluster. We see that three files X,Y and Z have been split across with a replication factor of 3 across the different Datanodes.

Namenode is the node,it keeps all the files location information of the file in HDFS.so it store the metadata for HDFS. The Secondary NameNode is not a failover node for

theNameNode. The secondary name node is responsible for performing periodic housekeeping functions for NameNode. It only creates checkpoints of the file system present in the NameNode. The DataNode is responsible for storing the files in HDFS. It manages the file blocks within the node. It sends information to the NameNode about the files and blocks stored in that node and responds to the NameNode for all filesystem operations. JobTracker is responsible for taking in requests from a client and assigning TaskTrackers with tasks to be performed. The JobTracker tries to assign tasks to the TaskTracker on the DataNode where the data is locally present (Data Locality). If that is not possible it will at least try to assign tasks to TaskTrackers within the same rack. If for some reason the node fails the JobTracker assigns the task to another TaskTracker where the replica of the data exists since the data blocks are replicated across the DataNodes. This ensures that the job does not fail even if a node fails within the cluster. TaskTracker is a daemon that accepts tasks (Map, Reduce and Shuffle) from the JobTracker. The TaskTracker keeps sending a heartbeat message to the JobTracker to notify that it is alive. Along with the heartbeat it also sends the free slots available within it to process tasks. TaskTracker starts and monitors the Map & Reduce Tasks and sends progress/status information back to the JobTracker

The following algorithm is exploiting the entire process of communication between Namenode and different cluster nodes on the intra communication system model.

Algorithm: A algorithm for the Storing, Tracing and Retrieving

Aim: able to store and access the data in distributed environment

1. Take input → structure, unstructured or semi-structured file
2. Store → copy the file in three different nodes
3. Namenode → file location information is stored in Namenode
4. Datanode → storing the files in HDFS.. it manages the file blocks within the node.
5. Task Tracker → Accept the task from the JobTracker
6. JobTracker → accept the request from the client and assigning Task Tracker with tasks are performed. JobTracker assigns the task to another TaskTracker where the replica of the data exists since the data blocks are replicated across the DataNodes. This ensures that the job does not fail even if a node fails within the cluster.
7. Results: selected results will be stored in the specified file in HDFS
8. If Name node Files → Secondary Namenode was initiate temporarily until wakeup the Namenode.
9. If Namenode == available

10. Goto step 1

3. Related Work

As Damasio et al. have noted, many of the annotated RDF approaches do not expose how- provenance (i.e., how a query result was constructed)]. The most comprehensive implementations of these approaches were presented by Annotated approaches have also been used for propagating “trust values”. Other recent work has looked at expanding the theoretical aspects of applying such a semi ring-based approach to capture SPARQL. In contrast, our work focuses on the implementation aspects of using annotations to track provenance within the query processing pipeline.

The concept of a provenance query was defined by Simon Miles in order to only select a relevant subset of all possible results when looking up the provenance of an entity. A number of authors have presented systems for specifically handling such provenance queries. Bit on et al. showed how user views can be used to reduce the amount of information returned by provenance queries in a workflow system.. Provenance is computed by using standard relational query rewriting techniques, e.g., using lazy and eager provenance computation models. Recently, Glavic and his team have built on this work to show the effectiveness of queryrewriting for tracking provenance in databases that support audit logs and “time travel” . Our approach is different, in that it looks at the execution of provenance queries in conjunction with standard queries within a graph database. Widom et al. presented the Trio system that supports the joint management of data, uncertainty and lineage. Lineage is an integral part of the storage model in this system, i.e., it is associated with its corresponding record in the database. Trio persists lineage information in a separate lineage relation where each record corresponds to a database tuple and contains provenance-specific attributes. However, Trio is not a database for semi-structured data, which is specifically our target. TripleProv leverages the specific requirements of RDF data and queries to enable efficient tracking and querying of provenance information. In that respect, our work is related to the work on annotated RDF , which developed SPARQL query extensions for querying over annotation metadata (e.g., provenance). Halpin and Cheney have shown how to use SPARQL Update to track provenance within a triplestore with no modifications to the SPARQL specification. Our focus is different, however, since we propose and empirically evaluate different execution strategies for running queries that take advantage of provenance metadata.

Our system partially builds upon dynamic query execution approaches, which have been studied in different contexts by database researchers. Graefe and Ward focused on

determining when re-optimizing a given query that is issued repeatedly is necessary. Subsequently, Colde and Graefe proposed a new query optimization model, which constructs dynamic plans at compile-time and delays some of the query optimization until run-time. Kabra and DeWitt proposed an approach collecting statistics during the execution of complex queries in order to dynamically correct suboptimal query execution plans. Ng et al. studied how to re-optimize suboptimal query plans on-the-fly for very long-running queries in database systems. Avnur and Hellerstein proposed Eddies, a query processing mechanism that continuously reorders operators in a query plan as it runs, and that merges the optimization and execution phases of query processing in order to allow each tuple to have a flexible ordering of the query operators. More recently, Madden et al. have extended Eddies to continuously adapt query plans over streams of data. Our work is different in the sense that we dynamically examine or drop data structures during query execution depending on provenance information.

4. Methods for Hadoop Cluster

4.1 Apache Pig

Apache Pig is a high-level procedural language for querying large data sets using Hadoop and the Map Reduce Platform. It is a Java package, where the scripts can be executed from any language implementation running on the JVM. This is greatly used in iterative processes. Apache Pig simplifies the use of Hadoop by allowing SQL-like queries to a distributed dataset and makes it possible to create complex tasks to process large volumes of data quickly and effectively. The best feature of Pig is that, it backs many relational features like Join, Group and Aggregate.

4.2 Apache Pig Latin

Apache Pig create a simpler procedural language abstraction over Map Reduce to expose a more Structured Query Language (SQL)-like interface for Hadoop applications called Apache Pig Latin, So instead of writing a separate Map Reduce application, you can write a single script in Apache Pig Latin that is automatically parallelized and distributed across a cluster. In simple words, Pig Latin is a sequence of simple statements taking an input and producing an output. The input and output data are composed of bags, maps, tuples and scalar.

4.3 Storage Modes

Apache Pig has two execution modes:

4.3.1 Local Mode

In „Local Mode“, the source data would be picked from the local directory in your computer

```
system. The MapReduce mode can be specified using „pig -x local” command
clodera@cloudera-vm:~$ pig -x local
2019-01-04 17:10:28,088 [main] INFO org.apache.pig.Main - logging error messages to:
/home/cloudera/pig_1385104108087.log
  2019-01-04 17:10:28,208 [main]
  INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
at:file:///
grunt>
```

4.3.2 MapReduce Mode:

To run Pig in MapReduce mode, you need access to Hadoop cluster and HDFS installation.

The MapReduce mode can be specified using the „pig” command

```
clodera@cloudera-vm:~$ pig
2019-01-04 17:10:27,956 [main] INFO org.apache.pig.Main - logging error messages to:
/home/cloudera/pig_1385103264962.log
  2019-01-04 17:10:28,152 [main]
  INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
at:file:///localhost:8080
2019-01-04 17:10:28,405 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map- reduce job
tracker at:localhost:8081
grunt>
```

5. Query Execution

We specified three queries to load and searching and retrieving the data from Hadoop cluster

1. **LOAD** operator is used to load data from the file system or HDFS storage into a Pig relation.
2. **FOREACH**: This operator generates data transformations based on columns of data. It is used to add or remove fields from a relation. Use **FOREACH-GENERATE** operation to work with columns of data
3. **Filter**: This operator selects tuples from a relation based on a condition.

We create two CSV files with name login.csv and rating.csv

The first file login.csv contain username, url & id

Login.csv

```
AASHRITH, GMAIL,10 NUSHANTH,
FLIPKART,9 SAIJYOSHIKA,
AMAZON,8 DHARANI,
```

DST Sponsored National Conference on Recent Advancements on Computer
Science (CONRACS 2019)- 26 to 28 July 2019

FACEBOOK,9

KARTHIK,

QUICKER,8

VISWANADH,

TWITTER,7

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

PRASADH, PSC.AP.GOV.IN,6
BHAVANI, GOV,4
KARUNA, IEEEEXPLORE,8 PADMA,
SLIDESHARE,5 KRISHNASRI,
SCRIBD,6 AASHRITH, AMAZON,9
PRASADH, QUICKER,5 KARUNA,
QUICKER,6 SAIJYOSHIKA,
AMAZON,10
NUSHANTH,FLIPKART,9 PADMA,
FACEBOOK,8 KRISHNASRI,
FLIPKART,10

The second file containing two fields url and rating

Rating.csv

GMAIL,5 FLIKART,4
AMAZON,8
FACEBOOK,9
QUICKER,8
TWITTER,7 GOV,6
APGOV,4
IEEEEXPLORE,8
SLIDESHARE,5
SCRIBD,6

Load the data files

The following query is used load the in to cluster both files

The first file login.csv to form relation store1.the field names are username,uri,id.

```
grunt> store1=load „/login“ using PigStorage(„,“)as (username:chararray,url:chararray,id:int);
```

The second file rating.csv to form relation store2. The field names are url and rating grunt>

```
store2=load „/rating“ using PigStorage(„,“)as (url:chararray,rating:int);
```

Retrieving the data from data file based on column data

```
grunt>results=foreach store1 generate url,id;
```

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

For executing above query by using dump;

```
grunt> dump results;
```

output:

```
(GMAIL,10)      (FLIPKART,9)
  (AMAZON,8)
  (FACEBOOK,9)
  (QUICKER,8) (TWITTER,7)
  (APGOV,6)
(GOV,4)
  (IEEEEXPLORE,8)
  (SLIDESHARE,5)
  (SCRIBD,6)
  (AMAZON,9)
  (QUICKER,5)
  (QUICKER,6)
  (AMAZON,10)
  (FLIPKART,9)
  (FACEBOOK,8)
  (FLIPKART,10)
```

Retrieving the data from data file based on tuple data.

```
grunt>t1=filter store1 by id>9
```

```
grunt>dump t1;
```

Output:

```
(AASHRITH,    GMAIL,10)    (NUSHANTH,
  FLIPKART,9) (DHARANI,  FACEBOOK,9)
(AASHRITH,  AMAZON,9) (SAIJYOSHIKA,
  AMAZON,10)  (NUSHANTH,FLIPKART,9)
(KRISHNASRI, FLIPKART,10)
```

6. Results

Store is used to save results to the file system.

```
grunt>store store1 into „./outputresults“;  
Input(s)  
  
Successfully read 18 records (899bytes) from:”/store1”  
Output(s):  
  
Successfully stored 7 records (61 bytes) in :”/outputresults
```

7. Conclusion and Future Work

We presented the first process to storing, retrieving the data and results from the hadoop cluster data nodes. The query is executed on static data on cluster nodes on few records on data file. We research that ability to work a single query in multiple data nodes without a interrupt. We will give future work for query evaluation, performance and graph theory on dynamic data on cluster nodes.

REFERENCES

1. Marcin Wylot, Philippe Cudre´-Mauroux, Manfred Hauswirth, and Paul Groth “Storing, Tracking, and Querying Provenance in Linked Data” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 29, NO. 8, AUGUST 2017
2. A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker, “LDIF- linked data integration framework,” in Proc. 2nd Int. Conf. Consuming Linked Data, 2011, pp. 125–130
3. M. Schmachtenberg, C. Bizer, and H. Paulheim, “ Adoption of the linked data best practices in different topical domains,” in The Semantic Web, P. Mika, et al., Eds. Berlin, Germany: Springer, 2014, pp. 245–260. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11964-9_16
4. M. Wylot, P. Cudre´-Mauroux, and P. Groth, “Executing provenance-enabled queries over Web data,” in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 1275–1285.
5. M. Wylot, P. Cudre´-Mauroux, and P. Groth, “A demonstration of TripleProv: Tracking and querying provenance over Web data,” Proc. VLDB Endowment, vol. 8, no. 12, pp. 1992–1995, 2015.
6. A. Zimmermann, N. Lopes, A. Polleres, and U. Straccia, “A general framework for representing, reasoning and querying with annotated Semantic Web data,” Web Semantics, vol. 11, pp. 72–95, Mar. 2012.

7. O. Hartig, “Querying trust in RDF data with tSPARQL,” in Proc. 6th Eur. Semantic Web Conf. Semantic Web: Res. Appl., 2009, pp. 5–20.
8. F. Geerts, G. Karvounarakis, V. Christophides, and I. Fundulaki, “Algebraic structures for capturing the provenance of SPARQL queries,” in Proc. 16th Int. Conf. Database Theory, 2013, pp. 153–164.
9. L. Moreau and I. Foster, “Electronically querying for the provenance of entities,” in Provenance and Annotation of Data, L. Moreau and I. Foster, Eds. Berlin, Germany: Springer, 2006, pp. 184–192.
10. O. Biton, S. Cohen-Boulakia, and S. B. Davidson, “Zoom*UserViews: Querying relevant provenance in workflow systems,” in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 1366–1369.
11. L. M. Gadelha, Jr, M. Wilde, M. Mattoso, and I. Foster, “MTCProv: A practical provenance query framework for many-task scientific computing,” *Distrib. Parallel Databases*, vol. 30, no. 5/6, pp. 351–370, Oct. 2012.
12. A. Chebotko, S. Lu, X. Fei, and F. Fotouhi, “RDFProv: A relational RDF store for querying and managing scientific workflow provenance,” *Data Knowl. Eng.*, vol. 69, no. 8, pp. 836–865, Aug. 2010.
13. G. Karvounarakis, Z. G. Ives, and V. Tannen, “Querying data provenance,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2010, pp. 951–962.
14. B. Glavic and G. Alonso, “The perm provenance management system in action,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 1055–1058.
15. B. Arab, D. Gawlick, V. Radhakrishnan, H. Guo, and B. Glavic, “A generic provenance middleware for queries, updates, and transactions,” in Proc. 6th USENIX Workshop Theory Practice Provenance, Jun. 2014. [Online]. Available: <https://www.usenix.org/conference/tapp2014/agenda/presentation/arab>
16. J. Widom, “Trio: A system for integrated management of data, accuracy, and lineage,” Tech. Rep., Stanford InfoLab, Aug. 2004.
17. H. Halpin and J. Cheney, “Dynamic Provenance for SPARQL Updates,” in *The Semantic Web*, P. Mika et al., Eds. Berlin, Germany: Springer, 2014, pp. 425–440. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11964-9_27

Acknowledgement

The authors would like to thank the anonymous reviewers for their careful reading and insightful comments that have helped in improving of this paper.

About the Authors



M.Sreerama Murthy pursuing Ph.D in Acharya Nagarjuna University,Guntur M.Tech in Computer Science and Engineering from University College of Engineering

,JNTU,Kakinada.B.Tech in Information Technology from JNTU,Hyderabad. And now presently working as Assoc Professor in Vikas Group of Institutions,Vijayawada .His research areas includes Data Mining and Big Data



Dr. N. Naga Malleswara Rao , working as Professor in Department of CSE, RVR & JC College of Engineering Chowdavaram, Guntur(Dt),Andhra Pradesh, India. He was 27 years of teaching experience and published few national and international journals and also attended national and international conferences. His research areas includes computer algorithms, compilers, image processing and data mining

Survey on various Process Models in Software Development

Mrs. K. Lalitha Kumari, Asst.Professor, Department of Computer Science and Engineering, AAR Mahaveer Engineering College

Ms. A. Manasa, Asst.Professor, Department of Computer Science and Engineering, AAR Mahaveer Engineering College

ABSTRACT

Software has been a significant part of modern society for a long time. In particular, this paper is concerned with various software development process models. Software process model is a description of the sequence of activities carried out in a software engineering project, and the relative order of these activities. This paper presents a comparative study of various process models in software development based on their applications and different aspects of each model to help the developers to select specific model at specific situation depending on customer demand.

Keywords: Software development, Process models, SDLC, Software Engineering

1. INTRODUCTION:

A software development process, also known as a software development life cycle (SDLC), is a structure imposed on the development of a software product. It is often considered as a subset of system development life cycle. The development models are the various processes or methodologies that [1] are being selected for the development of the project depending on the project's aims and goals. There are many development life cycle models that have been developed in order to achieve different required objectives. The models specify the various stages of the process and the order in which they are carried out. The selection of model has very high impact on the testing that is carried out. It will define the what, where and when of our planned testing, influence regression testing and largely determines which test techniques to use.

2. SOFTWARE PROCESS MODEL

A Process model will provide a way in which manner we are gathering requirements, how coding and testing techniques [5] are going to be performed finally how the product is delivered to the customer.

2.1 Waterfall Model:

This model is one of the oldest models and is widely used in government projects and in many major companies. As this model emphasizes planning in early stages, it ensures design flaws before they develop. The model begins with establishing system requirements and software requirements and continues with architectural design, detailed design, coding, testing, and maintenance. [2] The waterfall model serves as a baseline for many other lifecycle models.

Applications

- Requirements are very well documented, clear and fixed.
- Product definition is stable.
- Technology is understood and is not dynamic.
- There are no ambiguous requirements.
- The project is short.

2.2 Spiral Model

The model divided in four phases: Planning, Risk Analysis, Engineering and Evaluation. A software project frequently passes through these phases in iterations (called Spirals in this model). At the initial spiral, starting with the planning, requirements are gathered and risk is considered. Each consequent spiral builds on the initial spiral. Requirements are collect during the planning phase. [3]In the risk analysis phase, a process is going on to identify risk and their alternate solutions. A prototype is produced at the end of the risk analysis. The evaluation phase permits the customer to assess the output of the project to date before the project goes to the next spiral.

Applications

- When there is a budget constraint and risk evaluation is important.
- For medium to high-risk projects.
- Long-term project commitment because of potential changes to economic priorities as the requirements change with time.

- Customer is not sure of their requirements which is usually the case.
- Requirements are complex and need evaluation to get clarity.
- New product line which should be released in phases to get enough customer feedback.
- Significant changes are expected in the product during the development cycle.

2.3 RAD Model

In this model the components or functions are developed in parallel as if they were mini projects. The development are time boxed, delivered and then bring together into a working prototype. This can quickly give the customer something to see and use and to provide feedback regarding the delivery and their requirements.

Applications

- RAD should be used only when a system can be modularized to be delivered in an incremental manner.
- It should be used if there is a high availability of designers for modeling.
- It should be used only if the budget permits use of automated code generating tools.
- RAD SDLC model should be chosen only if domain experts are available with relevant business knowledge.
- Should be used where the requirements change during the project and working prototypes are to be presented to customer in small iterations of 2-3 months.

2.4 V-Model

In this model Each phase is compulsory to complete before the next phase begins.

Testing is highlighted in this model more than the waterfall model. The testing actions are developed early in the life cycle before any coding is done, during each of the phase's previous implementation. In this model testing is spotlighted on meeting the functionality specified in requirements gathering. [4]The main design phase spotlight on system architecture and design. The combine test plan is created in this phase in order to test the pieces of the software systems capacity to work jointly. Though, the low-level design phase lies where the real software components are designed, and unit tests are created in this phase. The accomplishment phase is, again, where all coding is generated. After coding is complete, the way of execution continues up the right side of the V where the test plans developed earlier can use.

Applications

- Requirements are well defined, clearly documented and fixed.
- Product definition is stable.
- Technology is not dynamic and is well understood by the project team.
- There are no ambiguous or undefined requirements.
- The project is short.

2.5 Agile Model

Agile Methods break the product into small incremental builds. These builds are provided in iterations. Each iteration typically lasts from about one to three weeks. [5]Every iteration involves cross functional teams working simultaneously on various areas i.e Planning, Requirements, Analysis, Design, Coding, Testing. At

the end of the iteration, a working product is displayed to the customer and important stakeholders.

Applications

- Agile model is useful when system have rapidly new changes are needed to be implementation required. The freedom agile gives to change is very important. New changes can be implemented at very less cost because of the frequency of new increments that are produced.
- When to implement a new feature and the development team required losing only the work of a few days, or even only hours, to roll back and implement it.
- When both system developers and stakeholders alike to interact with the under development project.
- Applicable when developer required more freedom of time and options than if the software was developed in a more rigid sequential way.
- Provide ability to the authorized development team member about to leave important decisions until more or better data or even entire hosting programs are available; meaning the project can continue to move forward without fear of reaching a sudden standstill.

3. CONCLUSION

A study is given about different development models and their applications. Waterfall model which provides base for other development models. Then its enhanced models are explained in Iterative model, Spiral model, V shaped model and finally, Agile development model. The study includes the applications of different models which can help to select specific model at

specific situation depending on customer demand.

4. REFERENCES

[1] CTG. MFA – 003, "A Survey of System Development Process Models", Models for ActionProject: Developing Practical Approaches to Electronic Records Management and Preservation, Center for Technology in Government University at Albany / Sunny, 1998.

[2] Steve Easterbrook, "Software Lifecycles", University of Toronto Department of Computer Science, 2001.

[3] Analysis of various Software Process Models Ashwini Mujumdar, Gayatri Masiwal, P. M. Chawan /International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, May-Jun 2012, pp.2015-2021

[4] Roger Pressman, Software Engineering: A Practitioner's Approach, Sixth Edition, McGraw-Hill Publication.

[5] I. Sommerville, "Software Engineering", Addison Wesley, 7th edition, (2004).

A COMPLETE SURVEY ON ITERATIVE CLUSTERING METHODS

M.VIJAYALAXMI, Assistant Professor, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AAR MAHAVEER ENGINEERING COLLEGE

A.MANASA Assistant Professor, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AAR MAHAVEER ENGINEERING COLLEGE

ABSTRACT

A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. A cluster is an aggregation of points in the space such that the *distance* between two points in the cluster is less than the distance between any point in the cluster and any point not in it. Clusters may be described as connected regions of a multidimensional space containing a relatively *high density* of points, separated from other such regions by a region containing a relatively low density of points.” The intention of this report is to present a special class of clustering algorithms, namely iterative-based methods. After the introduction and a review on iterative relocation clustering algorithms and partition around medoids and multivariant .

Keywords:

Iterative, k-mean, medoids, centroids, clustering, multi-dimensionality

1.Introduction

Clustering can be considered as the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. The intention of this report is to be an introduction into specific parts of this methodology called cluster analysis. So called partitioning-based clustering methods are flexible methods based on iterative relocation of data points between clusters. The quality of the solutions is measured by a clustering criterion. At each iteration, the iterative relocation algorithms reduce the value of the criterion function until convergence. by changing the clustering criterion, it is possible to construct robust clustering methods that are more insensitive to outliers and missing

The multi-dimensionality of the data objects (observations, records etc.). This is an important notion, since the grouping of objects that possess more than three variables is no easy matter for a human being without automated method. Naturally, most of the definitions address the notion of similarity. Similarity is one of the key issues of cluster analysis, which means that one of the most influential elements of cluster analysis is the choice of an appropriate similarity measure.

2. literature survey

Typically clustering methods yield a data description in terms of clusters that possess strong internal similarities [8]. Often one defines the cluster in terms of internal cohesion (homogeneity) and external isolation (separation). Hence, the cluster is often simply considered as a collection of objects, which are similar to one another within the same cluster and dissimilar to the objects in other clusters [12]. An interesting connection to the software engineering is recognized, when we notice that the principle is very similar with the common software architecture rule on "*loose coupling and strong cohesion*". Such architecture aims to localize effects caused by code modifications (see, e.g., Bachmann et al. [3]). The software components with a large number of mutual links can be considered close to each other. Hence, a good software architecture should contain clearly separated "component clusters".

Some common definitions are collected from the clustering literature and given

"A Cluster is a set of entities which are alike, and entities from different clusters are not alike."

"A cluster is an aggregation of points in the space such that the *distance* between two points in the cluster is less than the distance between any point in the cluster and any point not in it."

“Clusters may be described as connected regions of a multidimensional space containing a relatively *high density* of points, separated from other such regions by a region containing a relatively low density of points.

“Although the cluster is an application dependent concept, all clusters are compared with respect to certain properties: density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data points when compared to the other areas of space. From compactness and tightness, it follows that the degree of dispersion (variance) of the cluster is small. The shape of the cluster is not known a priori. It is determined by the used algorithm and clustering criteria. Separation defines the degree of possible cluster overlap and the distance to each other. Fuzzy clustering methods produce overlapping clusters by assigning the degree of the membership to the clusters for each point. Traditional partitioning clustering methods, such as K-Means, and hierarchical methods produce separated clusters, which means that each data point is assigned to only one cluster. A cluster is defined in a dimension of its variables and, if having a round shape, it is possible to determine its radius. These are the measurable features for any cluster, but it is not possible to assign universal values or relations to them. Perhaps, the most problematic features are shape and size.

3. Iterative-based algorithms

The aim of the partition-based algorithms is to decompose the set of objects into a set of disjoint clusters where the number of the resulting clusters is predefined by the user. The algorithm uses an iterative method, and based on a distance measure it updates the cluster of each object. The most representative partition-based clustering algorithms are

- k-Means

- Partition Around Medoids(K-Medoids)
- CLARA
- CLARANS

The advantage of the partition-based algorithms that they use an iterative way to create the clusters, but the drawback is that the number of clusters has to be determined in advance and only spherical shapes can be determined as clusters.

3.1 K-means algorithms

K-means type grouping has a long history. For instance, already in 1958, Fisher[10] investigated this problem in one-dimensional case as a *grouping problem*. At that time, algorithms and computer power were still insufficient for larger-scale problems, but the problem was shown to be interesting with concrete applications. Hence, more efficient procedures than exhaustive search was needed. The seminal versions of the K-means procedure were introduced in the Sixties by Forgy [11] (c.f. discussion in [16]) and MacQueen [16] (see also [1] and [5]). These are perhaps the most widely used versions of the K-means algorithms [13, p.112]. The main difference between Forgy's and MacQueen's algorithms is the order, in which the data points are assigned to the clusters and the cluster centers are updated. The MacQueen's K-means algorithm updates the "winning" cluster center immediately after every assignment of a data point and all cluster centers one more time after all data points have become assigned to the clusters. The Forgy's method updates the cluster centers only after all data points are assigned to the closest cluster centers. Moreover, another difference is that the Forgy's method iterates until converged while the MacQueen's basic algorithm performs only one complete pass through data. The starting points of the MacQueen's algorithm are often the first K data points in the data set.

In 1982, Lloyd [15] presented a quantization algorithms for pulse-code modulation (PCM) of analog signals. The algorithm is often referred to as *Lloyd's algorithm* and it is actually equivalent with the Forgy's K-means algorithm in a scalar case. Although Lloyd's paper was published not until 1982, the unpublished manuscript from 1957 is referred, for example, in articles from 1977 and 1980 by Chen [7] and Linde et al. [14], respectively¹. A basically similar algorithm for multidimensional cases was presented by Chen in [7]. Linde et al. Generalized the Lloyd's algorithm to a vector quantization algorithm [14]. This algorithm is often referred to as the *Generalized Lloyd's Algorithm* (GLA) in signal and image processing context. Hence, two main types of the K-means method have been discovered more than once on different disciplines.

The time complexity of the Forgy's K-means is $(npKt)$ (t is the number of iterations) [9]. A convergence proof for the algorithm is given by Selim et al. [17]. Because of the algorithmic details, the MacQueen's and Forgy's algorithms are also referred to as online- and batch-K-means algorithms, respectively (see, e.g., [6, 18]). One should note that many times, as in [6, 18], the convergent variant [1] of the MacQueen's K-means algorithm is behind the online clustering, although the MacQueen's basic K-means algorithm is referred. In [6], the numerical experiments suggest that the online K-means algorithm converges faster during the first few passes through the data and, thereafter, batch version outperforms it. However, the online clustering may be useful in real-time applications, which have to respond to inputs in extremely short time, or receive the data in a stream of unknown length, or if there is not enough memory available to store a data set as a complete block [4].

Drawbacks

Drawbacks of the ordinary K-means algorithms, there are some significant defects

that have led to development of numerous alternative versions during the past years :

Sensitivity to initial configuration. Since the basic algorithms are local search heuristics and K-means cost function is non-convex, it is very sensitive to the initial configuration and the obtained partition is often only suboptimal (not the globally best partition).

Lack of robustness. As the sample mean and variance are very sensitive estimate against outliers. So-called breakdown point is zero, which means that one gross errors may distort the estimate completely. The obvious consequent is that the k-means problem formulation is highly non-robust as well.

Unknown number of clusters. Since the algorithm is a kind "flat" or "non-hierarchical" method [9], it does not provide any information about the number of clusters.

Empty clusters. The Forgy's batch version may lead to empty clusters on unsuccessful initialization.

Order-dependency. The MacQueen's basic and converging variants are sensitive to the order in which the points are relocated. This is not the case for the batch versions.

Only spherical clusters. K-means presumes the symmetric Gaussian shape for cluster density functions. From this it follows that a large amount of clean data is usually needed for successful clustering.

- *Handling of nominal values.* The sample mean is not defined for nominal values.

In order to solve the previous problems many variants for the original versions have been developed.

3.2 Enhanced variants of K-means algorithm

It seems that the development of the clustering algorithms has been very intensive during the sixties. As we know, the rapid development of PC computer systems

during the eighties and still growing data storages led to the invention of knowledge discovery and data mining concepts. It seems that this developed has led again to the growing interest in clustering algorithms. Hence, a lot of variants for the traditional K-means algorithms have emerged during the last ten years. Many of these try to solve the known drawbacks of the K-means procedures. K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The main idea is to define k centroids, one for each cluster. The better choice is to place the Centroids as much as possible far away from each other. This algorithm aims at minimizing an objective function.

<p>1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.</p> <p>Assign each object to the group that has the closest centroid.</p> <p>When all objects have been assigned, recalculate the positions of the K centroids.</p> <p>Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be</p>

3.3 k-Medoids Clustering Algorithm

The k-means algorithm is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data. Instead of taking the mean value of the objects in a cluster as a reference point, the medoid can be used, which is the most centrally located object in a cluster. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster.

The algorithm is composed of the following steps

Input: The number of clusters k and a database containing n objects

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Method:

Arbitrarily choose k objects as the initial medoids;

- Repeat
- Assign each remaining object to the cluster with the nearest medoid
- Randomly select a non medoid object, o_{random}
- Compute the total cost, S of swapping o_j with o_{random}
- If $S < 0$ then swap o_j with o_{random} to form the new set of k medoid
- Until no change

3.4 Comparison k-means and k-medoids

The k-medoids method is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k-mean method.

3.4 From k-medoids to CLARANS

A typical k-medoids partitioning algorithm works effectively for small data sets, but does not scale well for large data sets. To deal with larger data sets, a sampling-based method, called Clara (clustering large applications) can be used.

The idea behind CLARA is as follows:

Instead of taking the whole set of data into consideration, a small portion of the actual data is chosen as a representative of the data. Medoids are then chosen from this sample Partitioning Around Medoids. If the sample is selected in a fairly random manner, it should closely represent the original data set. The representative objects (medoids) chosen will likely be similar to those that would have been chosen from the whole data set. Clara draws multiple samples of the data set, applies PAM on each sample, and returns its best clustering as the output. [1]

The effectiveness of CLARA depends one the sample size. Notice that PAM

searches for the best k medoids among a given data set, whereas CLARA searches for the best k medoids among the selected sample for the data set. CLARA cannot find the best clustering if any sampled medoid is not among the best k medoids. A k -medoids type algorithm called CLARANS (Clustering Large Applications based upon RANdomized Search) was proposed that combines both sampling technique with PAM. However, unlike CLARA, CLARANS does not confine itself to any sample at any given time. While CLARA has a fixed sample with some randomness in each step of the search, CLARANS draws a sample with some randomness in each step of the search. The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of k medoids. The clustering obtained after replacing a single medoid is called the neighbor of the current clustering. If a better neighbor is found, CLARANS moves to the neighbor's node and the process starts again; otherwise the current clustering produces a local optimum. [2]

3.5 Clustering on multivariate

There is a broad group of multivariate analyses that have as their objective the organization of individual observations (objects, sites, individuals), and these analyses are built upon the concept of multivariate distances (expressed either as similarities or dissimilarities) among the objects.

The organization generally takes two forms:

- the arrangement of the objects in a lower-dimensional space than the data were originally observed on;
- the development of "natural" clusters or the classifications of the objects.

These analyses share many concepts and techniques (both numerical and practical) with other procedures such as principal components analysis, numerical taxonomy, discriminant analysis and so on.

The analyses generally begin with the construction of an $n \times n$ matrix \mathbf{D} of the distances between objects. For example, in a two dimensional space, the elements d_{ij} of \mathbf{D} could be the Euclidean distances between points,

$$d_{ij} = [(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2]^{1/2}$$

The Euclidean distance, and related measures are easily generalized to more than two dimensions.

3.5.1 Basic distances

As an example of the calculation of multivariate distances, the following script will calculate the Euclidean distances, in terms of pollen abundance, among a set of (modern) pollen surface-samples in the Midwest that were used for fitting regression equations for reconstructing past climates from fossil-pollen data. (Note: because the pollen data were transformed by taking the square roots of pollen abundance data, what is actually being calculated here is the so-called “squared chord-distance” or SCD.)

The distance matrix, each element of which displays the distance between two points in “pollen space” (as opposed to geographical space) can be displayed using the `image()` function. The looping in the code below is inefficient, but illustrates what is going on.

3.5.2 Mahalanobis distances

The basic Euclidean distance treats each variable as equally important in calculating the distance. An alternative approach is to scale the contribution of individual variables to the distance value according to the variability of each variable. This approach is illustrated by the Mahalanobis distance, which is a measure of the distance between each observation in a multidimensional cloud of points and the centroid of the cloud. The Mahalanobis distance D^2 is given by

$$D^2 = (\mathbf{x} - \mathbf{m})\mathbf{V}^{-1}(\mathbf{x} - \mathbf{m})$$

where \mathbf{x} is a vector of values for a particular observation, \mathbf{m} is the vector of means of each variable, and \mathbf{V} is the variance-covariance matrix.

The following code illustrates the calculation of Mahalanobis distances in a “climate space” described by two climate variables from the Midwest pollen-climate data set. The graduate circle around each point is proportional to the Mahalanobis distance between that point and the centroid of scatter of points.

4. Conclusion

The choice of clustering algorithm depends both on the type of data available and on the particular purpose and application. The partition based algorithms work well for finding spherical- shaped clusters in small to medium-sized databases.

The k-medoids method is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. But its processing is more costly than the k-means method. The k-medoids method works effectively for small data sets, but does not scale well for large data sets. To deal with larger data sets, a sampling-based method, called CLARA can be used. The effectiveness of CLARA depends on the sample size. CLARA cannot find the best clustering if any sampled medoid is not among the best k medoids. CLARANS is the most effective partitioning method among all. It enables the detection of outliers. For the future enhancement, these algorithms are combined together to form the hybrid algorithm which is more efficient to form the clusters than all other algorithms. if we use mahalanobis distance formulae instead of Euclidean distance then we can find efficient clustering for multivariate objects

5. References

- [1] M. R. ANDERBERG, *Cluster analysis for applications*, Academic Press, Inc., London, 1973.
- [2] L. ÁNGEL GARCÍA-ESCUADERO AND A. GORDALIZA, *Robustness properties of k means and trimmed k means*, *Journal of the American Statistical Association*, 94 (1999), pp. 956–969.
- [3] F. BACHMANN AND L. BASS, *Managing variability in software architectures*, in *SSR '01: Proceedings of the 2001 symposium on Software reusability*, New York, USA, 2001, ACM Press, pp. 126–132.
- [4] A. BARALDI AND P. BLONDA, *A survey of fuzzy clustering algorithms for pattern recognition I*, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 29 (1999), pp. 778–785.
- [5] P. BERKHIN, *Survey of clustering data mining techniques*, tech. report, Accrue Software, San Jose, CA, 2002.
- [6] L. BOTTOU AND Y. BENGIO, *Convergence properties of the K-means algorithms*, in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, eds., vol. 7, The MIT Press, 1995, pp. 585–592.
- [7] D. CHEN, *On two or more dimensional optimum quantizers*, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '77.*, vol. 2, Telecommunication Training Institute, Taiwan, Republic of China, May 1977, pp. 640–643.
- [8] R. DUDA AND P. HART, *Pattern Classification and Scene analysis*, John Wiley & Sons, Inc., NY, 1973.

[9] R. O. DUDA, P. E. HART, AND D. G. STORK, *Pattern classification*, John Wiley & Sons, Inc., 2001

[10] W. D. FISHER, *On grouping for maximum homogeneity*, Journal of the American Statistical Association, 53 (1958), pp. 789–798.

[11] E. FORGY, *Cluster analysis of multivariate data: Efficiency versus interpretability of classifications*, Biometrics, 21 (1965), pp. 768–769. Abstracts in Biometrics.

[12] J. HAN AND M. KAMBER, *Data mining: concepts and techniques*, Morgan Kaufmann Publishers, Inc., 2001.

[13] L. KAUFMAN AND P. J. ROUSSEEUW, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, 1990.

[14] Y. LINDE, A. BUZO, AND R. GRAY, *An algorithm for vector quantizer design*, IEEE Transactions on Communications, 28 (1980), pp. 84–95.

[15] S. P. LLOYD, *Least squares quantization in PCM.*, IEEE Transactions on Information Theory, 28 (1982), pp. 129–136.

[16] J. MACQUEEN, *Some methods for classification and analysis of multivariate observations*, in Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297

[17] S. SELIM AND M. ISMAIL, *K-means-type algorithms: A generalized convergence theorem and characterization of local optimality*, PAMI, 6 (1984), pp. 81–87.

[18] A. SHADEMAN AND M. ZIA, *Adaptive vector quantization of mr images using*

on-line k-means algorithm, in Proceedings of SPIE, 46th Annual Meeting, Application of Digital Image Processing XXIV Conference, vol. 4472, San Diego, CA, USA, July-August 2001, pp. 463–470.

Big Data Analysis

P.Sowmya sree
Assistant professor
Computer science department
AV College of arts, science and commerce

Abstraction:

Big data is very large, distributed aggregations of loosely structured data that often incomplete and inaccessible. On a broad scale, data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance. Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs.

Keywords: data scientists, modelers, statisticians, business intelligence.

What is Big data?

Big data analytics is the often complex process of examining large and varied data sets -- or big data -- to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions. On a broad scale, data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance.

Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems.

Evolution of big data analytics

The concept of big data has been around for years; most organizations now understand that if they capture all the data that streams into their businesses, they can apply analytics and get

significant value from it. But even in the 1950s, decades before anyone uttered the term “big data,” businesses were using basic analytics (essentially numbers in a spreadsheet that were manually examined) to uncover insights and trends.

The new benefits that big data analytics brings to the table, however, are speed and efficiency. Whereas a few years ago a business would have gathered information, run analytics and unearthed information that could be used for future decisions, today that business can identify insights for immediate decisions. The ability to work faster – and stay agile – gives organizations a competitive edge they didn’t have before.

Why is big data analytics important?

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things.

Emergence and growth of big data analytics

The term *big data* was first used to refer to increasing data volumes in the mid-1990s. In 2001, Doug Laney, then an analyst at consultancy Meta Group Inc., expanded the notion of big data to also include increases in the variety of data being generated by organizations and the velocity at which that data was being created and updated. Those three factors -- volume, velocity and variety -- became known as the 3Vs of big data, a concept Gartner popularized after acquiring Meta Group and hiring Laney in 2005.

Separately, the Hadoop distributed processing framework was launched as an Apache open source project in 2006, planting the seeds for a clustered platform built on top of commodity hardware and geared to run big data applications. By 2011, big data analytics began to take a firm hold in organizations and the public eye, along with Hadoop and various related big data technologies that had sprung up around it.

Initially, as the Hadoop ecosystem took shape and started to mature, big data applications were primarily the province of large internet and e-commerce companies such as Yahoo, Google and Facebook, as well as analytics and marketing services providers. In the ensuing years, though, big data analytics has increasingly been embraced by retailers, financial services firms, insurers, healthcare organizations, manufacturers, energy companies and other enterprises.

Big data analytics technologies and tools

Unstructured and semi-structured data types typically don't fit well in traditional data warehouses that are based on relational databases oriented to structured data sets. Further, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently -- or even continually, as in the case of real-time data on stock trading, the online activities of website visitors or the performance of mobile applications.

As a result, many of the organizations that collect, process and analyze big data turn to NoSQL databases, as well as Hadoop and its companion tools, including:

- **YARN:** a cluster management technology and one of the key features in second-generation Hadoop.
- **MapReduce:** a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
- **Spark:** an open source, parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.

- **HBase:** a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
- **Hive:** an open source data warehouse system for querying and analyzing large data sets stored in Hadoop files.
- **Kafka:** a distributed publish/subscribe messaging system designed to replace traditional message brokers.
- **Pig:** an open source technology that offers a high-level mechanism for the parallel programming of MapReduce jobs executed on Hadoop clusters.

-

How big data analytics works

In some cases, Hadoop clusters and NoSQL systems are used primarily as landing pads and staging areas for data before it gets loaded into a data warehouse or analytical database for analysis -- usually in a summarized form that is more conducive to relational structures.

More frequently, however, big data analytics users are adopting the concept of a Hadoop data lake that serves as the primary repository for incoming streams of raw data. In such architectures, data can be analyzed directly in a Hadoop cluster or run through a processing engine like Spark. As in data warehousing, sound data management is a crucial first step in the big data analytics process. Data being stored in the Hadoop Distributed File System must be organized, configured and partitioned properly to get good performance out of both extract, transform and load (ETL) integration jobs and analytical queries.

Once the data is ready, it can be analyzed with the software commonly used for advanced analytics processes. That includes tools for data mining, which sift through data sets in search of patterns and relationships; predictive analytics, which build models to forecast customer behavior and other future developments; machine learning, which taps algorithms to analyze large data sets; and deep learning, a more advanced offshoot of machine learning.

Text mining and statistical analysis software can also play a role in the big data analytics process, as can mainstream BI software and data visualization tools. For both ETL and analytics

DST Sponsored National Conference on Recent Advancements on Computer
Science (CONRACS 2019)- 26 to 28 July 2019

applications, queries can be written in MapReduce, with programming languages such as R, Python, Scala, and SQL, the standard languages for relational databases that are supported via SQL-on-Hadoop technologies.

References: <https://searchbusinessanalytics.techtarget.com>

Challenges in Product Development using Scrum Model

Dr. K. Sunil Manohar Reddy¹, Mrs. G. Pratibha²

¹Associate Professor, Dept. of CSE, Matrusri Engineering College, Hyderabad.

²Assistant Professor, Dept. of CSE, Matrusri Engineering College, Hyderabad.

¹sunil186@gmail.com, ²pratibhareddy19@gmail.com

Abstract

Software Product Development plays a very important role in today's world. Recently more and more software development organizations are using agile software development methods and techniques for product development. Successful agile adoption leads to producing higher quality software, enhance developers moral and at a lower cost than the conventional software development approaches. This study seeks to evaluate, synthesize and present aspects of research on Scrum model which is an agile method, approaches adopted and the criteria utilized for agile practice selection.

Keywords: Prototyping, Agile Software Development, Scrum, Backlog, Agile Testing

1. Introduction

Software Engineering is the platform to develop the software step by step and to produce a high quality product at the end. There are many models to implement Software Engineering. All the models are actually based on Software Development Life Cycle (SDLC). The SDLC has phases or steps to be followed to achieve software development. These steps are customized to form a process model which will help the software engineers to follow a certain path. Some process models are Waterfall model, Spiral model, Prototyping, iterative and incremental model, rapid application development and agile model. Agile software development approach enables developing software in regular intervals, i.e., iterations, producing the software in increments^{1, 2}. This research concentrates on the major issues and challenges in Scrum method of agile model. Scrum is a process framework to deliver products with the highest possible value and handle complex problems or situations. Use iterative and incremental approaches to develop products using cross-functional teams^{1, 2}. The issues are listed and described below.

2. Issues with Prioritizing the user Stories

2.1 Assigning Product Backlog Items

In scrum method Product Backlog Item is central. The PBI contains a prioritized list of all items relevant to a specific product. Bugs, customer requested enhancements, competitive product functionality, competitive edge functionality and technology upgrades are available in this list³. The first major issue is to convert user stories into Product Backlog Items. There is lots of confusion in assigning PBIs from the Software Requirement Specification document. There are cases in reality wherein either PBI are assigned directly from the functionalities of the software that is requested by the client or the Scrum team assigning the PBIs or in some cases PBIs are assigned based on the sprint duration^{4, 5}. These cases are not the right method to assign PBI, because the functionality of software can be a PBI but not all functionalities are PBIs. A scrum team cannot assign PBI since the team does not have insights on the project, only the Product Owner and Scrum Master are responsible for assigning PBI. The assigning of PBI is the first task of Scrum, if that is not properly assigned then the whole project leads to failure or inappropriate delivery of increments.

In Scrum, the user-stories are converted into Product Backlog Items (PBI). These PBIs have to be prioritized so that it will benefit both the client and the company. Prioritization of user-stories is a difficult task in agile environment because of constant changes in the user stories. PBIs have to be prioritized with respect to some characteristics. The client and the company have their own priorities for the PBIs^{5, 6}. The method for prioritization of the PBI by the company for instance may be based on the criticality of the PBI, type of scrum team available to work on a particular PBI or which PBI will have a better software reuse. The client priorities are based on the business value of the PBI and the need for a PBI. These two different priorities might not be the same, so in that case there is a need to adopt a prioritization technique which values and also addresses the characteristics of both the client's and the company's priorities. Such a technique is not available and is a major issue in Scrum.

3. Distributed Environment

In a distributed scrum, spontaneous flow of communication gets compromised having no feasibility of sitting in the same work-room. Communication factors like Distance, Time Difference, Cultural Difference, Language, etc⁷ keeps on generating issues every day. There is no proper mechanism to communicate between Scrum teams placed in different geographical locations during the Sprint planning and analyzing phases. Team cannot be put together under a single roof in the current scenario. Pairing of off shore member and onsite member can solve this problem but the problem is, it's kind of a mentor-mentee relationship which can be done only between two members. This issue is faced by most of the software companies which are implementing Scrum. "Productivity Theft" is another challenge particularly for remote team which means engagement of members into several "beyond scrum" activities without knowledge of Scrum Master sometimes by offshore management and sometimes by even onsite members. The Scrum teams have to involve in various meetings like Scrum Planning Meeting, Scrum Review Meeting, Scrum Retrospective Meeting, etc. The Scrum team members have to participate in these meetings in every sprint cycle. Teams in

various geographical areas are the ones affected by this issue. There is no proper or practical solution for this issue.

4. Regression Testing

Regression testing is testing of software after it has been modified, to check whether the existing functionality gets affected or not. It is one of the most expensive activities that occur as the software is developed and maintained. Studies shows that a significant portion of development and maintenance costs go to this regression testing, which is known as *retesting*. Reports state that regression testing uses 80% of the overall testing cost and can use up to 50% of the cost of software maintenance^{8,9}. Rapidly changing software and computing environments present many challenges for effective and efficient regression testing in practice. Regression testing can be performed after the changes have been made to the software, before the new version of the software is released every time the software is saved and compiled. In an agile development environment or before patches, such as security patches are released. The goals of regression testing are to improve the confidence that the changes behave as required and that they have not affected the previous parts of the software. Since Scrum is about adapting changes of the software that is being developed, it is a must to test the software with regression test after the changes have been made. The regression testing is implemented on the every integration of PBI for the final delivery. It is important to develop PBI in such a way that it does not create lots of bugs on integration with other PBI. But in reality it is highly unlikely in cases of complex projects, which subsequently increase the cost and time of regression test, which in turn directly affects the agile delivery of software because of increased time taken to solve bugs to provide high quality software. It is a major issue among other issues as it directly affects the cost and time. There is no proper solution for this problem.

5. Integration of PBI

In Scrum method of agile methodology, after the completion of the development of PBI, it will be delivered to the client as an increment. After completion and delivery of all the PBIs, the Scrum team has to integrate each and every PBI. Here the significant issue is that, the analyzers need to perform relapse testing for every last combination, which builds the testing cost and time. More than 80% of the expense of testing is utilized just for relapse testing^{8,9}. No proper model or method is currently available for “how to integrate PBIs with reduced regression testing”.

6. Idle Team Members

During the identification and analysis of PBIs, the other team members like developers, testers and maintenance providers will be idle and it is a waste of time for the team. As in agile, there should not be any wastage in time otherwise it will affect the delivery of the PBI.

7. The Developer-Tester Problem

In Scrum method, for a particular sprint the starting time and time to complete the PBI will be planned and allocated. The developers and testers have to perform according to that time. Suppose if developer couldn't complete a PBI in a given time then it will also affect the testers because the testers cannot start testing before the completion by developers^{9, 10}. This results in problem between developers and testers and there is no proper technical or generic solution for this problem yet.

8. Testing Issues in Scrum

Scrum life cycles are becoming common and every life cycle affects testing. Testing is done in each sprint cycle, when the PBI is developed it is immediately tested and released to client and when all the PBIs are developed and delivered, the final step is to integrate all the PBI and a final testing is implemented. The quantity of blunders¹¹ is likewise subject to the individual or the group of persons building up the product. In the event that the designers are experienced there are lesser possibilities of blunders. There may be special cases to this. Most of the issues in Scrum are the issues which are related to testing such as Volume and Speed of change, Inconsistent and Inadequate unit testing and many meetings.

Conclusion

These are the major issues and problems faced by software companies when they implement Scrum method. These issues do not have any proper technical or generic solutions to solve. Each software company is trying and implementing some methods which they believe will solve these problems on a temporary basis. But this customized method does not lead in the right path always.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

References

- [1] Campanelli AS, Parreiras FS. Agile methods tailoring – A systematic literature review 2015. *The Journal of Systems and Software*. 2015 Dec; 110:85–100.
- [2] Othmane LB, Angin P, Weffers H, Bhargava B. Extending the agile development process to develop acceptably secure software. *Dependable and Secure Computing. IEEE Transactions*. 2014 Nov-Dec; 11(6):497–509. DOI:10.1109/TDSC.2014.2298011.
- [3] K. Sunil Manohar Reddy, “A Review on Agile Unified Process Model”, *IJRECE*, Volume: 7, Issue: 2, pg. 2768-2770, June 2019.
- [4] Danevaa M, Veena EVD, Amrita C, Ghaisasb S, Sikkela K, Kumarb R, Ajmerib N, Ramteerthkarb U, Wieringaa R. Agile requirements prioritization in large-scale outsourced system projects: An empirical study. *The Journal of Systems and Software*. 2013 May; 86(5):1333–53.
- [5] Popli R, Chauhan N, Sharma H. Prioritizing user stories in Agile Environment. *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. 2014 Feb 7-8. p. 515–19. DOI:10.1109/ICICT.2014.6781336.
- [6] K. Sunil Manohar Reddy, “Requirements Engineering: An Overview”, *IJRECE*, Volume: 7, Issue: 2, pg. 2930-2939, June 2019.
- [7] Ghosh GK. Challenges in Distributed Scrum. *IEEE Seventh International Conference on Global Software Engineering*. 2012. p. 200. DOI: 10.1109/ICGSE.2012.46.
- [8] Parsons D, Susnjak T, Lange M. Influences on regression testing strategies in agile software development environments. *Software Quality*. 2014 Dec; 22(4):717–39. DOI: 10.1007/s11219-013-9225.
- [9] Anita, Chauhan N. A Regression test selection technique by optimizing user stories in an agile development. *IEEE International Advance Computing Conference (IACC)*. 2014 Feb 21-22; 1454–58. DOI:10.1109/IAdCC.2014.6779540.
- [10] Ekbal R, Srikanta P, Vandana B. Estimation and Evaluation of Change in Software Quality at a Particular Stage of Software Development. *Indian Journal of Science and Technology*. 2013 Oct; 6(10).
- [11] Akif R, Majeed H. Issues and challenges in Scrum implementation. *International Journal of Scientific and Engineering Research*. 2012 Aug; 3(8):1–4.1. ISSN 2229-5518.

A WEB BASED APPROACH FOR HOUSE CONSTRUCTION

1. B. SANJANA 2. G. YOGITHA 3. VIJAY KUMAR

1. STUDENT- JB INSTITUTE of ENGINEERING AND TECHNOLOGY
2. STUDENT- JB INSTITUTE of ENGINEERING AND TECHNOLOGY
3. ASSISTANT PROFESSOR- DEPT. IT- JB INSTITUTE of ENGINEERING AND TECHNOLOGY

ABSTRACT- To construct a new house who don't have any idea or plan. This is the solution for constructing a new house according to peoples view. A person may have many ideas running out of his/her mind but incompetent to produce good output. This is the way where collaboration of all objectives for the construction includes information of engineers, workers, carpenters and also stores like sanitation and hardware, marble store, painting store etc. Cost estimation can also be provided according to 3BHK/2BHK/1BHK

INTRODUCTION

House is a place for protect people from outside influences, such as climate, wild animals, bad people, disease danger and so on. For using as the function, a house must be equipped with various facilities needed, such as electricity, clean water, ventilation, storage of important documents, sewerage. Along with the rate of population growth in Indonesia the need for homes is always increasing, therefore the construction continues to be done either massively by the developer or sole done by individuals, and a lot of people in Indonesia still do the construction their house by themselves or individual.

Everytime build the house, it is necessary to ensure the feasibility of investments in order to estimate the cost of construction at the planning stage is very important as has been disclosed. "Furthermore, cost estimating has functions with a very wide spectrum of planning and controlling resources, such as materials, labor, and equipment. Although the uses are same, but the emphasis for each of the participating organizations are different.

From all of the description cost estimating methods above have been recognized by scientists in their field and widely used in Indonesia either by Ministry of Public Works, developer or contractor, but from all referenced methods all require experience ,specific engineering scholarship and human expertise as estimators based on their

experience, the construction of a house has a different complexity than the school buildings or other buildings, according to the authors need to do research on the method of estimating the cost of housing construction, because the characteristics of society in Indonesia vary both socially, culturally and economically. Indonesian society is many variants, still very much we are witness the people in Indonesia who do the construction of his house without involving the expert estimator, many done alone without estimating the previous costs, due to economic limitations and the lack of engineering science, so still need to do research in order to produce a method that is easily understood By the people in Indonesia without having to have experience and engineering science specifically.

Based on the literature study, experience and direct observation on the construction of the residence that the variables that affect the cost of housing construction is the component of the house construction that is; 1) Foundation, 2)Structure, 3) Wall, 4) door frame, 5) Roof, 6) Mechanical and electrical, 7) Sanitary, 8) Floor, 9) Ceiling, 10) Painting, [1, 3, 4, 7].

The greater the volume of construction components will be the greater the cost of development. Construction components are arranged so as to form a building consisting of rooms, there is a closed room and there is open space, enclosed space between them, room, living room, dining room, family room, kitchen, bathroom and others, Open space like terrace and carport. From the engineering arrangement of construction components that form the rooms will eventually be known the building area.

II. LITERATURE SURVEY

Various authors have defined project risk in different way. Project risk has positive and negative effects on project objectives (Adetokunbo, Ogunsemi, Aje, Awodele, & Dairo, 2013; Wang, Dulaimi, & Aguria, 2004). It is the measure of the probability, severity and the exposure to all hazards of an activity (Sarkar & Panchal, 2015). Risk is closely connected to uncertainty and is a commonly used term in all kinds of contexts but is often related to the negative outcome of a certain event.

Moreover, risks have stochastic nature (Hamzaoui, Taillandier, Mehdizadeh, Breyse, & Allal, 2015).

There are different causes of risks in construction such as size, organizational and technical complexities, speed of construction, location of the project, technology being used and familiarity with the work (Dey & Ogunlana, 2004). In addition to the organizational and technical complexities, project managers have to consider a growing number of parameters (e.g. environmental, social, safety and security) and stakeholders, both inside and outside the project. The complexity of a project leads to the existence of a network of interdependent risks (Fang & Marle, 2012), where complex phenomena may occur, hard to anticipate and hard to keep under control (Fang & Marle, 2013).

III. SYSTEM ANALYSIS

3.1 Functional Requirements

The functional requirement refers to the system needs in an exceedingly computer code engineering method. The key goal of determinant “functional requirements” in an exceedingly product style and implementation is to capture the desired behavior of a software package in terms of practicality and the technology implementation of the business processes.

1. Good communicative social network with user friendly.
2. Load the information by the owner to the users.
3. Search for the objectives required to the user.
4. Select the objective and contact with the owner

3.2 Nonfunctional requirements

All the other requirements which do not form a part of the above specification are categorized as Non-Functional needs. A system perhaps needed to gift the user with a show of the quantity of records during info. If the quantity must be updated in real time, the system architects should make sure that the system is capable of change the displayed record count at intervals associate tolerably short interval of the quantity of records dynamic. Comfortable network information measure may additionally be a non-functional requirement of a system.

1. performance.
2. speed.
3. cost of storage and maintenance should be affordable.

3.3 software Requirements

The software requirements specify the use of all required software products like data management system. The required software product specifies the numbers and version. Each interface specifies the purpose of the interfacing software as related to this software product

1. Operating system: windows 10.
2. Database: MySQL
3. Webserver: XAMPP.
4. Web technologies: HTML,CSS,PHP.
5. client Applications: Chrome

3.4 Hardware Requirements

The hardware requirement specifies each interface of the software elements and the hardware the elements of the system. These hardware requirements include configuration characteristics

1. processor : Intel(R) Core(TM) i5
2. Hard Disk : 60GB
3. RAM : 4GB

The Systems Development Life Cycle (SDLC), or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies that people use to develop these systems. In software engineering the SDLC concept underpins many kinds of software development methodologies

3.5 Existing system

At present if we want to construct a house with our own ideas, we have to consult all workers, shops and etc and consulting them is a very huge and time taking process. Who don't have any idea how to initiate house construction, they have to consult everyone to take data about house construction.

Existing system disadvantages

- People are unable to fulfill their dreams without having any idea, how to construct a house.
- people waste lot of time in consulting different shops and workers.

3.6 Proposed system

A website with different objectives which are needed for house construction. The objectives like engineers, bricks shop, marble shop, workers and etc. The user can contact the required objective owner.

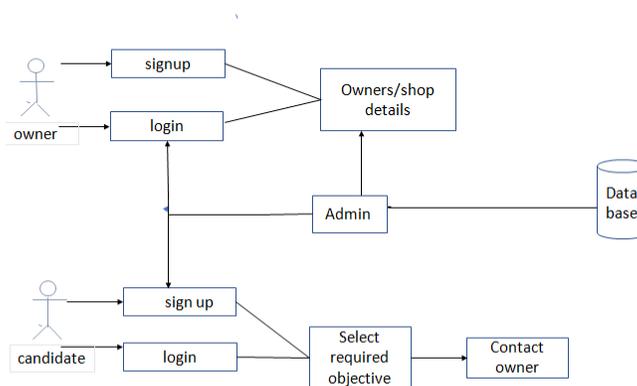
Proposed system disadvantages

- People can construct their dream home with the help of this website.
- They can see details of different shops and different persons in website like plumbers, electricians, engineers etc.

This makes easy in fulfilling peoples dreams and makes easy

IMPLEMENTATION

System architecture:



Introduction to Modules

There are three modules in the system

1. Admin.
2. Candidate.
3. Owner.

Admin:

Admin is the controller of website. In this module, admin is the main authority of this site .So he has all privileges to upload, delete files and update of any account. Admin will perform all controlling activities are performed by admin.

candidate:

In this module the candidate can get registered providing their personal details and get logged into their candidate home page, from where they can access the data provided by admin.

Owner:

Many owners were present in this. Owners fill their details of their shops and candidate can able to see owner's data and able to contact with owners.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

The web-based approach for house construction is successfully completed. The goal of the system is achieved, and the problems are solved where the users can contact in a short period. This application triumphs the over reducing gap between a user and owner .this project is developed in a manner that is user friendly.

The primary objective is to provide the interactive service to all the general users in this contemporary world. Different types of objectives are provided in this application which are used for house construction.

FUTURE SCOPE

1. This web application can be improving by adding extra features like Developing this application in Android
2. By including the GPS service for efficient results.
3. By adding nearest shops in case of Emergency, users can get there.
- 4.By comparing same type of shop to know the best shops in the city.

DIABETES DISEASE PREDICTION USING DATA MINING

K .Sankeerthana , K .Lasya, B.Ashritha, V.Vidhu

JB institute of engineering&Technology

Sankeerthanakancharla4@gmail.com, lasyakashetty777@gmail.com Bompallyashritha1999@gmail.com, Vidhu.venishetty123@gmail.com

ABSTRACT:

Data mining is a subfield in the subject of software engineering. It is the methodical procedure of finding examples in huge data sets including techniques at the crossing point of manufactured intelligence, machine learning, insights, and database systems. The goal of the data mining methodology is to think data from a data set and change it into a reasonable structure for further use. Our examination concentrates on this part of Medical conclusion learning design through the gathered data of diabetes and to create smart therapeutic choice emotionally supportive network to help the physicians. The primary target of this examination is to assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, we propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease.

Index Terms— Disease, Diabetes, Prediction., Naïve Bayes, KNN

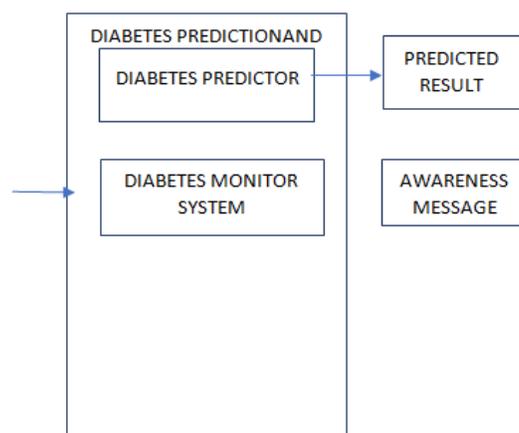
INTRODUCTION:

Data mining is the investigation of expansive datasets to separate covered up and beforehand obscure examples, connections and information that are hard to recognize with conventional measurable techniques. The territories where data mining is connected as of late incorporate designing, showcasing, human services and monetary anticipating. Data mining in social insurance is also a rising field of high significance for giving what we can say is high anticipation and a more profound comprehension of restorative data. The amount of accessibility of tremendous measure of patient's data which can be used to extricate valuable information, scientists have been utilizing data mining methods to help medicinal services experts in analysis of ailments. In the usually higher part of the papers, the diabetes forecast system chips away at a little dataset, however our point is to deal with expansive dataset. The quantity of medicinal test required may influence to execution of system in this way we additionally concentrate on diminishing the therapeutic test. It relies on upon which parameter or quality is taken in the system for foreseeing diabetes. Our expectation system will take a shot at a bigger dataset and number of therapeutic testing test required will overcome. Our system utilizes two calculations which we will apply on the same dataset for anticipating diabetes

LITERATURE REVIEW:

Following is some of the search which has been reviewed for the proposed system: - Sadegh et al. [6] have proposed, this system that comes under the category of data

mining. The system performs data mining on patterns and correlation to predict the economic events. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and also economic and financial institutions. Mohamed EL Kourdi et al. [3] have proposed this system in which Naive Bayes (NB) which is a factual machine learning algorithm is utilized to order Arabic web documents. This system utilizes K-Nearest Neighbor for estimating values that will maintain a strategic distance from financial distress and bankruptcy. In the current review k-Nearest Neighbor characterization technique, have been examined for economic estimating. Lately, after the situation of worldwide financial emergency, the quantity of bankrupt organizations has risen. Since organizations' financial distress is the principal phase of bankruptcy, utilizing financial proportions for anticipating financial distress have pulled in a lot of consideration of the scholastics and economic and financial institutions. Marius et al. [5] have proposed this system that implements rather fast generating nearest neighbor and appropriate algorithm configuration. Kevin Beyer et al. [10] have proposed this system that tries to explain what happens when dimensionality increases.



You expect to discover the class of the blue star (BS). BS can either be RC or GS and that's it. The "K" is KNN calculation is the closest neighbors we wish to take vote from.

1. Take patient dataset
2. Filter the data according to the requirements
3. Read the excel file of patient dataset
4. Perform normalization on the resultant dataset and execute
5. Perform naïve Bayesian or knn on the resultant dataset and execute
6. Obtain predicted result and awareness message.

ALGORITHMS

A. Working of naïve Bayes:

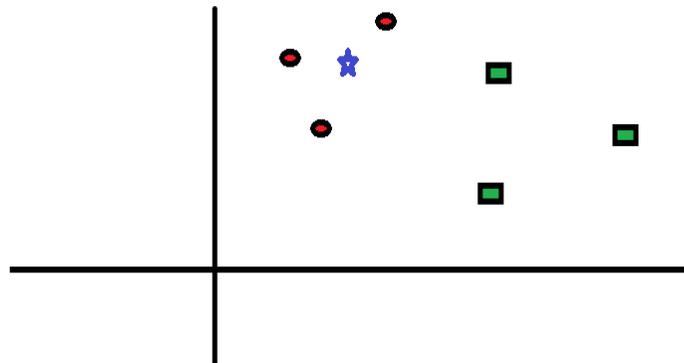
Steps for Solving Naïve Bayes: -

Step 1: Conversion of training set into a frequency table. Step 2: Creating what is called Likelihood table that finds the probability which is like Overcast probability = 0.29 and the probability of playing is 0.64.

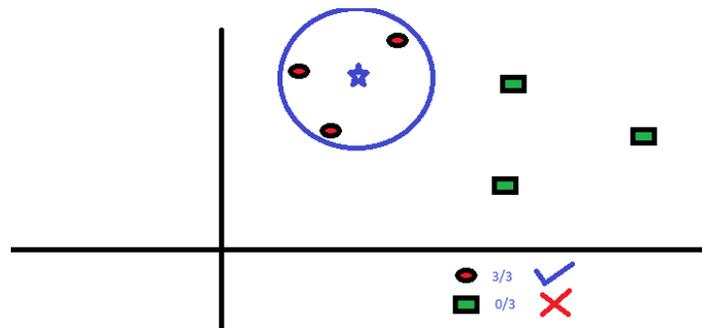
Step 3: Now, using Naive Bayesian equation, calculate probability for each possible class. The class that has the highest probability among the others is the result of prediction.

B. Working of k-nearest neighbor:

Steps For KNN: The example is in below Fig. 2 to acknowledge this algorithm. Following is a wide spread of red circles (RC) and green squares (GS):



Suppose $K = 3$. Consequently, construct a hover with BS as focus similarly as large as to encase just three data points on the same plane. Allude to taking after graph for more points of interest.



The three nearest focuses in above Fig 3 to BS is all RC. Subsequently, with great certainty level we can state that the BS ought to have a place with the class RC. Here, the decision turned out to be exceptionally clear as every one of the three votes from the nearest neighbor went to RC. The decision of the parameter K is exceptionally important in this algorithm.

Bayesian model is anything but difficult to manufacture and especially helpful for substantial data sets. Alongside effortlessness, Naive Bayes is known to beat even profoundly advanced grouping techniques. Bayes theorem provides a simple method of calculating posterior probability $P(c \text{ in } x)$ from $P(c)$, $P(x)$ and $P(x \text{ in } c)$.

Look at the equations below:

$$P(C \text{ in } X) = P(X \text{ in } C) P(C) / P(X)$$

$P(C \text{ in } x)$ is the posterior probability of class (C, target) given predictor (x, attributes).

$P(C)$ is the prior probability of class.

$P(X \text{ in } C)$ is the likelihood which is the probability of predictor given class.

$P(X)$ is the prior probability of predictor

EXPECTED RESULT:

The goal of our project is to know whether patient is diabetic or not, patient will be diagnosed and it will be depending on the attributes that we are going to take, such as age, pregnancy, pg concentration, tri fold thick, serum ins, body mass index (bmi), dp

function, diastolic bp i.e. the factors which are majorly responsible for diabetes. So, to reduce the correctly know whether the patient is diabetic or not, we are developing a system which will be a prediction system for the diabetes patients. Another best thing about the system is it is will give accurate results whether the patient is diabetic or not with the help of the knowledge base of the larger dataset that we are going to use added the recommendations we are going to provide based on the diabetic levels of the patients. Also, the prediction of the disease will be done with the help of Bayesian algorithm and K-NN algorithm.

CONCLUSION:

By our in-depth analysis of literature survey, we acknowledged that the prediction done earlier did not use a large dataset [12]. A large dataset ensures better prediction. Also what it lacks is recommendation system. When we predict we will give some recommendation to the patient on how to control or prevent diabetes in case of minor signs of diabetes. The recommendations would be such, that when followed it will help the patient. Thus we will build up a system which will anticipate diabetic patient with the assistance of the Knowledge base which we have of dataset of around 2000 diabetes patients and furthermore to give suggestions on the premise of the nearness of levels of diabetes patients. Prediction will be done with the help of two algorithms Naïve Bayes and K-Nearest Neighbor and also we will compare which algorithm gives better accuracy on the basis of their performance factors. This system which will be developed can be used in HealthCare Industry for Medical Check of diabetes patients

FUTURE ENHANCEMENT:

The proposed system can be developed in many different directions which have vast scope for improvements in the system. These includes: 1. Increase the accuracy of the algorithms. 2. Improvising the algorithms to add more efficiency of the system and

enhance its working. 3. Working on some more attributes so to tackle diabetes even more. 4. To make it as a complete healthcare diagnosis system to be used in hospitals.

DEEP LEARNING APPLICATIONS IN MRI-based IMAGE ANALYSIS

¹Vijay Kumar.B

²B Deepthi Reddy

¹ Assistant Professor, Department of IT, J. B Institute Of Technology, Hyderabad, Telangana.

² Assistant Professor, Department of IT, J. B Institute Of Technology, Hyderabad,
Telangana.

ABSTRACT

Brain tumor segmentation is an important task in medical image processing. Early diagnosis of brain tumors plays an important role in improving treatment possibilities and increases the survival rate of the patients. Manual segmentation of the brain tumors for cancer diagnosis, from large amount of MRI image sgenerated in clinical routine, is a difficult and time consuming task. There is a need for automatic brain tumor image segmentation. The purpose of this paper is to provide a review of MRI-based brain tumor segmentation methods. Recently, automatic segmentation using deep learning methods proved popular since these methods achieve the state-of-the-art results and can address this problem better than other methods. Deep learning methods can also enable efficient processing and objective evaluation of the large amounts of MRI-based image data. There are number of existing review papers, focusing on traditional methods for MRI-based brain tumor image segmentation. Different than others, in this paper, we focus on the recent trend of deep learning methods in this field. First, an introduction to brain tumors and methods for brain tumor segmentation is given. Then, the state of- the-art algorithms with a focus on recent trend of deep learning methods are discussed. Finally, an assessment of the current state is presented and future developments to standardize MRI-based brain tumor segmentation methods into daily clinical routine are addressed.

Introduction:

Cancer can be defined as the uncontrolled, unnatural growth and division of the cells in the body. Occurrence, as a mass, of these unnatural cell growth and division in the brain tissue is called a brain tumor. While brain tumors are not very common, they are one of the most lethal cancers¹.

Depending on their initial origin, brain tumors can be considered as either primary brain tumors or metastatic brain tumors. In primary ones, the origin of the cells are brain tissue cells, where in metastatic ones cells become cancerous at any other part of the body and spread into the brain. Gliomas are type of brain tumors that originate from glial cells.

They are the main type of brain tumors that current brain tumor segmentation research focuses on. The term glioma is a general term that is used to describe different types of gliomas ranging from low-grade gliomas like astrocytomas and oligodendrogliomas to the high grade (grade IV) glioblastoma multiform (GBM), which is the most aggressive and the most common primary malignant brain tumor. Surgery, chemotherapy and radiotherapy are the techniques used, usually in combination, to treat gliomas.

Early diagnosis of gliomas plays an important role in improving treatment possibilities. Medical Imaging techniques such as Computed Tomography (CT), Single-Photon Emission Computed Tomography (SPECT), Positron Emission Tomography (PET), Magnetic Resonance Spectroscopy (MRS) and Magnetic Resonance Imaging (MRI) are all used to provide valuable information about shape, size, location and metabolism of brain tumors assisting in diagnosis. While these modalities are used in combination to provide the highest detailed information about the brain tumors, due to its good soft tissue contrast and widely availability MRI is considered as the standard technique. MRI is a non-invasive in vivo imaging technique that uses radio frequency signals to excite target tissues to produce their internal images under the influence of a very powerful magnetic field.

Images of different MRI sequences are generated by altering excitation and repetition times during image acquisition. These different MRI modalities produce different types of tissue contrast images, thus providing valuable structural information and enabling diagnosis and segmentation of tumors along with their subregions. Four standard MRI modalities used for glioma diagnosis include T1-weighted MRI (T1), T2-weighted MRI (T2), T1-weighted MRI with gadolinium contrast enhancement (T1-Gd) and Fluid Attenuated Inversion Recovery (FLAIR). During MRI acquisition, although can vary from device to device, around one hundred and fifty slices of 2D images are produced to represent the 3D brain volume. Furthermore, when the slices from the required standard modalities are combined for diagnosis the data becomes very populated and complicated. Generally, T1 images are used for distinguishing healthy tissues, whereas T2 images are used to delineate the edema region which produces bright signal on the image. In T1-Gd images, the tumor border can easily be distinguished by the bright signal of the accumulated contrast agent (gadolinium ions) in the active cell region of the tumor tissue.

Since necrotic cells do not interact with the contrast agent, they can be observed by hypo intense part of the tumor core making it possible to easily segment them from the

active cell region on the same sequence. In FLAIR images, signal of water molecules are suppressed which helps in distinguishing edema region from the Cerebrospinal Fluid (CSF). Before applying any therapy, it is crucial to segment the tumor in order to protect healthy tissues while damaging and destroying tumor cells during the therapy. Brain tumor segmentation involves diagnosing, delineating and separating tumor tissues, such as active cells, necrotic core and edema from normal brain tissues including Gray Matter (GM), White Matter (WM) and CSF. In current clinical routine, this task involves manual annotation and segmentation of large amount of multimodal MRI images. However, since manual segmentation is a very time consuming procedure, development of robust automatic segmentation methods, to provide efficient and objective segmentation, became an interesting and popular research area in recent years. Current high segmentation performances obtained by deep learning methods make them good candidates for achieving this task. The rest of the paper is organized as follows: First we briefly review methods for brain tumor image segmentation in one section. Then, in another section, we especially focus on methods based on deep learning algorithms, which provide the state-of-the-art results in recent years. In particular, we compare designs of different deep learning methods and their performances.

SYSTEM ANALYSIS

Existing System

There is a myriad of imaging modalities, and the frequency of their use is increasing. Smith-Bindman et al. looked at imaging use from 1996 to 2010 across six large integrated healthcare systems in the United States, involving 30.9 million imaging examinations. The authors found that over the study period, CT, MRI and PET usage increased respectively.

The symbolic AI paradigm of the 1970s led to the development of rule-based, expert systems. One early implementation in medicine was the MYCIN system by Short life, which suggested different regimes of antibiotic therapies for patients. Parallel to these developments, AI algorithms moved from heuristics-based techniques to manual, handcrafted feature extraction techniques, and then to supervised learning techniques. Unsupervised machine learning methods are also being researched, but the majority of the algorithms from 2015-2017 in the published literature have employed supervised learning methods.

Drawbacks of Existing System

In image processing processing techniques used different types of filters and

Fourier and discrete transform it increases the complexity the cost of those equipment also high.

To know the result of the tumor concerned person has to be there.

This diagnosis perform some particular equipment only

PROPOSED SYSTEM

Detection, sometimes known as Computer-Aided Detection is a keen area of study as missing a lesion on a scan can have drastic consequences for both the patient and the clinician. The task for the Kaggle Data Science Bowl of 2017 involved the detection of cancerous lung nodules on CT lung scans. Approximately 2000 CT scans were released for the competition and the winner Fangzhou achieved a logarithmic loss score of 0.399. Their solution used a 3-D CNN inspired by U-Net architecture to isolate local patches first for nodule detection. Then this output was fed into a second stage consisting of 2 fully connected layers for classification of cancer probability. Shin *et al.* evaluated five well-known CNN architectures in detecting thoracoabdominal lymph nodes and Interstitial lung disease on CT scans. Detecting lymph nodes is important as they can be a marker of infection or cancer. They achieved a mediastinal lymph node detection AUC score of 0.95 with a sensitivity of 85% using GoogLeNet, which was state of the art. They also documented the benefits of transfer learning, and the use of deep learning architectures of up to 22 layers, as opposed to fewer layers which was the norm in medical image analysis. Overfeat was a CNN pre-trained on natural images that won the ILSVRC 2013 localization task . Ciompi applied Overfeat to 2-dimensional slices of CT lung scans oriented in the coronal, axial and sagittal planes, to predict the presence of nodules within and around lung fissures. They combined this approach with simple SVM and RF binary classifiers, as well as a Bag of Frequencies, a novel 3-dimensional descriptor of their own invention.

Advantages of Proposed System

1. It not only helps us in predicting the outcome but also gave us valuable insights about the nature of data, which can be used in future to train our classifiers in a much better way.
2. The data obtained can also be used as training data in future

LITERATURE SURVEY

Kharrat et al (2009) introduced an efficient detection of brain tumor from cerebral MRI images. The methodology consists of three steps: enhancement, segmentation and classification. To improve the quality of images and limit the risk of distinct regions fusion in the segmentation phase an enhancement process is applied. Mathematical morphology was adapted to increase the contrast in

MRI images. Then Wavelet Transform was applied in the segmentation process to decompose MRI images. At last, the k-means algorithm is implemented to extract the suspicious regions or tumors. Some of experimental results on brain images show the feasibility and the performance of the proposed approach.

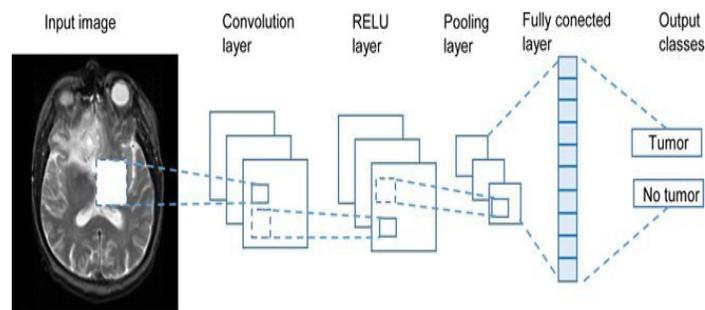
Akram and Usman (2011) proposed a method for automatic brain tumor diagnostic system from MR images. The system consists of three stages to detect and segment a brain tumor. In the first stage, MR image of brain is acquired and preprocessing is done to remove the noise and to sharpen the image. In the second stage, global threshold segmentation is done on the sharpened image to segment the brain tumor. In the third stage, the segmented image is post processed by morphological operations and tumor masking in order to remove the false segmented pixels. Results and experiments show that this technique accurately identifies and segments the brain tumor in MR images.

Leung et al (2003) proposed a new approach to detect the boundary of brain tumor based on the Generalized Fuzzy Operator (GFO). Boundary detection in MR image with brain tumor is an important image processing technique applied in radiology for 3D reconstruction. The non homogeneities in density tissue of the brain with tumor can result in achieving the inaccurate location in any boundary detection algorithms. Some studies using the contour deformable model with regional base technique, show that the performance is insufficient to obtain the fine edge in the tumor, and the considerable error in accuracy exist. Moreover, even in some of the normal tissue region, edge created by this method has also been encompassed. One typical example is used for evaluating this method with the contour deformable model.

Perner (2002) discussed about image mining framework as a standard tool and its application to medical-image analysis. A tool and a technique for data mining in picture-archiving systems are provided. It is expected to determine the suitable knowledge for picture examination and identification from the data base of image descriptions. Knowledge-engineering methods are used to acquire a list of attributes for symbolic image descriptions. An expert describes images based on this list and accumulates descriptions in the database.

Jaba Sheela and Shanthi (2007) described the image mining approaches for categorization and segmentation of brain MRI data. Image segmentation plays a vital role in several medical imaging applications by computerizing or assisting the description of anatomical arrangements and additional regions of interest. Automatic recognition of tumors in several medical images is encouraged by the

requirement of better accuracy when handling with a human life. Also, the computer assistance is demanded in medical institutions owing to the reality that it possibly will progress the results of humans in such a domain where the false negative cases must be at a very low rate. It has been confirmed that double reading of medical images possibly will show the way for enhanced tumor detection. But the cost implied in double reading is extremely huge, that's why better software to assist humans in medical institutions is of interest at the present time. In their approach they developed a system which uses image mining approaches to categorize the images either as normal or abnormal and then divide the tissues of the anomalous Brain MRI to recognize brain related diseases.



IMPLEMENTATION

MODULES

“Deep Learning Applications In Medical Field Analysis” mainly consists of four modules:

- Importing Dataset
- Splitting the data-types
- Building neural networks
- Prediction

MODULE DESCRIPTION

Importing Dataset:

Initial step in our project is to import image dataset which are having both tumor contained images, and non tumor x-ray images.

Splitting the data:

After importing the dataset need to split the data into two parts like training and testing datasets. Based on split size need to split the data.

Building neural networks:

We need to feed the convolution neural networks with trained with different layers need to give, after that giving the testing the data for analyzing the result

Prediction:

Finally give on MRI image to neural networks it will checks that weather that image having tumor or not and then gives the result

Algorithms:

Convolutional Neural Networks:

Currently, CNNs are the most researched machine learning algorithms in medical image analysis. The reason for this is that CNNs preserve spatial relationships when filtering input images. As mentioned, spatial relationships are of crucial importance in radiology, for example, in how the edge of a bone joins with muscle, or where normal lung tissue interfaces with cancerous tissue. As shown in, a CNN takes an input image of raw pixels, and transforms it via Convolutional Layers, Rectified Linear Unit (RELU) Layers and Pooling Layers. This feeds into a final Fully Connected Layer which assigns class scores or probabilities, thus classifying the input into the class with the highest probability.

a: Convolution Layer

A convolution is defined as an operation on two functions. In image analysis, one function consists of input values (e.g. pixel values) at a position in the image, and the second function is a filter (or kernel); each can be represented as array of numbers.

Computing the dot product between the two functions gives an output. The filter is then shifted to the next position in the image as defined by the stride length. The computation is repeated until the entire image is covered, producing a feature (or activation) map. This is a map of where the filter is strongly activated and 'sees' a feature such as a straight line, a dot, or a curved edge. If a photograph of a face was fed into a CNN, initially low level features such as lines and edges are discovered by the filters. These build up to progressively higher features in subsequent layers, such as a nose, eye or ear, as the feature maps become inputs for the next layer in the CNN architecture. Convolution exploits three ideas intrinsic to perform computationally efficient machine

learning: sparse connections, parameter sharing (or weights sharing) and equivariant (or invariant) representation . Unlike some neural networks where every input neuron is connected to every output neuron in the subsequent layer, CNN neurons have sparse

connections, meaning that only some inputs are connected to the next layer. By having a small, local receptive field (i.e., the area covered by the filter per stride), meaningful features can be gradually learnt, and the number of weights to be calculated can be

drastically reduced, increasing the algorithm's efficiency. In using each filter with its fixed weights across different positions of the entire image, CNNs reduce memory storage requirements. This is known as parameter sharing. This is in contrast to a fully

connected neural network where the weights between layers are more numerous, used once and then discarded. Parameter sharing results in the quality of equivariant representation to arise. This means that input translations result in a corresponding feature

map translation. The convolution operation is defined by the * symbol. An output (or feature map) $s(t)$ is defined below when input $I(t)$ is convolved with a filter or kernel $K(a)$.

b: Rectified Linear Unit (RELU) Layer

The RELU layer is an activation function that sets negative input values to zero. This simplifies and accelerates calculations and training, and helps to avoid the vanishing gradient problem. Mathematically it is defined as:

c: Pooling Layer

The Pooling layer is inserted between the Convolution and RELU layers to reduce the number of parameters to be calculated, as well as the size of the image (width and height, but not depth). Max-pooling is most commonly used; other pooling layers include

Average pooling and L2-normalization pooling. Max-pooling simply takes the largest input value within a filter and discards the other values; effectively it summarizes the strongest activations over a neighborhood. The rationale is that the relative location of a strongly activated feature to another is more important than its exact location

d: Fully Connected Layer

The final layer in a CNN is the Fully Connected Layer, meaning that every neuron in the preceding layer is connected to every neuron in the Fully Connected Layer. Like the convolution, RELU and pooling layers, there can be 1 or more fully connected layers depending on the level of feature abstraction desired. This layer takes the output from the

preceding layer (Convolution, RELU or Pooling) as its input, and computes a probability score for classification into the different available classes. In essence, this layer looks at the combination of the most strongly activated features that would indicate

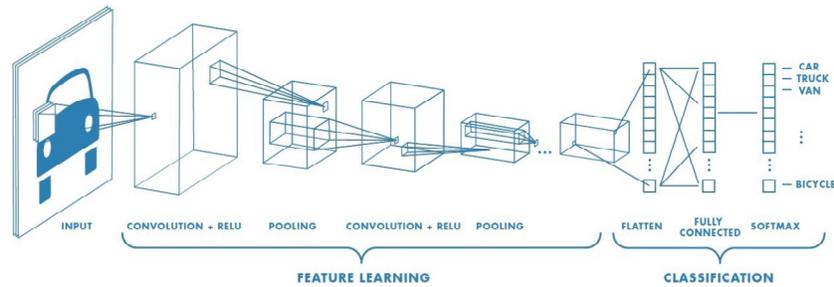
the image belongs to a particular class. For example, on histology glass slides, cancer cells have a high DNA to cytoplasm ratio compared to normal cells. If features of DNA were strongly detected from the preceding layer, the CNN would be more likely to

predict the presence of cancer cells. Standard neural network training methods with back propagation and stochastic gradient descent help the CNN learn important associations from training images.

CNN(Convolutional Neural Network) Algorithm:

Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernels), Pooling, fully connected layers (FC) and apply Soft ax function to classify an object with probabilistic values

between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.



CONCLUSION AND FUTURE SCOPE

Automatic segmentation of the brain tumors for cancer diagnosis is a challenging task. Recently, availability of public datasets and the well-accepted BRATS benchmark provided a common medium for the researchers to develop and objectively evaluate their methods with the existing techniques. In this paper, we provided a review of the state-of-the-art methods based on deep learning, and a brief overview of traditional techniques. With the reported high performances, deep learning methods can be considered as the current state-of-the-art for glioma segmentation. In traditional automatic glioma segmentation methods, translating prior knowledge into probabilistic maps or selecting highly representative features for classifiers is challenging task. However, convolutional neural networks (CNN) have the advantage of automatically learning representative complex features for both healthy brain tissues and tumor tissues directly from the multi-modal MRI images. Future improvements and modifications in CNN architectures and addition of complementary information from other imaging modalities such as Positron Emission

Tomography (PET), Magnetic Resonance Spectroscopy (MRS) and Diffusion Tensor Imaging (DTI) may improve the current methods, eventually leading to the development of clinically acceptable automatic glioma segmentation methods for better diagnosis.

BIBLIOGRAPHY

1. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for imagerecognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770_778.
2. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200_205, Jul. 2016.
3. S. M. Plis *et al.*, "Deep learning for neuroimaging: A validation study," *Front Neurosci.*, vol. 8, p. 229, Aug. 2014.

4. H. I. Suk, C. Y. Wee, S. W. Lee, and D. Shen, "State-space model with deep learning for functional dynamics estimation in resting-state fMRI," *Neuroimage*, vol. 129, pp. 292_307, Apr. 2016.
5. Z. Yan *et al.*, "Bodypart recognition using multi-stage deep learning," in *Information Processing in Medical Imaging*, vol. 24. Cham, Switzerland: Springer, Jun. 2015, pp. 449_461.
6. H. R. Roth *et al.*, "Anatomy-specific classification of medical images using deep convolutional nets," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 101_104.
7. F. Ciompi *et al.*, "Automatic classification of pulmonary perissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Med. Image Anal.*, vol. 26, no. 1, pp. 195_202, 2015.
8. F. Ciompi *et al.*, "Bag-of-frequencies: A descriptor of pulmonary nodules in computed tomography images," *IEEE Trans. Med. Imag.*, vol. 34, no. 4, pp.962_973, Apr. 2015.
9. A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115_118, 2017.
10. D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2013, pp. 411_418.
11. X. Yang *et al.*, "A deep learning approach for tumor tissue image classification," in *Proc. Int. Conf. Biomed. Eng.*, Calgary, AB, Canada, 2016. [Online]. Available: <http://dx.doi.org/10.2316/P.2016.832-025>

An Enhanced Deep Learning Method for Video Retrieval

N. Saddam Hussain, M.tech (CS), JNTU Anantapur, India,
saddamhussain0536@gmail.com

C. Shoba Bindu, Professor of CSE, JNTU Anantapur, India,
shobabindhu@gmail.com

ABSTRACT

Recently, with extremely the very large in size increasing of online videos, Fast Video Retrieval Research attracted growing consideration. The expansion of image hashing systems, As the usual video hashing techniques, it depends on the features and rendered actual-value features the binary hash codes. Therefore, low-level and high-level semantic extract features to refer to these videos in its place. This paper, retrieving video from the dataset by using image query based on deep convolution neural network is proposed to compare the frame object identity capture and high-level semantic feature extracted is involved this structure to arrive the optimization at the end of the binary hash function. In particular, our approach relies on relative image similarity to the video and then retrieving video while matching the object identity features. The outcomes determine the better worth of our proposed strategy to compare with the other video retrieval methods.

KEYWORDS

Video Retrieval, Deep Convolutional Neural Network, Hash Mapping Function, Fisher Vector.

1. INTRODUCTION

Nowadays with improvement of innovation, video is wherever on the web and the spread of mobiles, cameras, and devices, are a quickly increasing capacity of video. The related information collected in many uses are video find, approval, distribution, and promoting the websites. In recent years, the outcomes are different visual search applications have appeared as an image-to-video and image-to-image retrieval.

The issue of retrieving a peak 'k' at best appropriate video in a database of huge videos based on an image query, that is, giving an image query, our target is to retrieving the 'k' peak related videos, because based on the frames while

similarity to the image query. But suppose, If a user gives an image of Tom and Jerry he can retrieve all the related videos of Tom and Jerry if found or else we can't retrieve any video when there are no similar videos found. Other uses, a user which can find out a sports video website by giving the image. In the technology of video retrieval, an image feature extract the object identity and video frame feature extracts the object identity is based on size similarity. Traditional video retrieval techniques generally feature extracting the object identity are mainly effort, but we are using the image and video features extract the object identity to arriving the optimal- binary code, to find the convenient hash function. Ordinary hash techniques absorb the binary hash signals or code which the space interrelated for this similarity the novel input data. Locality-sensitive hash [LSH] [1] which adopting the arbitrarily estimates the map novel information into small dimensional features extract the object identity and then converting the actual-value into binary-code. To search the threshold and flagging the pixels of grayscale. I used this otus model to search the threshold. Otus model computes an optimal threshold by the minimum difference between two classes of pixels which are divided by threshold equality of intro classes. Grayscale is the image dimensionality has been decreased to the grayscale. However, the features are equally visible in two images.

Restricted Boltzmann Machines [RBM] to study compressed binary codes of input information. Iteratively quantization applies the repeated optimization approach [3] search the estimates the small binarization defeat. But this method has verified the fairly effective, binary hash code is still can't representing the input information exactly. This customary binary hash code techniques are objected identifying feature extract object identity to compare the novel information similarity, this not effectively during the variances

between to the high-level semantic similar that can observing to low- level graphical similarity.

2. RELATED WORK

Asymmetric comparison methods using Fisher vectors:

We improve approaches which check the video retrieval by using image query for retrieval structures that can use Fisher vectors. The baseline uses Fisher vector [FV] [9] collation systems are not optimized, suppose we present the technique that statement problem for configuration. We study two dissimilar irregular problems which arising the run- through, wherever the videos element capacity holds in one database cluster. To testing and display the significant performance improved by using the recommended methods.

The Fisher vector [9] a piece of stored data that indicates how other data is stored for video retrieving by using this method, but similarity the image and video are uniform, for each image and video.

With the immediate growth of many approaches depends on Deep Learning [DL]. It has been providing the view of effective and exact video recovery. The DL can an achieve complex semantically [2] attributes an incorporate lower-level graphical features based on [DCNN] Deep Convolutional Neural Networks [4] had shown an adaptable image demonstration with the robust simplification potentiality, however, makes DCNN an essential a share the effective video recovery method.

Recently, encouraged the victory of [DL] deep learning in video retrieval, Some video recovery models have been integrated hash mapping functions into a DL structure. For example, [5] applies three-tier hierarchical neural networks to learn an ability to recognize the estimated matrix.

However, This technique is not an ability to take the benefit of deep learning, According to these kinds for binary-code fewer efficient.

Further, [6] this proposed a DL an encoder-decoder structure, A two-tier [LSTM] Long Short Term Memory unit is following the binarization (Converting a pixel image to a binary image) level can straight with encoding the video extract features object identity into the binary- hash code. But this unbiased purpose for reducing error, to protect locality arrangement of information, which this significant the similarity of video retrieving.

Retrieval video founded on the DCNN deep convolutional neural networks is collecting large apperception. But hashing structure is recommended with the feature engendered and hash-function is coherent network to arrive the optimization.

Visual looking is like an index and image query graphical information, But the classify frames, according to its variants videos and database data. Large work in the scene related to the looking Image-to-Video (I2V) problem, The image query is against the video database. The Image-to-Video (I2V) problem, however, is related to growth truth refers to the looking the database videos based on image query.

3. PROPOSED APPROACH

The proposed initiate a structural of DCNN conceived for retrieval video. The training part is to design the accepts input an image. The input image query frames are selected the arbitrarily same number of frames. Given one image, the channel of the proposed structure contains few arrangements.

The unified features of every input image object identity extracting from the deep convolutional neural network and the image features extract

object identity by weighted a usual order to simplify the convolution. The complete fully connected layer is followed by the study similarity- protective binary-code, and the another complete fully layer as 'k' nodes are equal to the number of classification. The last layer part is binding the classification.

The retrieving binary hashing code produces the binarization, Which mapping binary outputs are zero/one. The binary codes for an image query and binary-codes videos are stored in a database for an finding the consistent Hamming distance and then the looking video with the maximum similar, while matching similarity features query image and video frame is same and then retrieve the video. For suppose, image query and video similarity features are different we can't retrieve the video.

In some recent research works, DCNN methods are acquired to produce visual characteristics from the middle convolutional network layers. This feature extracts the frames are computed through the forward propagation for an image query. DCNN network and aggregate the function on every convolutional network layer.

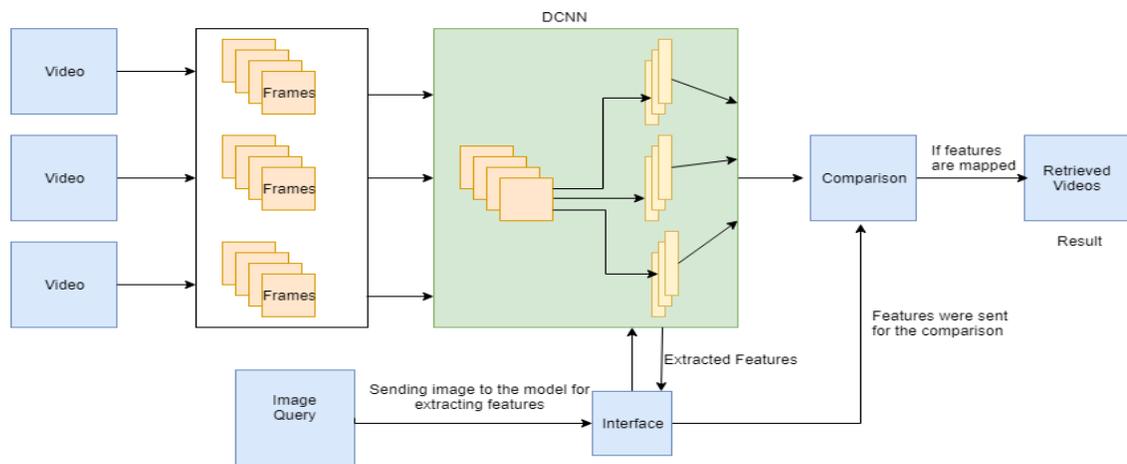


Fig. 1 Architecture of Deep Convolutional Neural Network for Video Retrievals

DCNN utilized the general-purpose feature extractors for an action other than classification. To produce deep characteristics from small video frame classification, with combine raw pixel intensities with the optical flow to extract video frames with DCNN.

Given an input image query which extracts an object identity and video is divided into frames. The Frames are extracted an object identity by using a deep Convolutional neural network. After that comparison between an image and video and video frame, an object identity similar then will be retrieval video or else, the video can't be retrieved.

4. DATASET

We observationally assess the recommend design on the UCF-101 [13] and HMDB-45

[14] datasets. UCF104 comprises about 1320 reasonable activity characteristic gathered starting with YouTube in the datasets, Hosting 104 activity Classes.

It blankets an expansive set about exercises, for example, sports, musical instruments, and human-object interactional. HMDB-45 holds something like 6,000 characteristics gathered starting with an assortment for sources going starting with digitized motion pictures on YouTube What's more need been ordered under 45 unique activity categories, each holding no less than 104 features.

5. RESULTS

Recently, It has been shown an object identify features produce based on deep convolutional neural networks (DCNN) arrive extraordinary performance in retrieval video difficulties, however when an approach is trained the classification action [7]. A comparison of such as DCNN characteristics against with FV-based method. DCNN has applied the model AlexNet [8]. Given the input image query is reduce the size 228*228, and features extract the object identity are produced from the FC3 and FC4 layers (before Rectified Linear Unit) as in preceding task [7], [8]. The frame-based test, the feature is produced by an object is identified for each frame.

The result is presented by delay and size. In this paper recommended methods enable more worth retrieval the video with considerably decrease the properties, comparison with the Frame FV and DCNN network technique.

TABLE 1
RETRIEVAL VIDEO RESULTS (mAP IN %) ON THE DATASETS,
COMPARE DCNN FEATURES EXTRACTED FOR EACH FRAME AGAINST THE FV
METHOD(USING
ASYMMETRIC COMPARISONS)

SNO	Different Techniques	mAP (%)	Latency	Memory(GB)
1	Frame Fisher Vector	70.44	0.4118	20.59
2	Scene Fisher Vector	49.71	0.1643	3.01
3	Our Method	76.54	0.106	11.25

6. PERFORMANCES ANALYSIS

As previous, we used mAP(mean Average Precision) to evaluate the quality of retrieval methods, Which are computed the rank 110 shots. The novel list of frames to produce the many exacting last lists of outcomes. This is used an evaluate quantity of appropriate database shots which is ordered amongst the upper outcomes are the primary list – irrespective ordering, hence these results are ranked [9] the succeeding period. The quantity types of outcomes, it uses the mean Recall @100. Existing approach disadvantage of mean average precision [mAP], For that considerable complex weight to the upper-rank outcomes. When the outcomes are using mean Recall @100, the shot placed at some rank the outcome list retrieving the same weight. So to overcome this drawback by using a deep convolutional neural network model, Which are computed over the number of the frames. We are against the baseline, and then generating the object identity features final list results. For this, we used an evaluated relevant database frames which are using the highest proportion of retrieving the database.

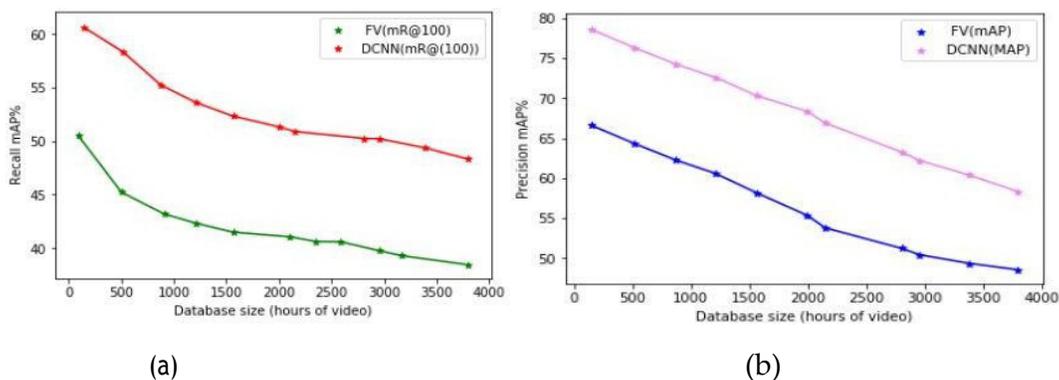


Fig 2. Retrieval video efficiency from database sizes, (a) mean Recall (R@100) (b) mean Average Precision (mAP). The top-performing Fisher Vector and DCNN are compared I2V-14M.

7. REFERENCES

- [1] Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbour in high dimensions [C]// Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on. IEEE, 2006: 459-468.
- [2] Salakhutdinov R, Hinton G. Semantic hashing[J].International Journal of Approximate Reasoning, 2009, 50(7): 969-978.
- [3] Gong Y, Lazebnik S, Gordo A, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916-2929.
- [4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural network[c] //Advances in neural information processing systems. 2012: 1097-1105.
- [5] Liong V E, Lu J, Tan Y P, et al. Deep Video Hashing[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1209-1219.
- [6] Zhang H, Wang M, Hong R, et al. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing[C]//Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016: 781-790.
- [7] Norouzi M, Fleet D J, Salakhutdinov R R. Hamming distances metrics learning[C] //Advances in neural information processing systems. 2012: 1061-1069.
- [8] Soomro K, Zamir A R, Shah M UCF101:
A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

- [9] A. Araujo, J. Chaves, R. Angst, and B. Girod “Temporal aggregation for large-scale query-by-image video retrieval,” in Proc. ICIP, Sep. 2015, pp.1519–1522.
- [10] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod. (Apr. 2016). “Large-scale Query-by-image video retrieval using Fisher Vectors.”[Online]. Available:<https://arxiv.org/abs/1604.07939>
- [11] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, “Neural codes for image retrieval,” in Proc. ECCV, 2014, pp. 584–599.
- [12] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in Proc. CVPR Workshops, Jun. 2014, pp.806–813.
- [13] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.
- [14] Kuehne H, Jhuang H, Stiefelhagen R, et al. HMDB51: A large video database for human motion recognition[M]//High Performance Computing in Science and Engineering '12. Springer, Berlin, Heidelberg, 2013: 571-582.

Multi Level KeyFrame Selection for Video Summarization

B.Sirisha, B.Sandhya

Department of Computer Science and Engineering
M.V.S.R Engineering and Technology, Nadergul,
Hyderabad, Telangana, India sirishavamsi@gmail.com.

Abstract—Due to exponential growth of video technology there is a huge multimedia content obtainable on the internet the main challenge for user is how to inspect and review rapidly these large multimedia data. Video Summarization is a Technique that permit rapid overview of multimedia data which is widely used in computer vision related applications like video browsing, video retrieval system . Video summarization aims to segment the input video to shots and extract the most informative video frames referred as key Frames. In our paper we proposed a new approach for video summarization by introducing BOW and Entropy model for extracting the informative and meaning full summary. Evaluation is done Using VSUMM Dataset by calculating fidelity category using Manhattan Distance between summarized key frames and total number of video frames.

Keywords: Video Summarization, BOW, Manhattan Distance, Fidelity.

I. INTRODUCTION

The explanation of video acquisition devices and user in-terests towards the access to video repositories have magnified the need for effectual and systematic methods in managing and retrieving such multimedia data. Structure of video is composed of a sequence of stories, each story is depicted with a sequence of video shots, and each shot is again composed of a sequence of image frames . Due to the existing continuity of the consecutive frames within a video shot, and exists redundant information among the image frames.

Video Summarization has recently interest of many re-searchers due to its importance in several applications such as information browsing and retrieval. Video summarization is a process that facilitates faster browsing of large videos and also more efficient access and indexing of content. The main objective of video summarization is to provide users with a concise video representation so that the user can have a quick idea about the content of the video[6]. The steps involved in video summarization are: converting video to frames, local feature extraction , Global feature extraction,and based on the entropy final video summary is generated.

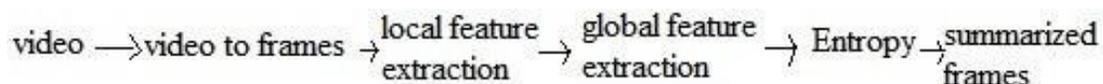


Fig. 1: introduction

There are two main video summarization techniques:

- Static video summarization [16] (video summary)

- Dynamic video summarization[10] (video skimming) Static video summarization:
- The static video summaries consist set of frames extracted from the original video, frame sets are not restricted by any timing or synchronization issues and, therefore in browsing and navigation they are more flexible and organized.

Dynamic video Summarization:

The dynamic video summaries are collection of video segments extracted from original video. Video skim include audio and motion that contains more information. In addition, it is often more entertaining and interesting to watch a skim than a slide show of frames[2]

Three principles of Video summarization for creating summary of frames should satisfy the following

- The video summary must contain high priority entities and events from the video.
- The summary itself should exhibit reasonable degrees of continuity.
- The summary should be free of repetition.

II. RELATED WORK

In this section ,some of existing methods on the video summarization, which can be found in the literature are briefly discussed in .

Wen-Sheng Chu et. al [15] proposed a video summarization technique by visual co-occurrence. The Main aim of the video co summarization is to identify the shots that occur frequently across videos of the same topic are explained in five steps. The first step, performs the video segmentation by measuring the amount of changes between two consecutive frames in two color spaces. In the second step, the collection of videos and their associated shots as a weighted bipartite graph. In the third step, performs co-clustering approach to tackle video co-summarization. In the fourth step, formulated the maximal Biclique Finding(MBF)[13]algorithm for visual co - occurrence, finally, computed the score for each shot and selected top ranked shots as the final summary.

Supriya Kamoji et. al [8] has done video summarization by implementing the use of motion activity descriptors which generate relative motion between consecutive frames. The aim of this motion activity algorithm is to provide a summary with capturing the motion of a video is explained in two stages, input video to frames and converted into gray scale. The first stage is to divide the frames into fixed number of blocks[16*16] and compares the consecutive frame blocks, this is implemented by block matching algorithm, the block matching algorithm utilizes two algorithms 1. Diamond search and 2. Three step search algorithm. the second stage is block comparison by matching algorithm to get the resultant motion activity descriptors. Diamond search has an advantage over three step search, it achieves higher precision.

Sandip [12] has given an approach for a key frame extraction based on the block based x^2 histogram difference and edge matching rate is proposed for shot boundary detection[5]. The method for the keyframe extraction consists of three stages: input a video, second stage is to calculate the block based histogram difference of each consecutive frame and extracted the edges of the candidate key frame by Prewitt operator. Finally the edges of the adjacent frames are matched.

S.Almeida et. al [1] proposed a method called VSUMM(video summarization)is propped. In the First step the video frames are pre sampled by selecting one frame per second. In the second step,the color features of video frames are extracted from the Hue component only in the HSV color space. In the third step,the meaningless frames are eliminated. In the fourth step,the frames are clustered using k-means algorithm where the number is estimated by computing the pair wise Euclidean distances between video frames and a keyframe is extracted from each cluster. Finally the key frames are compared among themselves through color histogram to eliminate that similar key frames in the produced summaries.

K.M. Mahmoud [11] has given a method for generating static video summaries by utilizing a modified agglomerative hierarchical clustering algorithm is proposed[9]. In the first step,the color features of video frames are extracted, in the second the modified dynamic modeling based hierachical clustering algorithm is done.In the third step,one frame per cluster is selected as a key frame. Finally ,the extracted key frames are arranged in the original order for the apperance of the summarization.

G. Guan et. al [7] has construct a keypoint pool for each scene and identify a set of frames which best covers the keypoint pool. The goal of keyframe selection is to best to represent a scence with a minimal number of frames for that a greedy algorithm is developed to select suitable keyframes based on the Local Coverage and Redundancy within each scene and the experimental results on the dataset , from the open video project demonstrate that the proposed method is able to achive similar results and more effective as those of the ground truth story board and the state of the art. It has advantage over them by using the keypoint based keyframe selection techniqe.

III. IMPLEMENTATION **Stage 1: Local Frame Extraction Stage 2: Global Frame Extraction Stage 3: Entropy**

1: Input video

2: Converting video to frames

3: Extracting features by using SIFT for all frames.

4: Using SIFT descriptors of all frames,calculate BOW.

5: By taking BOW histogram of all frames, calculate Euclidean Distance between two consecutive frames and generating Dissimilarity Matrix.

6: Plotting the Dissimilarity Matrix and considering the peaks of Dissimilarity Matrix as local frames formed from BOW.

Algorithm 1: Local Feature Extraction

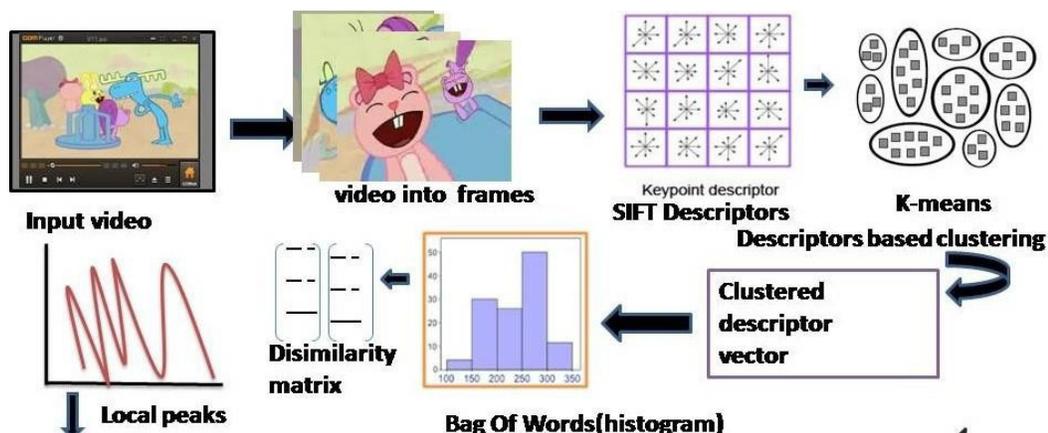


Fig. 2: Local Frame Extraction flow

- From the above global frame algorithm , we calculate Entropy of each peak frame.
- In each cluster formed in step 10 we pick Maximum Entropy value frame which is one of the summarized keyFrame of the input video.

IV. DATASET

The VSUMM1[4] Dataset contains of 60 videos each with different frame rate and frame per second.

- 1: Considering peak frames,we convert each frame into different Color space and Split in to channels.
- 2: For each channel of the frame we plot a histogram of a 50 bins and combine all the respective channels of the frame in to a vector.
- 3: Repeat above step for all peak frames and form a single vector of histogram values.
- 4: We apply K-means clustering of 15 clusters on the histogram vector and form 15 clusters.

Algorithm 2: Global Feature Extraction

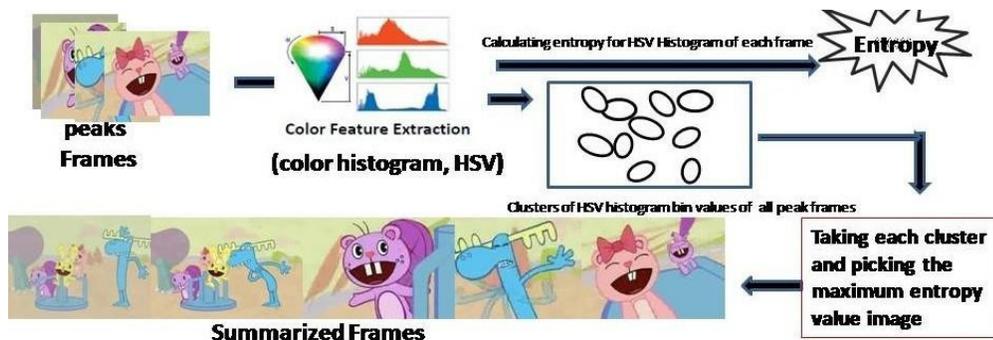


Fig. 3: Global Frame Extraction flow

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

Author Name	Proposed Method	Paper Result	Authors Future Work	Reference
Wen-Sheng Chu ,Yale Song ,Alejandro Jaimes	Maximal Biclique Finding (MBF) algorithm	Parallelized unsupervised video summarization with co-occurrence	The method may improve the using active learning or weakly-supervised learning for bipartite graphs	[15]
Supriya Kamal, Mankar, Aditya Naik, Mankar, Abhishek	The block matching techniques uses 2 approaches : 1 Diamond Search 2. Three step search	Diamond Search		[8]
S.de Almeida ,Antonio Carlos de Nazare Junior,Arnaldo de Albuquerque Araujo, David Menotti	VSUMM approach for parallelization	The hybrid version gives best summary for all resolutions and video lengths[14]		[1]
Karim M. Mahmoud,, Nagia M. Ghanem, Mohamed A. Ismail	Modified dynamic modeling-based agglomerative hierarchical clustering algorithm by time	Based on ground truth higher quality video summaries	Enhance the video skim by other features extractions in global method	[1]
Genliang Guan1, Zhiyong Wang1, Kaimin Yu1, Shaohui Mei2, Mingyi He2, and Dagan Feng	video summarization framework by local and global methods and greedy algorithm is implemented for keypoint pool for each scene and frames	keypoint based keyframe selection technique	Quantitative evaluation	[1]
Mr. Sandip, T. Dhagdi, Dr. P.R. Deshmukh	Block based 2 Histogram algorithm for shot boundary detection and keyframe extraction for image segmentation	Edge detection algorithm improved time to summarize the video will also increase the efficiency	GraphPartition Model SupportVectorMachine	[1, 2]

TABLE I: Related Work

Dataset	videos
Cartoons	10 videos
TV Show	5 videos
Home	1 video
News	15 videos
Commerci	2 videos

TABLE II: Dataset videos

summary with the other frames in the video sequence. We consider a video, summarized frames in that video and calculate manhattan distance between them. The distance between the set of key frames KF_t and a frame F belonging to the video S_t can be computed as:

$$d(F(t+n), KF_t) = \min_j \{Diff(F(t+n), F_{KF}(t+n_j))\} \quad j = 1, 2, \dots, \gamma_{NKF}$$

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

V. RESULTS

We have selected the summarized frames from the cluster- ing by calculating entropy from the histogram of a frame in the cluster. We have calculated entropy in RGB HSV and Lab color spaces and selected summarized frames.

V i d e o	RG B	HS V	Lab
V 1 1	4 64 115 195 270 315 352 489 522 663 814 831 932 109 5 110 4	10 15 164 208 270 315 337 484 495 721 754 871 945 109 9 110 7	4 60 103 198 323 385 479 620 707 814 820 845 101 5 109 5 110 4
v 1 5	103 150 234 257 262 521 571 679 693 784 106 2 109 6 130 7 131 8	81 103 152 265 334 427 458 542 647 703 793 991 117 1 131 5	15 93 163 259 277 385 518 549 679 690 759 793 106 9 130 5 131 5
v 9 3	176 296 338 354 374 407 427 444 460 479 517 528 592 649 661	14 230 294 354 374 414 457 484 528 581 592 640 647 652 725	44 241 274 331 390 447 452 471 526 543 599 617 642 649 721
v 6 8	70 127 130 133 194 363 632 656 688 819 100 3 120 8 166 2 182 2	23 130 145 245 327 357 632 656 718 756 864 133 9 138 7 182 7	27 142 220 253 420 500 658 682 756 794 108 9 121 0 164 2 193 7

TABLE III: Summarized frame numbers of videos in dataset

VI. EVALUATION METRIC

Evaluation is done by calculating fidelity measure [3]. Fidelity measure , which compares each key frame in the

Where Diff() is a suitable frame difference measure. The distance between the video S_t and the set of key frames KF_t is finally defined as:

$$d(S_t, KF_t) = \max_n \{d(F(t+n), KF_t) | n = 0, 1, \dots, \gamma_{NF} - 1\}$$

We can then compute the Fidelity measure as:

$$Fidelity(S_t, KF_t) = d(S_t, KF_t)$$

		summarization	
		hsv summ lab summ rgb summ	
carto	1	1.48	1
on	.	9	.
v11	8		4
	2		8
	4		9
v12	1	1.96	1
	.	9	.
	9		7
	8		7
	2		2
v13	0	0.95	0
	.	1	.
	9		9
	6		5
	7		2
v14	0	0.33	0
	.	2	.
	3		3
	7		6
	1		8
v15	0	0.88	0
	.	3	.
	8		6
	8		1

4	4
5	9
4	1
	<hr/>
	v107 0.852 0.908 0.88

TABLE VII: Fidelity values of videos in the dataset

VII. CONCLUSION AND FUTURE WORK

Video summarization is done using entropy in different color spaces and evaluated using fidelity Measure and we observed that in RGB colorspace summarization gives a better

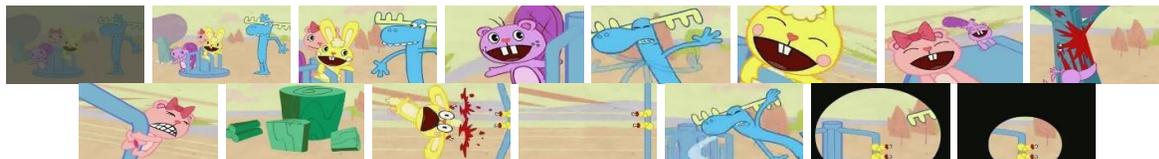


Fig. 4: RGB entropy summarized frames of V11 video



Fig. 5: HSV entropy summarized frames of V11 video



Fig. 6: Lab entropy summarized frames of V11 video



Fig. 7: RGB entropy summarized frames of V93 video



Fig. 8: HSV entropy summarized frames of V93 video

Fig. 9: Lab entropy summarized frames of V93 video



DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

								9	...	1160	
k1			0.007	0.002	1.984	1.984	1.984	1.988	...	2	
k2	1.999	1.999	1.981	1.981	1.582	1.581	1.582	0.557	0.559	...	1.999
k3	1.997	1.997	1.94	1.939	1.655	1.655	1.655	0.723	0.722	...	1.997
k4	1.992	1.992	1.897	1.897	1.699	1.699	1.698	0.871	0.869	...	1.992
k5			1.953	1.953	1.583	1.583	1.583	0.674	0.672	...	2
k6	1.99	1.99	1.972	1.972	1.746	1.746	1.747	0.814	0.815	...	1.99
k7	1.995	1.995	1.931	1.931	1.723	1.723	1.724	0.72	0.725	...	1.995
k8			1.949	1.949	1.531	1.531	1.531	0.706	0.707	...	2
k9	1.998	1.998	1.961	1.961	1.79	1.79	1.79	0.602	0.601	...	1.998
k10			1.855	1.856	1.489	1.488	1.489	0.859	0.854	...	2
k11	1.997	1.997	1.963	1.963	1.783	1.783	1.783	0.614	0.61	...	1.997
k12			1.996	1.996	1.863	1.863	1.863	0.611	0.607	...	2
k13			1.979	1.978	1.611	1.611	1.612	0.797	0.794	...	2
k14	1.501	1.501	1.978	1.978	1.676	1.676	1.676	0.995	0.994	...	1.501
k15	0.557	0.557	1.989	1.989	1.876	1.876	1.876	1.499	1.499	...	0.557
min	0.557	0.557	0.007	0.002	1.489	1.488	1.489	0.557	0.559	...	0.557
fidelity	1.489										

TABLE IV: Manhattan distance and fidelity of v11 video RGB summarization

								9	...	1160	
k1			1.988	1.988	1.988	1.587	1.587	1.588	0.015	...	2
k2			1.987	1.986	1.987	1.594	1.594	1.595	0.568	...	2
k3	1.998	1.998	1.933	1.933	1.933	1.665	1.665	1.666	0.726	...	1.998
k4	1.993	1.993	1.9	1.9	1.9	1.606	1.606	1.606	0.808	...	1.993
k5			1.953	1.953	1.953	1.583	1.583	1.583	0.674	...	2
k6	1.99	1.99	1.972	1.972	1.972	1.746	1.746	1.747	0.814	...	1.99
k7	1.996	1.996	1.934	1.934	1.934	1.709	1.709	1.709	0.676	...	1.996
k8			1.949	1.949	1.949	1.522	1.521	1.522	0.701	...	2
k9	1.998	1.998	1.961	1.96	1.96	1.789	1.788	1.788	0.591	...	1.998
k10	1.999	1.999	1.823	1.824	1.824	1.602	1.601	1.602	0.835	...	1.999
k11	1.998	1.998	1.991	1.991	1.991	1.715	1.715	1.715	0.486	...	1.998
k12			1.91	1.909	1.909	1.674	1.674	1.673	0.919	...	2
k13			1.954	1.954	1.954	1.686	1.686	1.687	0.787	...	2
k14	1.127	1.127	1.987	1.987	1.987	1.723	1.723	1.723	1.153	...	1.127
k15	0.342	0.342	1.992	1.992	1.992	1.93	1.93	1.93	1.69	...	0.342
min	0.342	0.342	1.823	1.824	1.824	1.522	1.521	1.522	0.015	...	0.342
fidelity	1.824										

TABLE V: Manhattan distance and fidelity of v11 video HSV summarization

	n	v	summarization		r
			hsv summ	lab summ	
e	8	.	0	0	g
w	8	5	.	4	b
s		1	9	9	s
				9	u
					m
					m
					0
					.
					4
					4
					9
	v	0		0.723	0
	8	.			.
	9	5			6
		6			6
		6			3
	v	0		0.24	0
	9	.			.
	0	1			1
		6			4
		3			2
	v	1		1.4	1
	9	.			.
	1	5			2
		5			4
		7			6
	v	0		0.382	0
	9	.			.
	2	4			3
		4			8
		7			3
	v	0		0.475	0

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

fidelity, compared with Lab and HSV colorspace summarizations. Our future work, we focus mainly to improve summarization by using different feature extraction methods and by changing clustering techniques.

v60 0.655 0.677 0.469

TABLE IX: Fidelity values of videos in the dataset

hsv	summarization lab	rgb
Home 0.709	0.813	0.733
Tvshows 0.903	0.918	0.922
News 0.756	0.761	0.774
Story board 1 1.043	1.05	0.995
		Story board 21.066 1.047 0.997

TABLE X: Mean Fidelity values of videos in the dataset

k1		0.007	0.002	1.984	1.984	1.984	1.988	1.988	...	1160
k2		1.978	1.978	1.593	1.593	1.593	0.555	0.557	...	2
k3	1.999	1.999	1.929	1.928	1.638	1.638	1.638	0.726	0.718	1.999
k4	1.993	1.993	1.9	1.9	1.605	1.605	1.605	0.804	0.802	1.993
k5	1.989	1.989	1.976	1.976	1.735	1.735	1.735	0.828	0.83	1.989
k6	1.995	1.995	1.932	1.932	1.704	1.703	1.704	0.726	0.728	1.995
k7			1.96	1.96	1.489	1.489	1.489	0.71	0.709	2
k8	1.997	1.997	1.963	1.963	1.738	1.737	1.737	0.593	0.594	1.997
k9	1.999	1.999	1.823	1.824	1.6	1.599	1.6	0.821	0.816	1.999
k10	1.997	1.997	1.963	1.963	1.783	1.783	1.783	0.614	0.61	1.997
k11			1.992	1.992	1.812	1.813	1.813	0.563	0.557	2
k12			1.895	1.894	1.667	1.668	1.667	0.978	0.977	2
k13	1.999	1.999	1.947	1.947	1.678	1.678	1.678	0.848	0.841	1.999
k14	1.501	1.501	1.978	1.978	1.676	1.676	1.676	0.995	0.994	1.501
k15	0.557	0.557	1.989	1.989	1.876	1.876	1.876	1.499	1.499	0.557
min	0.557	0.557	0.007	0.002	1.489	1.489	1.489	0.555	0.557	0.557
fidelity	1.489									

TABLE VI: Manhattan distance and fidelity of v11 video Lab summarization

REFERENCES

- [1] S. Almeida, A. Nazare, A. Araujo, G. Chavez, and D. Menotti. Speeding up a video summarization approach

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

- using GPUs and multicore CPUs. In *International Conference on Computational Science*. 2014.
- [2] E. Asadi and N. Charkari. Video summarization using fuzzy c-means clustering. In *Proc*, pages 690–694. May 2012.
- [3] Gianluigi Ciocca and Raimondo Schettini. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing*, 1:69–88, 2006.
- [4] Sandra E. F. de Avila. Antonio da Luz Jr. Arnaldo de A, 'VSUMM, 2008.
- [5] W. A. C. Fernando, C. N. Canagarajah, and D. R. Bull. A unified approach to scene change detection in uncompressed and compressed video. *IEEE Transactions on Consumer Electronics*, 46(3):769779, August 2000.
- [6] Marco Furini, Filippo Geraci, Manuela Montangero, and Marco Pellegrini. Stimo: Still and moving video storyboard for the web scenario. *Multimedia Tools Appl.*, 46(1):47–69, 2010.
- [7] G. Guan, Z. Wang, K. Yu, S. Mei, M. He, and D. Feng. Video summarization with global and local features. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops*, pages 570–575. IEEE Computer Society, Washington, DC, USA, 2012.
- [8] A. Kamoji, R. Mankame, A. Masekar, and A. Naik. Key frame extraction for video summarization using motion activity descriptors. *IJRET*, 62:291–294, 2014.
- [9] G. Karypis, E. Han, and V. Kumar. chameleon: Hierarchical clustering using dynamic modeling, *IEEE Computer*. 32(8):6875, August 1999.
- [10] Y. Li et al. *An overview of video abstraction techniques*. Tech. Rep, 2001.
- [11] K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail. Unsupervised video summarization via dynamic modeling based hierarchical clustering. In *Machine Learning and Applications (ICMLA)*, volume 2, page 303308. December 2013.
- [12] Mr. Sandip t. Dhagdi, Dr. P. R. Deshmukh, (2012). 'Key frame Based Video Summarization Using Automatic Threshold and Edge Matching Rate', *international Journal of Scientific and Research Publications*, Vol, 2, 2012.
- [13] N. Nagarajan and C. Kingsford. Uncovering genomic reassortments among influenza strains by enumerating maximal bicliques. In *Intl. Conf.* 2008.
- [14] J. Palacios and J. Triska. *A Comparison of Modern GPU and CPU Architectures: And the Common Convergence of Both*. March, 2011.
- [15] W. S. Chu. Y. Song, and A. Jaimes. Video co-summarization, 2015.
- [16] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans*, 3:3, 2007.

Intensifying the Accuracy of Finger-Vein Image Identification using Convolutional Neural Networks

Vineeth Reddy. K, M.tech (CSE-AI), JNTU Anantapur, India, nanivineeth2244@gmail.com Shoba Bindu.
C, Professor of CSE, JNTU Anantapur, India, shobabindhu@gmail.com

Abstract- Biometric identification is commonly used identification systems. This identity is classified based on the physical body parts and its behavior. Commonly used biometry's are a fingerprint, finger- veins, face. Describing the individual biometric identifiers is used as unique and measurable characteristics. The design of economical identity verification systems is to measure the distinctive Physical or behavioral symptoms of the particular person for their secure recognition. The identification of finger vein images with more accuracy is a challenging task in real-time scenarios. The CNN based finger vein image identification works can achieve greater accuracy in biometric identification. However this has produced better identification results but still, it is unable to defend spoofing attacks of finger- vein images. This paper mainly focuses on achieving high accuracy by using deep CNN model when considering multiple session's data over two publicly available datasets.

Keywords: Biometrics, identification, convolutional neural network, finger-vein.

I. INTRODUCTION

Personal identification is key to secure the data of individuals. Personal identification techniques can be extended to large scale applications such as Computer login, Bank account authorization, controlling access, e- commerce, etc. Biometrics identification getting more attention than traditional identification systems like maintain passwords, PINs, and security keys. In Biometric identification accessibility and protection of the system is considered as important. The biometric identification system requires fast response time and high accuracy. There are many methods of biometric identification based on physiological and behavioral characteristics. The most commonly used biometrics are fingerprint, face, iris, hand geometry. The above systems are not secure because these systems are disclosed outside the body of human beings. All these biometric identification systems are open for burlesque.

Nowadays making an effective model for biometrics identification by using the behavioral and physical characteristics for individual confidential detection is a tough job for industry associates and scientists. To overcome the problems for the above system, a new method of a biometric system using the veins present inside the finger of the human body. When compared to other systems this Biometric system is gaining more importance because it requires less cost and it will provide greater results in terms of achieving high accuracy for identification.

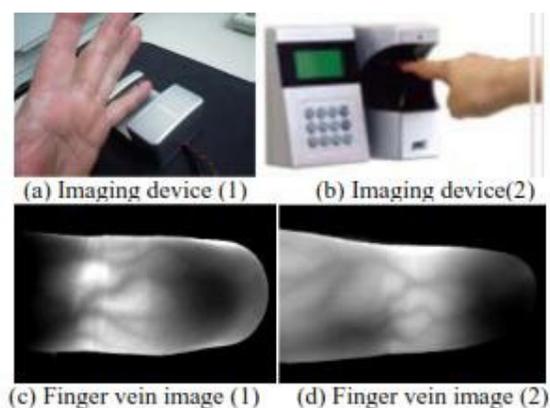


Fig 1: Prototype of Finger-vein images (a, b) and examples of infrared images of fingers (c, d)

This high accuracy rates of finger-vein biometrics are because of the vein patterns are Unable for spoof attacks and finger vein showed that patterns are different for every finger so that for individual identification it plays a crucial role. There are different methods used for the extraction of finger-veins. Despite recent advances in technology. To capture a finger-vein pattern a person needs to expose the finger to the near-infrared LED light and a monochrome CCD as shown in fig.1. The hemoglobin present inside the blood will observe the infrared rays, the pattern of veins inside the finger is taken as a shadow of patterns by these patterns the system will be made. The monochrome CCD camera as shown in fig.1 will record raw data of patterns and sent them to the database as a registered pattern image. This entire process of recording and storing the raw data of patterns will take two seconds. Repeated line tracking and mean curvature method has good performance but under some hypothesis, the grade of the finger vein images is low this is due to the optical blurring and scattering problems on the skin. During the feature extraction segmentation problems may occur and this is due to the image adaption, turn, and uneven brightness, scaling which may result in degrading of the performance of the system.

II. RELATED WORK

A Related number of modern finger-vein based biometric methods are tested on internal finger-vein datasets so that it is difficult to observe a relation with the newly proposed methods. Moreover, the significant performance may very often encounter when applying sophisticated procedures for different finger-vein databases, particularly when low-quality images are there in the database. Testing on different databases which are available publicly and compare the results obtained in the same datasets with the most sophisticated methods is therefore of great importance to evaluate the effectiveness of the proposed method.

In the following section a detailed description of the publicly available finger

vein datasets in Section II-A. Particularly identification of finger-vein and verification of finger-veins are quite different when coming to performance and these are explained in Section II-B and Section II-C.

A. Finger-Vein Datasets

There is a limited number of datasets that are available publicly for finger-vein based detection. The capability of the present Convolutional neural network system is evaluated over two publicly available datasets. The reason behind considering these databases is all the models or systems which are developed previously are considered on these finger-vein images. A brief introduction of the databases are described in the below Table: I and in the section below it will provide more details about them 1). FV-USM dataset

Bakhtiar A. R et. al[9]University Sains Malaysia collected finger-veins of the FV-USM database. This database consists of images around 123 people aged between 20 to 52 years old in which 83 are male and 40 are female. Each subject in the database asked to provide two fingers from left hand i.e., index and middle finger and two fingers from the right hand. A collection of 492 image classes are obtained from the above subjects. Geometry and vein pattern are the two important features from the captured finger images. Six times for each figure in every session and each subject participated in two sessions. In First Session 2952 images and from the second session total of 2952 images are captured, all these images have 640x480 and 256 grey level resolution.

2). SDUMLA database

Yilong Yin et. Al[10] Shandong University of China collected the SDULA database. In this database, each subject is asked to participate once and provide three fingers from each hand. Six images from one subject three from the left hand and three images from the right hand are collected and a sum of 3816 images were gathered and stored in the database. The collected images in the database are in 'bmp' format with 320x240 pixels in size.

TABLE 1: Details of publicly-available finger-vein image databases.

Database	Subjects	No. of Fingers	Details of Fingers	Images per Finger	Sessions	Image Size	Total Images
FV-USM	123	4	Left & right hand index & middle finger	12	2	640 × 480	5904
SDUMLA	106	6	Left & right hand index, middle & ring finger	6	1	320 × 240	3816

A. Finger-Vein Identification:

Van et al [10].used the MFRAT for the tendency of discrimination feature

extraction over the SDUMLA dataset. Grid PCA [2] has also been enforced to eliminate more redundant information. ETS comparison methods are used to solve the translations and by computing Euclidean distance can achieve a CIR =95.67% between test and training templates. Ong et al. proposed two credible steps based on the different cases for finger vein identification system on micro comparison [3]. They used SDUMLA dataset for their research, and mixed minutiae were used features derived from multiple cases of finger veins. For their proposed systems the authors choose the ROI and fuzzy contrast enhancement and CLAHE performed along with mean filtering and directional Dilation (DD).

B. Finger-Vein Verification:

Mainly the study of this paper focuses only on finger-vein biometric identification but some relevant information about the overview of the finger-vein biometric verification system also provided. Bakhtiar et al. [4] have improved finger-vein images of FVUSM dataset by using MGF (modified Gaussian filter) [5] and by correcting the image displacements. BLPOC [6] has been used for measuring the correspondence between registered images and test images as it is tough to noise, obstruction, and rescaling factors. In [7] authors used ASAVE and elastic matching, for SDUMLA databases for achieving an EER of 1.39%.

III PROPOSED SYSTEM

A convolutional neural network is a multilayer perceptron type of network with some special topology which has one or more than one hidden layers inside the network. CNN's are used to detect the objects present within the image while processing, recognition of handwritten characters, and recognition of speech as all of these will automatically extract the particular features within the input raw data without using any specific normalization techniques. This type of models is important while finding constant features when there is an inner structure like images that are present in the input data. CNN's won't consider the user hand design input features, which cannot be extracted by considering common problems. In present days, applications like CNN are introduced in the area of finger vein image identification and verification. The below sections will describe the different layers present in Convolutional neural network.

1. CNN description:

The below subsection will provide a detail description of the CNN model.

A. Convolutional Layer:

In the convolutional layer [1], a group of two-dimensional convolutional

operation is implemented on the inputs x_l here m and l are levels and map indexes and through kernels, the filters are denoted as w_l where n is considered as filter index. From the above inputs y_l (n th output of layer l) is calculated as :

$$Y_l = \sum M^{(l-1)} w_l * x_l + b_l$$

In the above equation $*$ represents convolutional, b_l can be considered as the n th output of level l and $M_l - 1$ is considered as no. of input maps.

B. ReLu Layer:

The ReLu is an activation layer (or non- linear layer) it is used after the convolutional layer in the network. The purpose of the ReLu layer is to enable non-linearity in the system. Previously, some non-linear functions like sigmoid and tanh are being used but later to train network faster researchers found ReLu layer and this layer will train the network without defecting the accuracy. It will be useful to reduce the invisible gradient problem, the problem of slow training of the lower layers of the network due to the decrease of gradient over the layers. The ReLu layer will convert the negative activations present in the layer to 0 by applying the function $f(y) = \max(0, y)$.

C. Pooling Layer:

The pooling layer is also called as the downsampling layer. The most popular used layer is Maxpooling. Generally, it will be 2×2 in size filters and the length of strides will be the same. It is applied to a volume of input which results in max value for every subregion that convolves filters involve. There are some other pooling options available those are Average pooling and L2 norm pooling. The obvious logic behind polling layer is that we know that a unique feature is in the actual input volume. That is, the higher the activation value, its current position is not as important to other relative places. This layer desperately decreases the spatial range of the input volume. Consequently, neither parameters nor weights are significantly reduced by this cost of computational and over-fitting will be controlled.

D. Fully Connected Layer:

Fully connected layer takes input as output of the previous convolution layer or pooling layer or ReLu layer and generates an N size vector, where N is the no. of classes that a program should select. The Softmax classifier is commonly used to estimate the likelihood of an image input related to a specific label. Let m th input map of the output layer be as x_m , then linear combination defined as:

$$O_n = \sum M (w_{n, m} * x_m + b_n).$$

Where $m = 1$ and * represents the convolutional layer.

The CNN architecture is shown in Fig 2. The network we have proposed has 5 convolutional layers, 4 max-pooling, 3 ReLu, 1 Global average pooling, and a softmax layer.

2. CNN Training and Testing:

The size of the CNN with hidden layers will depend on the image size i.e. if the size of the image is bigger than it will lead to bigger CNN model with more hidden layers. To reduce the complexity in the CNN model initially we consider the size of the image to 65×153 . Our approach for generating the training and testing templates for the CNN model is by considering the images from multiple session. For our experiment, we consider the four fingers of each individual as a separate class for FV-USM data, since 123 persons with four fingers each and there is a sum of 492 classes present for training. For SDUMLA dataset, since 106 persons with six fingers each and there is a sum of 636 classes available for training.

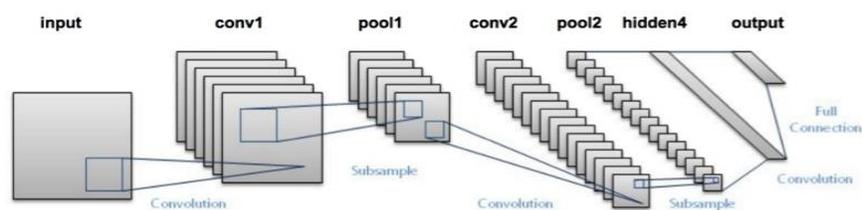


Fig 2: Architecture of CNN Model

For training the CNN, we used 80% of finger-veins from both databases and for testing the remaining 20% finger-vein from both datasets consider for testing. Our CNN model set to 0.00001 learning rate with a batch size of 32 samples for FV-USM and SDUMLA datasets, so that the loss can be reduced when the repetitive execution of epochs increased. When the number of epochs is large then it allows the network to be well-trained. For our experiment, we consider 40 epochs. The purpose of using the smallest learning rate and large epochs is the network will train slowly and every small detail will be considered from every class.

For each test image sample, the trained CNN model provides a probability value to the available classes or finger-veins. For our experiment, we set a threshold value to match the given test image sample. For the input mage sample, if the probability value return by our model is greater than 60% the image is classified as “identified” otherwise it will be classified as “not identified”.

III RESULTS

To evaluate the proposed network, the accuracies achieved by previous methods for finger-vein based biometric systems and compared the result is discussed in Table II, with our proposed CNN based approach when using the same training and testing strategies.

All the experiments for training and testing have been performed in Anaconda with a system configuration of 16GB RAM; NVIDIA GeForce GT 710 2 GB DDR3 graphics card; i7, 3.40GHz processor and Windows 10 operating system.

TABLE II: CNN-based identification accuracy over the considered publicly-available databases.

Database	Training	Testing	Das et al. CNN with CLAHE enhanced images	Proposed CNN with Multiple Sessions (raw data)
FV-USM	3 images from each session	3 images from each session	97.05%	-
FV-USM	8 images from each session	4 images from each session	-	98.06%
SDUMLA	4 images	2 images	95.13%	-
SDUMLA	4 images	2 images	-	98.78%

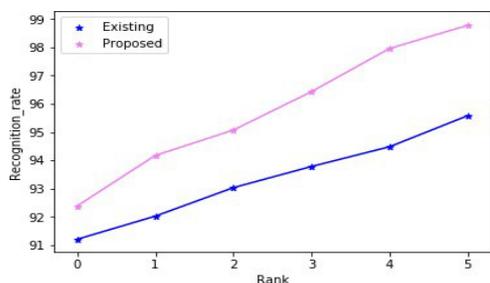


Fig 3: SDUMLA dataset Comparison of Recognition rate between Existing and proposed CNN models

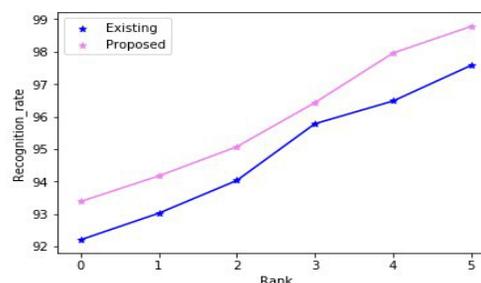


Fig 4: FV-USM dataset Comparison of Recognition rate between Existing and proposed CNN models

IV CONCLUSION

In this paper, the proposed model for CNN which can perform an effective identification of finger-veins irrespective of the conditions. We conducted a set of experiments over two publicly available datasets. The results showed that it is possible for achieving the accuracy of identification will be higher than 97% for considered datasets available publicly using the proposed Network CNN model. The proposed model showed that if the images for training increases the accuracy of the identification also increases when considering the different quality images and with multiple session. In the future consider more than two datasets and increase the quality and quantity of the images may give better results.

ACKNOWLEDGMENT

We would like to express our thanks to the MLA Lab of Shandong University for SDUMLA-HMT Database and FV-USM Database.

REFERENCES

- [1] Rig Das, Emanuela Piciuccio, Emanuele Maiorana, and Patrizio Campisi, “Convolutional Neural Network for Finger- Vein-Based Biometric Identification” *IEEE Trans. on Information Forensics and Security* (Volume: 14, Issue: 2, Feb. 2019)
- [2] H. T. Van, P. Q. Tat, and T. H. Le, “Palmprint verification using gridpca for Gabor features,” in *Proceedings of the Second Symposium on Information and Communication Technology*, ser. SoICT '11. New York, NY, USA: ACM, 2011, pp. 217– 225.
- [3] T. S. Ong, J. H. Teng, K. S. Muthu, and A. B. J. Teoh, “Multi-instance finger vein recognition using minutiae matching,” in *2013 6th International Congress on Image and Signal Processing (CISP)*, vol. 03, Dec 2013, pp. 1730– 1735.
- [4] M. S. M. Asaari, S. A. Suandi, and B. A. Rosdi, “Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics,” *Expert Systems with Applications*, vol. 41, no. 7, pp. 3367 – 3382, 2014.
- [5] E. C. Lee, H. Jung, and D. Kim, “New finger biometric method using near-infrared imaging,” *Sensors*, vol. 11, no. 3, pp. 2319– 2333, 2011.
- [6] K. Takita, T. Aoki, Y. Sasaki, T. Higuchi, and K. Kobayashi, “High-accuracy subpixel image registration based on phase-only correlation,” *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences*, vol. E86-A, no. 8, pp. 1925–1934, 2003.
- [7] L. Yang, G. Yang, Y. Yin, and X. Xi, “Finger vein recognition with anatomy structure analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, March 2017.
- [8] Mohd Shahrime Mohd Asaari, Shahrel A. Suandi, Bakhtiar Affendi Rosdi, Fusion of Band-Limited Phase-Only Correlation and Width Centroid Contour Distance for finger- based biometrics, *Expert Systems with Applications*, Volume 41, Issue 7, 1 June 2014, Pages 3367-3382, ISSN 0957-

4174, <http://dx.doi.org/10.1016/j.eswa.2013.11.033>.

[9] Yilong Yin, Lili Liu, and Xiwei Sun. SDUMLA-HMT: A
Multimodal

[10] H. T. Van, T. T. Thai, and T. H. Le, “Robust finger vein identification base
on discriminant orientation feature,” in 2015 7th Int. Conf. on Knowledge and
Systems Engineering (KSE), Oct 2015, pp. 348–353.

Image similarity with cosine distance

C. Siva Jyothi

CSE Dept, MVSR Engg College,
Hyderabad, India
sivajyothi.chandra@gmail.com

B. Sandhya

CSE Dept, MVSR Engg College,
Hyderabad, India

Abstract— Image similarity, is the process of identifying correspondences between same scene images that differ due to different acquisition parameters(illumination, view point, multi modal) or noise(ex: blur). Image patch matching gives two patches as input to a deep neural network, find similarity between them. Our objective is to design a convolution neural network that classifies image patches by finding similarities between images of same scene. Similarities of images are measured from the feature maps that are extracted from raw patches. A model is developed that maps the patch to low-dimensional feature vector and distance is calculated using cosine distance. Threshold is applied on the similarity distance resulting '1' for similar patches and '0' for dis-similar patches. The results are collected by training the model with Hpatches dataset and evaluating the model with Hpatches dataset. Promising results of Mean Absolute Error has been shown.

Keywords- Image matching; similarities; feature maps; classification; convolution neural network; cosine distance;

Introduction

Identifying the correspondences between different images of the same scene is one of the key operations in several computer vision applications such as object detection and recognition, capturing of image from motion, image fusion etc. Computing similarity between patches across images is quite challenging due to photometric and geometric variations between them such as illumination of a scene, viewpoint, shading, occlusions etc. Similarity between patches is computed by measuring the similarity between the feature descriptors of the patches.

Conventionally image matching has been addressed using hand crafted feature detectors and descriptors such as SIFT[3], SURF[4], MSER [5]etc. Using detectors, key points are detected and a patches are generated from images. Extracted patches are represented as a vector using descriptors such as SIFT, HOG which commonly use gradient information. The descriptor vectors are matched using similarity measures such as Euclidian, cosine. Recently due to success of deep nets across applications of computer vision, deep CNN's are being used for learning features instead of traditional hand crafter features. In this paper, a CNN is designed, trained and tested for finding similarity between patches of H Patches (Homography Patches) images.

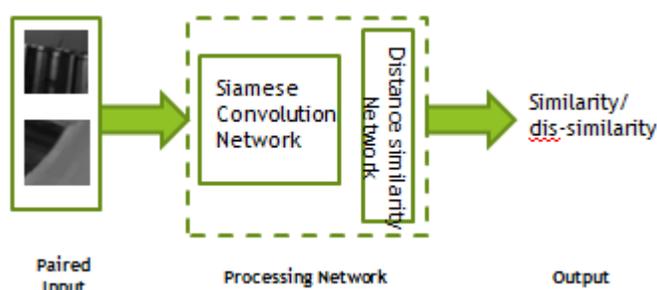


Fig1. Block diagram of paired input Image similarity

The paper is organized as follows: Section 2 discusses related work, focusing on descriptors, dataset, metric learning and performance key points. Section 3 details the different network architectures used for patch matching. Section 4 details with training of network, datasets utilized, similarity function used, explains how the pre-processing, learning, pairing of datasets, evaluation pipeline are performed. Section 5 showcases the experimental methodology and results on a standard dataset. Follows the conclusion and extension of future work

RELATED WORK

Deep nets have been proved to be effective in the field of computer vision and have become a state of art for image recognition and detection purposes.

AlexNet[1] is a deep convolution neural network which is proved to be successful in image classification. It is trained with Imagenet dataset of over 15 million high resolution images belonging to roughly 22,000 categories. Alexnet won the Imagenet Large-Scale Visual Recognition Challenge (ILSVRC) with roughly 1000 images in each 1000 categories. It contains 8 (5 convolution ,3 fully connected) layers. ReLU activation function is applied after every convolution and fully connected layer. The network takes 224X224X3 as the input image size. Data augmentation is one of the technique used in reducing overfitting.

VGG16 [2] (Visual Geometry Group) is a CNN which consists of 13 convolution and 3 fully connected layers and is trained with weights of imagenet. Network is trained with varying image scales, such that it can recognize objects over a wide range of scales.

Matchnet [6] uses Deepnet architecture and learns jointly the feature network and metric network. Feature network maps a patch to a feature representation and a metric network computes similarity between the pairs of features. The network is trained on UBC patch dataset by minimizing the cross-entropy error with plain stochastic gradient descent(SGD) with a learning rate of 0.01. Performance is evaluated at different quantization levels of bottleneck layer, resulting in reducing the average error rate and increasing accuracy.

Srikar Appalaraju and VineetChaoji [7] approach uses deep siamese network trained with CIFAR10 datasets for evaluation of accuracy of Image similarity. Network is trained with curriculum learning by selecting easier samples first and then selecting “hard image pairs”. Sampling is done based on selection of pair of images for training with online pair mining strategy(OPMS) to reduce the convergence and performance. Contrastive loss function with L2 norm is used in training. Multi-scale siameseconv network resulted better at finding fine-grained image similarities.

Zagoruyko and Nikos Komodakis [8] proposes architecture Spatial pyramid pooling(SPP) network for comparing patches of arbitrary sizes by adjusting the size of spatial pooling regions same as input patch. Hinge-based loss term and squared l2-norm regularization is used in learning. Asynchronous Stochastic

Gradient Descent (ASGD)[9] with constant learning rate, momentum of 0.9 and weight decay is used to train the model. Standard benchmark datasets Yosemite, Notre Dame and Liberty extracted from DoG descriptor are used in training. Results show that the estimated depth maps exhibit more fine details than DAISY[10]. MSER [5] detector is applied to both pair of images to extract keypoints and then patches(from ellipses provided by detector).

Melekhov and Kannala [11] approach proposes a siamese network that accepts two patches of images as input. L2 distance measure with contrastive loss, optimizer SGD with learning rate 0.01, weight decay 0.001 with a mini batch size 100 are used between image patches. Pre-processing technique used is mean-standard deviation over the whole training data set to normalize intensity of every pixel of gray scale input image. Training is performed on Liberty patches of MSC dataset, testing on Notredame dataset. Histogram equalization for adjusting patch contrast along with Spatial Transform(ST) layers improved accuracy and potential future performance that makes the descriptor more robust to geometric transformation.

Akilapemasiri, Sridhasridharam [12] proposed an approach of Patch Matching with sparse representation which results in better accuracy to capture structures and patterns in images. Evaluation is based on Liberty and Notredame subsets of UBC patch dataset. Training the network in supervised manner by minimizing the binary cross entropy using Adam optimizer. Features from sparse coding are trained to CNN on which raw image patches are classified as 0 and 1.

ARCHITECTURES

CNN Architecture Employed

Three different architectures are tested for computing the similarity between the patches of images:

Setup 1: Pre trained VGG16 model used as feature extractor

Setup 2: Siamese CNN which outputs distance between the patches simulating the behavior of distance computation between the feature descriptors

A VGG16 as feature extractor

A pre-trained VGG (Visual Geometry Group) that has proved its challenge in 2014 came out with combination of convolution, max pooling layers along with three fully connected layers (FC1, FC2, FC3) with dropout layers totaling to 16 layers and has 1000 different classes of predicting the classifications they belong to. It has been proved in object classification, Image classification, Image reconstruction. Uses ReLU(Rectified Linear Units) and Sigmoid as activation functions for Dense Layers that are fully connected.

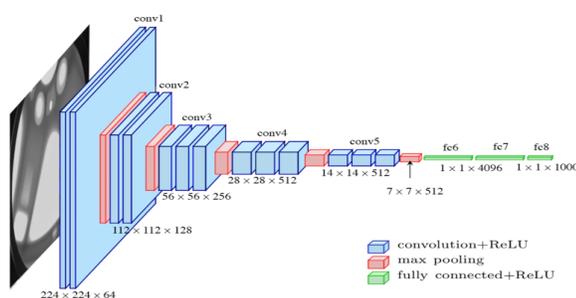


Fig 2. VGG16 architecture – pre-trained model

Before the image is fed to the network, it is normalized using the imagenet mean, standard deviation of the total images on which VGG16 layers are trained.

By extracting the local image features at fully connected layers (fc1, fc2, fc3) separately for two input images, similarity is computed using metrics such as Euclidean distance(L2 norm) and cosine-similarity. With Euclidean distance the output should be less than a threshold 'm' if the patches are similar(label '1') and distance is more for dis-similar patches(label '0'). With cosine-similarity, value is high for similar images('1') and less than threshold for dis-similar images('0').

Accuracy and Mean absolute error are measured with the evaluation of the network by applying a threshold.

B. Siamese CNN with similarity values as output

A Siamese CNN which takes two image patches as inputs and outputs a similarity value, similar to cosine similarity is designed. Three channel images are used to train the network, to improve performance even in the case of color images. The network has two branches that share exactly same architecture and the same set of weights.

Each branch takes one image patch as the input and passes through a series of convolution and max-pooling layers. The architecture of each network in the Siamese CNN is based on VGG16 model. However input size to a VGG16 model is $224 \times 224 \times 3$. Since we are designing a network for patch based matching, we modified the input size to $64 \times 64 \times 3$.

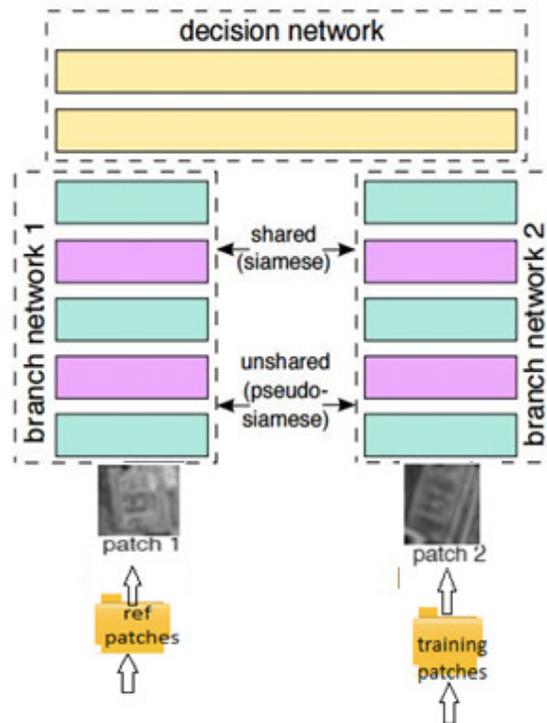


Fig 3. Siamese CNN with two branches

Hence only the convolution layers of VGG16 could be adopted without the fully connected layers. We have added 2 fully connected(Dense) layers after convolution output. The description of each layer is described in the table 1.

Two branches generate the local features map individually from each patch. At the final layer, we are experimenting by inserting a Lambda layer that takes two input vectors of same size and computes Euclidean distance /cosine-similarity based on choice. The images here are pre-processed by normalizing the intensity value to 0 to1.

Layer	Input Size	No of parameters
Input	64 X 64 X 3	0
Conv0	64 X 64 X 64	147584
Conv1	32 X 32 X 128	147584
Conv2	32 X 32 X 128	147584
MaxPool2	16 X 16 X 128	0
Block3_conv1	16 X 16 X 256	295168
Block3_conv2	16 X 16 X 256	590080
MaxPool3	8 X 8 X 256	0
Block4_conv1	8 X 8 X 512	1180160
Block4_conv2	8 X 8 X 512	2359888
MaxPool4	4 X 4 X 512	0
Block5_conv1	4 X 4 X 512	2359888
Block5_conv2	4 X 4 X 512	2359888
MaxPool5	2 X 2 X 512	0
Flatten	2048	0
Dense	700	1434300
Dropout	700	0
Dense	300	210300
		16,359,288

Table 1. Layers of Siamese CNN.

Flow Diagram

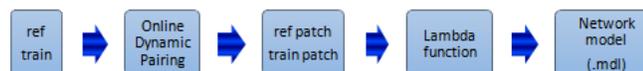


Fig 4. Training network model(.mdl) with paired inputs

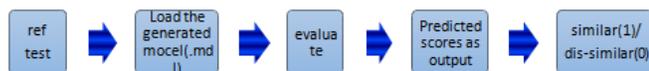


Fig 5. Evaluating the model for score predictions

4 TRAINING

4.1 HPatches Dataset

HPatches – Homography patches dataset contains patches extracted from image sequences collected from different sources of images.

Each image sequence contains images of same scenes differ by transformations, illuminations along with certain amount of noise. Illumination sequences are more challenging than geometric ones.

51 sequences are captured by camera, 33 scenes are captured from outdoor, 12 scenes from Interesting interest points, and others from fully affine invariant comparisons.

A total of 116 image sequences with 57 scenes have photometric changes, 59 scenes have significant geometric deformation.

Each sequence includes reference image and 5 target images with photometric and geometric variations. For each image patches are detected in the reference image and are projected on the target image using the ground truth homographies.



Figure 6. shows the image sequences with diversity in illumination, focus, reflections, other noises

Patches are sampled in the reference image with a combination of local feature extractors that proved successful results like Hessain, Harris[13], DoG detectors. Patches are extracted from the regions with magnified scale factor of 5 and preventing multiple detection overlap of regions. Now the set of corresponding image patches of sequence images are done with a certain amount of noise.

In patch extraction an affine jitter is applied on the target image to simulate the geometric repeatability error. With easy jitter the overlap with original image is 0.85, with hard jitter the overlap is 0.72.

Noise is simulated in image sequences with the settings of Easy(e), Hard(h), Tough(t). Images in the sequences are sorted by increasing transformations, resulting increase in noise and are segregated as e,h,t.

Every Image sequence is instantiated with Easy, Hard, Tough patches.

Set of reference patches ref.png, remaining 5 images in the sequence three patch sets eK.png, hK.png, tK.png containing the corresponding patches from ref.png as found in the k-th image with increasing amounts of geometric noise($e < h < t$). Affine adaptation is not included, so all patches are of square regions.

Ex: ref45.png – e145.png; e245.png; e345.png; e445.png; e545.png;



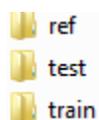
Figure 7 showing extracted easy patches of HPatches dataset



Figure 8 showing extracted hard patches of HPatches dataset

4.2 HPatches training – testing

Patches are extracted and segregated as ref, train, test folders.



ref - ref.png;

test - e2X,e4X, h2X,h4X, t2X,t4X;

train - e1X,e3X,e5X, h1X,h3X,h5X, t1X,t3X,t5X

For every patch in ref folder we randomly pick a patch in the corresponding train folder with 3 sets of images prefixes with 'e1','e3','e5'(easy patch) with same name of ref, for positive patch matching so that these 3 pairs are labeled with 1. For the same ref patch we select a patch with 'h1','h3','h5' prefixes with some other patch name for negative matching, and labeled as 0. Randomly selecting the patches for training is termed as online pair matching

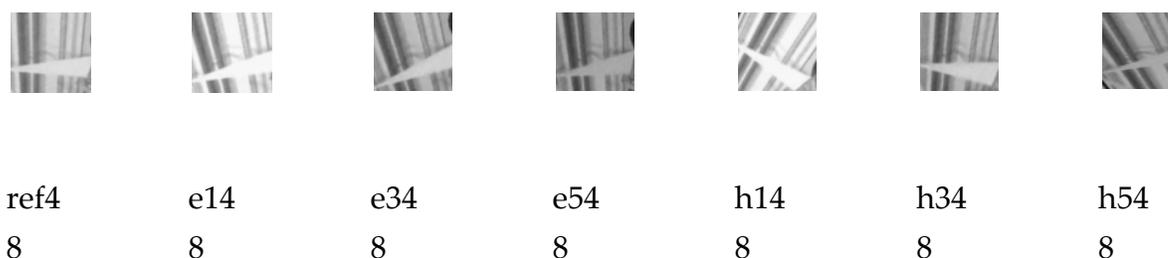


Fig 9: Positive matching pairs

(ref48.png,e148.png),(ref48.png,e348.png),(ref48.png,e548.png) -1



Fig 10: Negative matching pairs

(ref48.png,e198.png),(ref48.png,e3198.png),(ref48.png,e5298.png)-0

4.3 Data Augmentation Pre-processing

Data Pre-processing is required in order to deal with noise, standardizing, normalizing the data that is been used at the time of training the network.

Before feeding the Input data to the Convolution Neural Network it has to be processed by re-sizing, re-scaling, applying rotation. This is pre-processing of data. Here normalization of intensity values of each pixel in the input gray scale patch by calculating the mean, standard deviation over the whole training data, the image data by converting the scalar to vector that ranges from [0, 1] or [-1 ,1].

Along with weights are required to convolute the layers from the Input layer. The weights are used are of Imagenet[1] weights that are used in convolution operations in the Neural network.

Data Augmenting technique to avoid overfitting by apply rotation, scaling vertically and horizontally.

Histogram equalization technique that adjust the intensity of patches allows to improve the contrast of images is observed to improve performance.

4.4 Online Dynamic Pairing

In Siamese Network two patches [ref + train] are combined as single input and selected dynamically by the random seed. Suppose random number 85 is picked it selects ref85.png as reference image, then its positive training pairs are e185.png, e385.png, e585.png for easy patches. Similarly h185,h385,h585 for hard positive pairs, t185, t385, t585 for tough positive pairs.

For negative pairing ref85.png with e1796, e3954, e541png files are picked on the fly dynamically. Where it always changes its selection and make sure that it oesn't

take a positive pair as negative pair.

4.5 LOSS FUNCTION

The proposed network of Siamese CNN with predicted score as output is trained on **more than 85,000** samples, validate on **15,420** samples for 225 epochs. Initially the loss was **0.21224**, gradually the **loss** reduced to **0.16642**.

The network is trained by using contrastive loss function

$$\frac{1}{2}\{(Y)(D_w^2) + (1 - Y)\{\max(0, m - D_w)\}\}$$

Where Y is a binary label selected the input pair is positive (Y=1) or negative(Y=0),

$m > 0$ is the margin for negative pairs.

D_w is defined as the distance(Euclidean/cosine) between the outputs of the siamese networks

Contrastive loss function is used to learn discriminative features of images , during training the model with ground truth of images.

Model was trained using NVIDIA Tesla K80 GPU Frame work by using Tensorflow as back end.

4.6 Optimizer

In image matching we try to minimize the 'mean_absolute_error' in metrics. Mean absolute error is the average of errors in a set of predictions. It is the average of the absolute differences between the prediction and actual observation where all individual differences have equal weight.

$$mae = \frac{1}{n} \sum_{i=1}^n \llbracket abs(y_i - \lambda(x_i)) \rrbracket$$

n - number of image patches, y_j - actual value,

$\lambda(x_i)$ - predicted value

4.7 Cosine similarity

Similarity measure utilized in calculating whether two patches are similar / dis-similar is cosine similarity.

$$\text{Cosine similarity} = \cos \theta = \frac{A \cdot B}{|A| |B|}$$

Cosine similarity between two vectors is the measure of the cosine of the angle between the vectors. Computing the cosine similarity between 4096 values of each input set [P1,P2]. If the similarity value is nearing to 1 means they are similar and the angle is 0° . where as the measure is deviating from 1 then the magnitude increases and they are dis-similar. By performing dot product of vectors and dividing with product of magnitude of each vector separately.

5. Experiments

In patch based matching we have sets of patches each extracted from two images. To compute $N1 \times N2$ pair-wise matching scores with $N1$ and $N2$ as the number of patches from each image. Pushing each pair of image through the network generates the convolutions and other operation with the layers and executes the cosine-similarity over the pairs and predicts the values.

In order to speed up the scenario the features of First image patch is concatenated to **FirstR** array $n2$ times and all the Second image patches are concatenated to Second array(size $n2$). [**FirstR** , **Second**] is fed to the network for retrieval of scores. Scores are passed through threshold of m and selected if the score is greater than threshold and labeled as 1 for matching pair , else labeled as 0 for non-matching pair.

The above scenario is repeated $n1$ times.

Conducting results on VGG16 pre-trained network and custom created Siamese networks proposed in Section 3. Promising results are seen by reduction in mean absolute error and absolute error.

6. Results

Training Results

Below are the results drawn by training of the network architecture specified in Section III B with HPatches data set.

Architecture has been trained with 225 epochs by updating the weights of each epoch with more than 85,000 patches with supervised learning by labeling a true (positive) pair as '1' and false (negative) pair as '0'.

Training accuracy increased from **75.9%** to **98.56%**. **Validation accuracy** increased from **72.80** to **82.80**.

Testing Results

Below table gives the accuracy of trained and pre-trained model with the HPatches Dataset.

Data set	Pre-trained accuracy(%)	Trained model accuracy(%)	samples
i_nuts	50	52.39	4704
i_leuven	50.46	52.39	2384
i_greenhouse	50.09	53.47	3520
i_brooklyn	50.1	54.19	11988
i_steps	51.04	56.44	3260
i_fenis	50.18	56.49	2188
i_books	50.11	56.91	1896
i_resort	50.05	58.14	10092
i_greentea	50.05	58.28	4244
i_melon	50.05	60.72	2220
			46496

Table 2 showing the test results of accuracy of HPatches with pre-trained and trained model

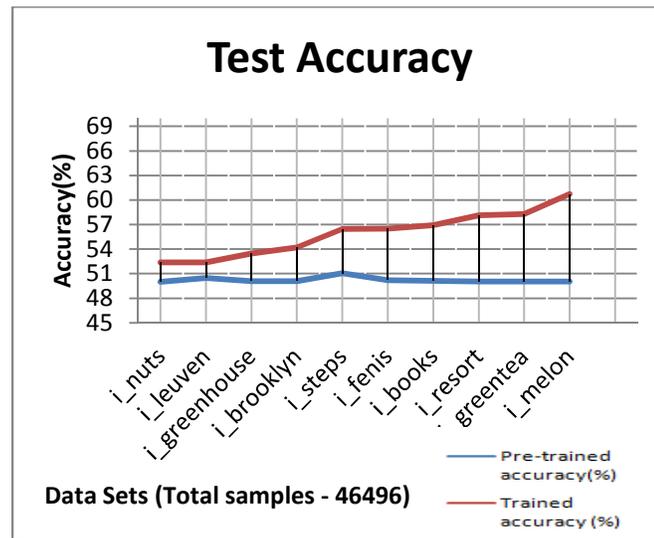


Fig 11. Graph showing the increase in accuracy from pre-trained to trained model during evaluation with HPatches Dataset

CONCLUSION

In this paper we design a deep neural network that compares raw image patches of same scene images and learn general similarity function for patches. A study has been performed on different network architectures resulting best performance on Siamese CNN with predicted similarity score as output. In training process we utilize matching of similar and dis-similar pairs of HPatches dataset with photometric changes on homography resulted best. Also the metric terms used are concentrated on reducing the mean absolute error. By increasing more training set, potential increase in overall performance.

References

- [1]. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: In ImageNet classification with deep convolutional neural networks: In NIPS, 2012, 1,2
- [2]. Karen Simonyan & Andrew Zisserman + Very Deep Convolutional Networks For Large-Scale Image Recognition
- [3]. Jindal et al., International Journal of Advance research, Ideas and Innovations in Technology. (Volume 1, Issue 1, October 2014)
- [4]. Herbert Bay¹, Tinne Tuytelaars², and Luc Van Gool^{1,2}: SURF - Speeded Up Robust Features: In

- [5]. J. Matas^{1,2}, O. Chum¹, M. Urban¹, T. Pajdla¹: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions: In
- [6]. Xufeng Han ; Thomas Leung ; Yangqing Jia ; Rahul Sukthankar ; Alexander: MatchNet: Unifying feature and metric learning for patch-based matching. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- [7]. Srikar Appalaraju, Vineet Chaoji, "Image similarity using Deep CNN and Curriculum Learning," In *arXiv:1709.08761*, 2017
- [8]. Sergey Zagoruyko, Nikos Komodakis: Learning to Compare Image Patches via Convolutional Neural Networks: In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- [9]. Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, Tie-Yan Liu : Asynchronous Stochastic Gradient Descent with Delay Compensation: In *Proceedings of the 34th International Conference on Machine Learning, PMLR 70:4120-4129*, 2017.
- [10]. Engin Tola, Vincent Lepetit, and Pascal Fua, Senior Member, IEEE DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo : In *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32, No.5, May 2010*.
- [11]. Iaroslav Melekhov, Juho Kannala, Esa Rahtu: Image Patch Matching Using Convolutional Descriptors with Euclidean Distance : In *ACCV 2016 Workshops proceedings (Workshop on Interpretation and Visualization of Deep Neural Nets)*
- [12]. Akila Pemasiri, Kien Nguyen, Sridha Sridharan, and Clinton Fookes : Sparse Over-complete Patch Matching: In *arXiv:1806.03556v2 [cs.CV]*
- [13]. Max Danielsson¹ , Thomas Sievert¹ , Håkan Grahn¹ and Jim Rasmusson: Feature Detection and Description using a Harris-Hessian/FREAK Combination on an Embedded GPU: In *ICPRAM 2016 - International Conference on Pattern Recognition Applications and Methods*.

“Internet of EVERY Thing” Applications, its Future and Challenges

Rayees Fathima
rfatimabb@gmail.com

Abstract: The Internet of Everything was listed as one of the top trends of 2015 by GRATNER The term Internet of Everything (IoE) is a fairly new term, and there is a confusion about the difference between the Internet of Everything (IoE) and the Internet of Things (IoT) , to clarify that, let’s start with definitions , applications and e

Keywords: Internet of Things; RFID,NFCHeterogeneous, Organizational Challenges, Security,Technology.xplore the future of this new concept.

1.INTRODUCTION

WHAT IS THE INTERNET OF EVERYTHING (IOE)?

The Internet of Everything (IoE) “*is bringing together **people, process, data, and things** to make networked connections more relevant and valuable than ever before-turning information into actions that create new capabilities, richer experiences, and unprecedented economic opportunity for businesses, individuals, and countries.*”, (Cisco, 2013).In simple terms: IoE is **the intelligent connection of people, process, data and things**. The Internet of Everything (IoE) describes a world where billions of objects have sensors to detect measure and assess their status; all connected over public or private networks using standard and proprietary protocols.

II.LITERATURE REVIEW

Pillars of the Internet of Everything (IoE)

- **People:** Connecting people in more relevant, valuable ways.
- **Data:** Converting data into intelligence to make better decisions.
- **Process:** Delivering the right information to the right person (or machine) at the right time.

- **Things:** Physical devices and objects connected to the Internet and each other for intelligent decision making; often called *Internet of Things (IoT)*.

The Internet of Things (IoT)

The Internet of Things (IoT) is the network of physical objects accessed through the Internet. These objects contain embedded technology to interact with internal states or the external environment. In other words, when objects can sense and communicate, it changes how and where decisions are made, and who makes them. For example Nest thermostats.



III.APPLICATIONS

The difference between IoE and IoT The Internet of Everything (IoE) with four pillars: people, process, data, and things build on top of The Internet of Things (IoT) with one pillar: things. In addition, IoE further advances the power of the Internet to improve business and industry outcomes, and ultimately make people’s lives better by adding to the progress of IoT. (*Dave Evans, Chief Futurist Cisco Consulting Services*)



IV FUTURE OF IoE

THE FUTURE? The Internet of Everything will re-invent industries at **three levels: business process, business model, and business moment.** “**At the first level, digital technology is improving our products, services and processes, our customer and constituent experiences, and the way we work in our organizations and within our partnerships,**” said Hung Le Hong, research vice president and Gartner Fellow.

IV CHALLENGES

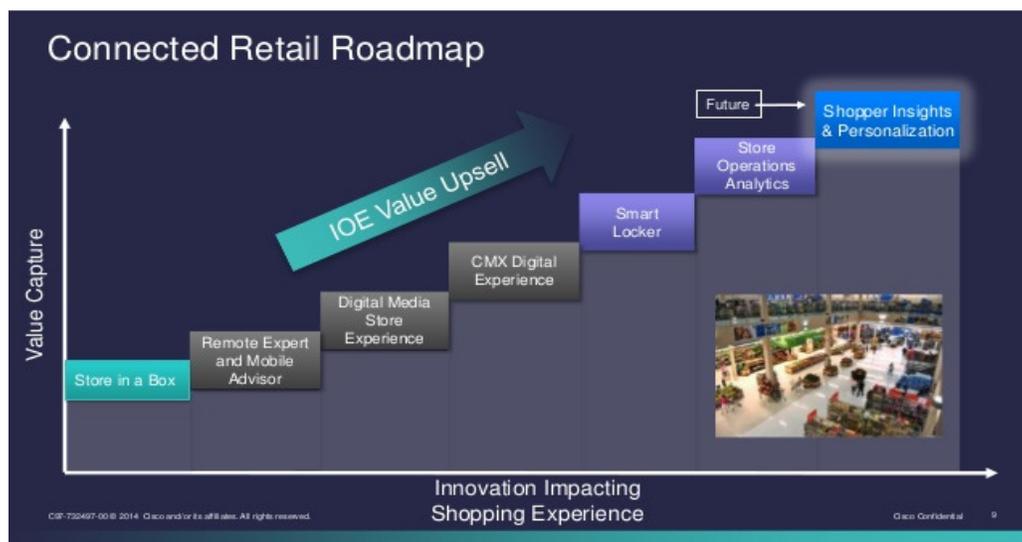
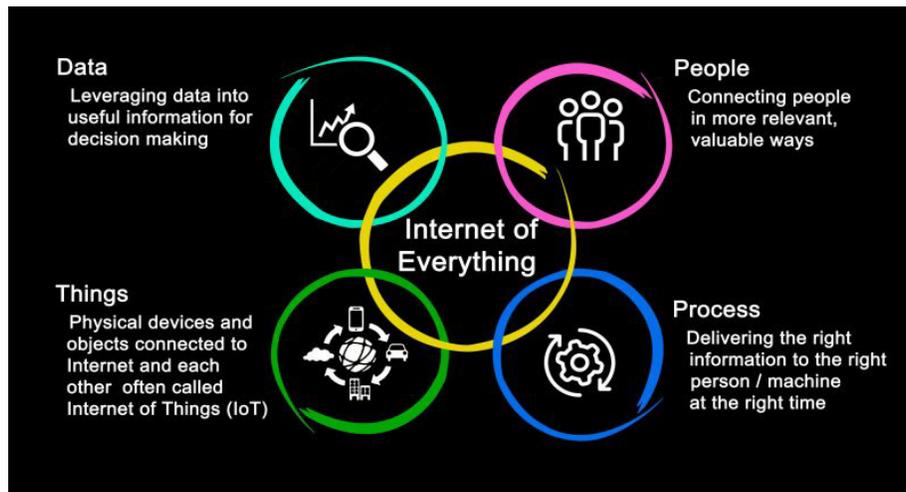
“We do what we normally do, but digitalization allows us to do it better or develop better products within our industry.” As companies digitalize products and process, completely **new ways of doing business in industries** emerge. Gartner analysts expect more transformational changes as digitalization re-invents industries at the business **model** level. Mr. Le Hong gave the examples of Nike, playing on the edge of the healthcare industry with its connected sporting clothes and gear, and Google having a visible presence in autonomous vehicles. “These organizations had no business in your industry, and are now re-inventing them,” said Mr. Le Hong.

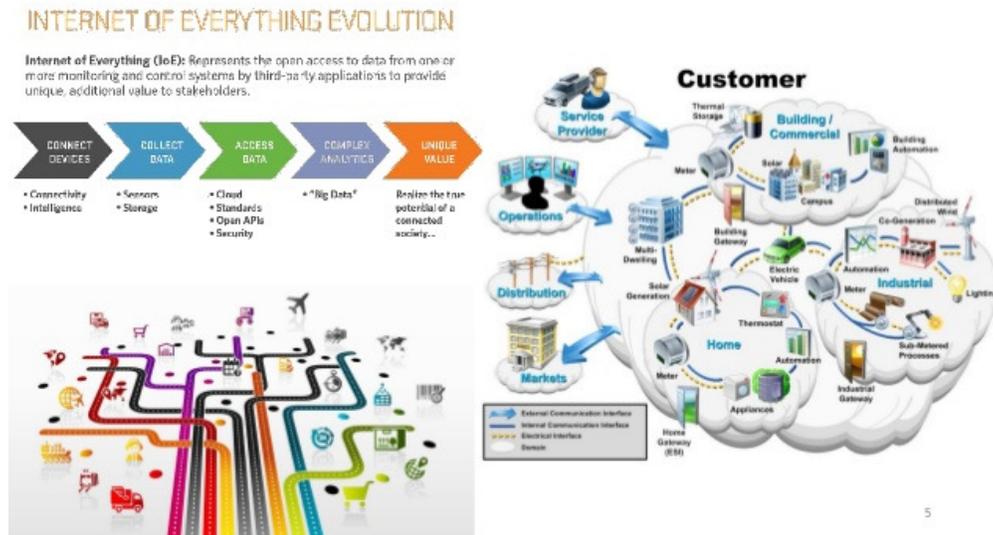
The third level of digital re-invention is created by **the need to compete with unprecedented business velocity and agility**. Gartner calls this the “business **moment**.”

Research Challenges for the Future IoE

The Internet of Everything will create tens of millions of new objects and sensors, all generating real-time data. “**Data is money**,” said Nick Jones, research vice president and distinguished analyst at Gartner. “Businesses will need BIG DATA and **storage technologies** to collect, analyze and store the sheer volume of information. Furthermore, to turn data into money business and IT leaders will need decisions. As they won’t have the time or the capacity to make all the decisions themselves they will need processing power.” “Now that digital is embedded in everything we do, every business needs its own flavor of digital strategy. Vanilla is off the menu,” said Dave Aron, research vice president and Gartner Fellow. “Digital is not an option, not an add-on, and not an afterthought; it is the new reality that requires a comprehensive digital leadership.” *Gartner* predicts that enterprises will make extensive use of IoE technology, and there will be a wide range of products sold into various markets. These will include advanced medical devices, factory automation sensors and applications in industrial robotics, sensor motes for increased agricultural yield, and automotive sensors and infrastructure integrity monitoring systems for diverse areas such as road and railway transportation, water distribution and electrical transmission; an endless list of products and services. But as devices get more connected and collect more data, **privacy** and **security** concerns will increase too. How companies decide **to balance customer privacy with this wealth of IoE data will be critical**.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019





References:

- <http://www.cisco.com/web/about/ac79/innov/IoE.html>
- <http://internetofeverything.cisco.com/>
- <http://www.cisco.com/web/solutions/trends/iot/overview.html>
- <http://time.com/#539/the-next-big-thing-for-tech-the-internet-of-everything/>
- <http://www.gartner.com/newsroom/id/2621015>
- <http://www.livemint.com/Specials/34DC3bDLSCItBaTfRvMBQO/Internet-of-Everything-gains-momentum.html>
- <http://www.tibco.com/blog/2013/10/07/gartners-internet-of-everything/>
- <http://www.eweek.com/small-business/internet-of-everything-personal-worlds-creating-new-markets-gartner.html>

INNOVATIVE TECHNOLOGY TO PROTECT FARMERS FROM SNAKE BITE USING IoT

1. CHENNUR KEERTHIKA REDDY

2. VADLAPUDI. POOJITHA

3. ARURU. GURUGAYATRI

1(computer science, Narayana engineering college, gudur, keerthikareddychnur@gmail.com)

2(computer science, Narayana engineering college, gudur, vadlamudipoojitha1120@gmail.com)

3(computer science, Narayana engineering college, gudur, gayatriaruru999@gmail.com)

ABSTRACT: The main objective of this paper is to protect the farmers from the snakes bite. Agriculture is the back bone to India and it generates more income to the country. More than 60% of the people are depending on the agriculture and working in the field day and late nights. They have to work in the crops, bushes and water. There are many dangerous snakes roaming in and around crops, bushes and water. There may be chances to face snakes bite and may be died because of lack of medical facilities nearby and it is very difficult to track the snakes. To overcome this problem here new method is proising based on IoT technology to track snakes automatically. This improved random forest decision tree technique can recognizes the snakes and classifies the snakes by extracting features and it also gives the warning alert message to the former so that he can be saved in the rural areas.

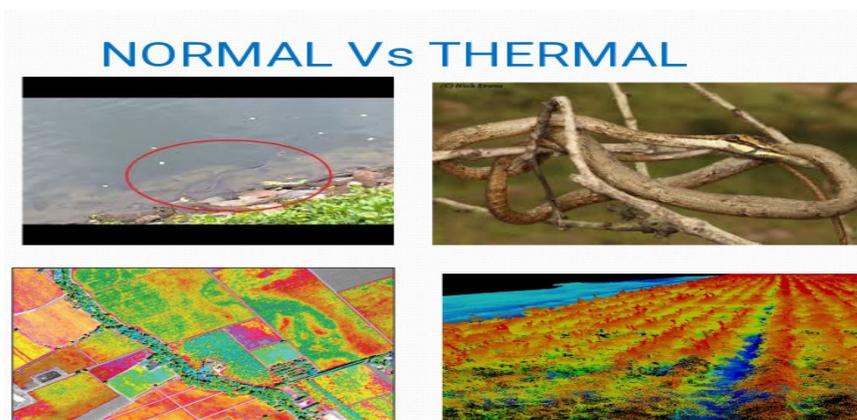
Keywords: Classification algorithm, feature extraction, Thermal imaginary, Raspberry Pi, Ultrasonic sensor.

1. INTRODUCTION :

We all are well known about farming .In India 60% of people depend on farming ,they need to work during day time and late nights in such case very harmful snakes will be roaming in and around crops,bushes,and water.so there may be changes to face snake bites which leads to death due to lack of medical facilities

near by. It is impossible to a human to identify snake which is hidden in crops ,s o to protect farmers from snake bites here new method is prosing based on IoT technology to track snakes automatically . This improved random forest decision tree technique can recognizes the snakes and classifies the snakes by extracting features and it also gives the warning alert message to the former so that he can be saved in the rural areas. **2.OBJECTIVES OF THIS TECHNOLOGY**

Monitoring the thermal images in different time Provide an appropriate algorithm to extract the features efficiently Detect the object and provide the precaution to the farmer



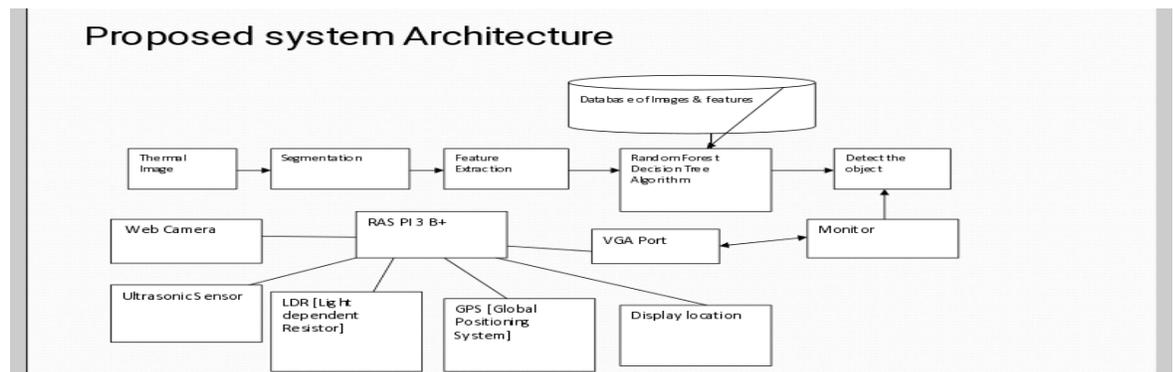
3.EXISTING SYSTEM :

HARR algorithm: Extracts the unique key features from all the wild animals.

- **Draw Backs:**
- Only detects the intrusion
- Does not ensures any prevention methods
- It does not efficient for all time

4. PROPOSED SYSTEM

Based on the drawbacks of the HARR algorithm, introducing a new technology which detects the snakes and give precautions to the farmers and protects the farmer from snake bi proposed system is given bel below so, lets have a tes The architecture for proposed system is given below so, lets have a look.



segmentation: An automatic region based approach and the color from the input image is segmented.

Feature objects Extraction : Thermal attributes are necessary to be computed on each band of the input image to further processing it.

Thermal_Mean, Thermal_Max,

Thermal_Min, Thermal_STD

Texture Attributes :Texture attributes are also derived from each band.

Texture_Range,Texture_Mean Texture_Variance,Texture_Entropy

Shape Feature

Area,Solidity,Major_Length, Minor_Length

Random forest Decision tree classification

Symmetric structure of tree

Further budgeting principle

- RAS Pi 3 B+

It consists Ultrasonic Sensor, Light dependent resistor,Global positioning system, Buzzer

In this method, use of group of images that are composed from the images captured by camera and through the internet sources are used to create a database for Snakes .

We also have the ultimate night vision which detects the snake very easily and gives the precautions to the farmer.

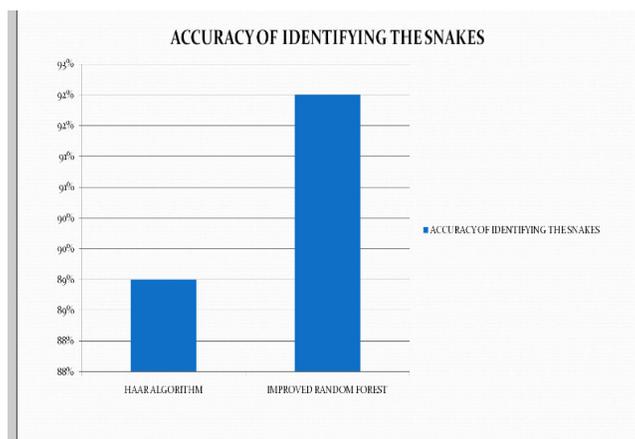


5. IMPLEMENTATION:

Thus once the snakes are identified as dangerous using the classification methods and produce alert buzzer in current location.

The proposed Improved Random Forest DT algorithm provides increased accuracy of identifying the objects than the all other approaches developed earlier. By implementing this technology we can detect the snakes very faster than the HARR algorithm .let us have an comparision between before and HARR algorithm.

6.STATISTICS:



Above graph shows the growth has taken place due to improved random forest technique

7.CONCLUSION AND FUTURE WORK:

Thus, novel technique to detect dangerous species i.e., the snakes in order to prevent Human-species Collision .To develop a faster method to cover large area of land, with a greater training dataset, the efficiency for the species detection and prevention system is likely to increase. The detection and tracking of motion of multiple or groups of species will also be an area of focus in the future studies.

8.REFERENCES:

- Muthukumar. N and Ravi. R, 'Hardware Implementation of Architecture Techniques for Fast Efficient loss less Image Compression System', Wireless Personal Communications, Volume. 90, No. 3, pp. 1291-1315, October 2016, SPRINGER.
- R. K. Vigneshwar, R.Maheswari –Development of Embedded Based System to Monitor Elephant Intrusion in Forest Border Areas Using Internet of Things| International Journal of Engineering Research. (1 July, 2016) Volume No.5, Issue No.7, pp : 594-598.
- Matthias Zeppelzauer and Angela S. Stoegaer –Establishing the fundamentals for an elephant early warning and monitoring system| IEEE International Conference on Image Processing .(4 September, 2015).

“Internet of Things” Applications, its Future and Challenges

Rayees Fathima

Abstract: The *Internet of things* (IoT) is the connection of physical devices, vehicles, home appliances, and other items embedded with electronics, software, sensors, actuators, and connectivity which enables these *things* to connect, collect and exchange data, creating opportunities for more direct integration of the physical world. IoT can also be considered as an extension of the Internet in which large numbers of “things”, including sensors, actuators and processors, in addition to human users, are connected together and able to provide high resolution data on their environment and can work out a degree of control over it. IoT is still in an early stage of growth, and there are many problems/research challenges must be resolved before it is widely adopted. Many of these are technical, including interoperability and scalability, as billions of heterogeneous devices will be interconnected, but determining on how to invest in the IoT is a challenge for business, and there are also major social, legal and ethical challenges, including security and privacy of data collection, which must be determined. As we know the future of IoT will be gaining a way into multi-national, multi-industry, multi-technology setup, this paper reviews the various applications, challenges and future of IoT as well discusses balancing the technical and non-technical research challenges which confront the IoT.

Keywords: Internet of Things; RFID, NFC, Heterogeneous, Organizational Challenges, Security, Technology.

1. INTRODUCTION

Internet of Things (IoT) term signifies a general concept for the ability of network devices to be aware of the devices around and collect data from around the world, and then share that data across the Internet where it can be processed and used for various interesting purposes. The IoT consists of smart devices interacting and communicating with other devices, objects, environments and setups. In today's world every person is connected with each other using different modes of communication. In all those the most popular method of communication is the internet. When objects like household appliances are interconnected to a network, they can work together in cooperation to provide the ideal service as a complete system and not as a collection of independently working devices. This is suitable for many of the real world applications and services, and one would for example apply it to build a smart home; windows can be closed automatically when the air conditioner is turned on, or can be opened for oxygen when the gas oven is turned on. The idea of IoT is especially very useful for persons with disabilities, as IoT technologies can support human

activities at larger scale like building or society, as the devices can mutually cooperate to act as a complete system. A profound evolution of the current Internet into a Network of interconnected objects that not only yields information from the environment (sensing) and interacts with the physical world (actuation/command/control), but also uses prevailing standards of Internet in order to provide services like information transfer, analytics, and communications. Powered by the pervasiveness of the devices supported by open wireless technology such as Bluetooth, radio frequency identification (RFID), WiFi, and telephonic data services along with embedded sensor and actuator nodes, IoT is marching out of its infancy and is on the approach to changing the current static Internet into a fully integrated Future Internet. The Internet revolution led to the interconnection between people at an extraordinary scale and pace. The next revolution will be the interconnection between devices to design a smart environment. Presently there are 9 billion interconnected devices and it is estimated to reach 24 billion devices by 2020. These days in every place like at railway station, shopping malls, in colleges an information desk is mandatory that provides information about the train schedule, promotional offers and important notice instantly. From educational organization viewpoint, the drawback is that it requires some staff that is devoted for that purpose and that must have up to date information about the institute and the latest activities in the institute. The second difficulty is that a person needs to go in the institute at the information desk in order to get information from them. A way out of this is to apply a technology and make technology accountable to answer all the queries asked by people. The best tool is Cell phones, which are available with everyone and which is connectable to internet to download latest information. With IoT in command you would be able to switch on air conditioning even before reaching home or switch off lights even after you have left home. Or you can even unlock the doors to friends for temporary access when you are not at home. With IoT's taking appearance and growing needs, companies are building products to make your life simpler and convenient. Smart Home has become the innovative ladder of accomplishment in the domestic spaces and it is predicted Smart homes will become as common as smartphones. There are three types of technologies that enable the internet of things, 1. Near-field Communication (NFC) and Radio Frequency Identification (RFID) In the 2000s, RFID was the prevailing technology. Few years later, NFC became influential. NFC have become common in smart phones during the early 2010s, with uses such as reading NFC tags or for access to public transportation.

- ii. Quick response codes and Optical tags are used for low cost tagging. Phone cameras decode QR code using image processing techniques. In reality QR advertisement canvassing gives less audience as users need to have another application to read QR codes.
- iii. Bluetooth and low energy This is one of the latest techniques. All newly releasing smartphones have BLE hardware in them. Tags based on BLE can signal their presence at a power budget that enables them to operate for up to one year on a lithium coin cell battery.

The area of Operations Research (OR) is an applied discipline which aims to help resolve the real world problems. It includes many mathematical tools and techniques, plus the sub-discipline of Systems Thinking, which itself has many varieties, both quantitative and qualitative. Both the forms of OR can assist in the design, management and use of the Internet of Things (IoT).

OR and the IoT have many functional areas, to which they may both be used jointly or separately, e.g., the “smart city”, where the OR techniques of routing, scheduling, discrete-event simulation, etc., may enable more capable traffic management, energy usage, etc. The “things” composing the IoT include processors which can carry out some of the computational tools and techniques of OR, so this part of OR can be considered part of the IoT. Many papers have been written on the IoT, mainly dealing with the accompanying technologies and technical research challenges, but increasingly also dealing with the IoT business ecosystem, and the social, legal and ethical problems that will arise with its adoption. Study of IoT as an entire system required knowledge from many technical disciplines, including distributed systems, mobile computing, human-computer interaction, cloud computing, and the many business, domestic and personal fields to which the IoT is or will be applied.

OR, including Systems Thinking, may make a significant contribution. The research approach we take is, in part, a survey of existing papers which apply particular OR tools, techniques and systems methodologies to the IoT. Some IoT research challenges may not have been tackled by OR as yet, or we may not have found any accounts thereof, and in such cases the authors use their knowledge of OR, systems methodologies and the IoT to outline how we think these approaches could support the IoT.

II.LITERATURE REVIEW

In literature [10] the IoT refers as intelligently connected devices and systems to gather data from embedded sensors and actuators and other physical objects. IoT is expected to spread fast in coming years a new aspect of services that improve the quality of life of customers and productivity of enterprises, opening an opportunity. This new trend of interconnectivity is going beyond tablets and laptops; to connected cars and buildings; smart meters and traffic control; with the prospect of intelligently connecting almost anything and anyone. This is what the GSMA refers to as the "Connected Life". The author in [11] describes the concept of sensor networks which has been made viable by the convergence of microelectro-mechanical systems technology, wireless communications. Firstly the sensor networks functioning and sensing task are explored, and according to that the evaluation factors guiding the design of sensor network is provided. Then the algorithms and protocols developed for each layer and the communication architecture for sensor networks is sketched. The literature [13] presents a three layered network construction of Internet of Things (IOT) communication method for high-voltage transmission line which includes the wireless self-organized sensor network (WSN), optical fiber compound overhead ground wire (OPGW), general packet radio service (GPRS) and the Beidou (COMPASS) navigation satellite system (CNSS). The function of each layer of network, application deployment and management of energy consumption are studied. The technique can meet the requirements of interconnection between the monitoring center and terminals, reduce the terminals' GPRS and CNSS configuration and OPGW optical access points, and guarantee the online monitoring of data transmission realtime and reliable under the condition of remote region, extreme weather and other environmental conditions.

III.APPLICATIONS

1. Smart home

Smart Home evidently stands out, ranking as highest Internet of Things application on all computed channels. More than 60,000 people currently search for the term "Smart Home" each month. The IoT Analytics company database for Smart Home embraces 256 companies and startups. More companies are active in smart home than any other application in the field of IoT. The total expanse of funding for Smart Home startups currently exceeds \$2.5bn.

2. Wearables

Wearable's remains a scorching topic too. There are plenty of wearable innovations to be thrilled about: like the Sony Smart B Trainer, the Myo gesture control, or LookSee bracelet. Of all the IoT startups, wearable's maker Jawbone is probably the one with the biggest funding to date. It stands at more than half a billion dollars!

3. Smart City

Smart city traverses a wide variety of uses, from traffic management to water supply, to waste management, urban security and environmental monitoring. Its reputation is fueled by the fact that many Smart City solutions assure to ease real pains of people living in cities these days. IoT solutions in the area of Smart City solve traffic jamming problems, reduce noise and pollution and help make cities safer.

4. Liquid Presence: Liquid detection in data centers, sensitive building grounds and warehouses to prevent breakdowns and corrosion.

5. Radiation Levels: In nuclear power stations surroundings distributed measurement of radiation levels to generate leakage alerts.

6. Explosive and Hazardous Gases: Detection of gas leakages and levels in industrial environments, surroundings of chemical factories and inside mines.

7. Smart grids: A future smart grid vows to use information about the actions of electricity suppliers and consumers in an programmed style to improve the efficiency, reliability, and economics of electricity.

8. Smart farming: Smart farming is an often unnoticed business-case for the internet of Things because it does not really fit into the well-known classes such as health, mobility, or industrial. However, due to the remoteness of farming operations and the large number of livestock that could be monitored the Internet of Things could transform the way farmers work. But this idea has not yet reached large-scale attention. Nevertheless, one of the Internet of Things

applications that should not be underestimated. Smart farming will develop the important application field in the largely agricultural-product exporting countries.

9.Structural Health: Monitoring of vibrations and material conditions in buildings, bridges and historical monuments which will help in avoiding and preventing dangerous situations like structure falling etc.

10.Smart Roads: Intelligent Highways with warning messages and diversions according to climate conditions with which we can prevent any unexpected events like accidents or traffic jams.

11.Industrial Control:Indoor Air Quality: Monitoring of oxygen levels and toxic gas inside chemical plants to ensure workers and goods safety. Temperature monitoring: Monitor the temperature inside the industry.

12.Medical field :All Detection: Assistance for elderly or disabled people living alone. Medical Fridges:Monitoring and Control of conditions inside freezers storing medicines vaccines, and organic elements.

Sportsmen Care: Vital signs monitoring in high performance centers and fields.

Patients Surveillance: Monitoring of conditions of patients inside hospitals and in old people's home.

Ultraviolet Radiation: Measurement of UV sun rays to warn people not to be exposed in certain hours.

IV CHALLENGES

For the IoT industry to thrive there are three categories of challenges to overcome and this is true for any new trend in technology not only IoT: technology, business and society.

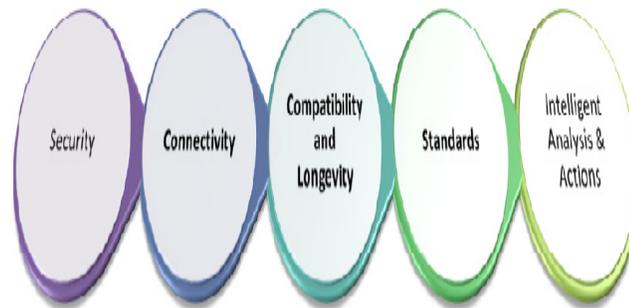


Figure 1: Technological Challenges

In Figure 1 we show the different technological challenges, including Security, Connectivity, Compatibility & Longevity, Standards and Intelligent Analysis & Actions.

Challenges facing the adoption of intelligent actions within IoT

- Machines' actions in unpredictable situations
- Information security and privacy
- Machine interoperability
- Mean-reverting human behaviors
- Slow adoption of new technologies

Business

The outcome is a big stimulus for starting, investing in, and operating any business, without a sound and solid business model for IoT we will have another effervesce, this model must satisfy all the requirements for all kinds of e-commerce; vertical markets, parallel markets, and consumer markets. But this class is always a victim of governing and legal scrutiny. End-to-end solution contributors operating in upright industries and providing services using cloud analytics will be the most successful at monetizing a large portion of the value in IoT.

Society

Understanding IoT from the customers and regulators prospective is not an easy task for the following reasons:

- Customer demands and requirements change constantly.

- New uses for devices—as well as new devices—sprout and grows at breakneck speeds.
- Inventing and reintegrating must-have features and capabilities are expensive and take time and resources.
- The uses for Internet of Things technology are expanding and changing—often in uncharted waters.
- Consumer Confidence: Each of these problems could put a dent in consumers' desire to purchase connected products, which would prevent the IoT from fulfilling its true potential.
- Lack of understanding or education by consumers of best practices for IoT devices security to help in improving privacy, for example change default passwords of IoT devices.

Research Challenges for the Future IoT

There are various research challenges associated with the IoT. We know some of them now, others will emerge in the future. These cover the complete field, including the technical challenges of designing, managing and using a multi-national, multi-industry, multi-technology infrastructure, the business challenges of developing IoT business models, and the organizational, political and social challenges of a new technology which promises to change the way we live and work in major ways.

Many current surveys of the IoT incorporate a section on research challenges, and we have tried to consolidate their results for our purposes. For example, a very high level research challenge might be “IoT design”, but this includes a number of lesser level research challenges such as “architecture”, “interoperability” and “scalability”. Each of these minor level research challenges may include other still subordinate level research challenges, e.g., IEEE’s Standard for an Architectural Framework for the IoT includes the research challenges of protection, security, privacy and safety [3]. Some authors contemplate IoT Standardization to be a research challenge in its own right, however we consider this to be a high level research challenge which encompasses many lower level research challenges.

We have split these research challenges into the classes of Design, Scientific/Engineering

and Management/Operations, though this is slightly artificial, as several research challenges belong to more than one group. For example, Reliability/Robustness is a challenge at both the design and operative stages, as is Security/Privacy.

Some of the technical and business challenges of the IoT may be resolved with the aid of the mathematical tools and techniques of Operations Research.

V FUTURE OF IoT

Soon IoT technology will turn out to be an evident part of our clothing as well. 968 thousand smart clothes made it to the consumer in 2015. This number is expected to grow to 24.75 billion by 2021. The adoption of connected home devices is expected to be higher than wearable's. More than two-thirds of consumers are likely to obtain IoT devices for their homes by 2019. Nearly half say the same for wearable tech. Of all the devices smart thermostats are expected to make it to 43% acceptance rate in the next 5 years.

Though the amount of implementation of smart devices is expected in the long run, most of the consumers (87%) aren't aware of the meaning of "The Internet of Things."

These inventions will bring a lot of savings in homes. Smart kitchens alone will provide minimum 15% savings in the F&B industry by 2020.

IOT is supposed to change the entire way people communicate, work and live. Now there will be connectivity for everyone, everything and everywhere. It is going to have an significant impact on how the businesses and government interact with the world.

REFERENCES

1. Mortenson, M.J.; Doherty, N.F.; Robinson, S. Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *Eur. J. Oper. Res.* 2015, 241, 583–595. [CrossRef]
2. Mingers, J.; White, L. A review of the recent contribution of systems thinking to operational research and management science. *Eur. J. Oper. Res.* 2010, 207, 1147–1161. [CrossRef]
3. IEEE. P2413—Standard for an Architectural Framework for the Internet of Things (IoT). 2016. Available online: <https://standards.ieee.org/develop/project/2413.html> (accessed on 12 March 2017).
4. Borgia, E. The Internet of Things vision: Key features, applications and open issues. *Comput. Commun.* 2014, 54, 1–31. [CrossRef]
5. Jain, R. Internet of Things: Challenges and Issues. In *Proceedings of the 20th Annual Conference on Advanced Computing and Communications (ADCOM 2014)*, Bangalore, India, 19–22 September 2014.
6. Stankovic, J.A. Research directions for the internet of things. *IEEE Internet Things J.* 2014, 1, 3–9. [CrossRef]

7. Mattern, F.; Floerkemeier, C. From the Internet of Computers to the Internet of Things. In *From Active DataManagement to Event-Based Systems and More*; Springer: Berlin, Germany, 2010; pp. 242–259.
8. Elkhodr, M.; Shahrestani, S.; Cheung, H. The Internet of Things: Vision & Challenges. In *Proceedings of the TENCON Spring Conference, Sydney, Australia, 17–19 April 2013*; pp. 218–222.
9. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* 2013, 29, 1645–1660. [CrossRef].
10. www.gsma.com/connectedliving/wpcontent/.../cl_iot_wp_07_14.pdf
11. http://www.libelium.com/top_50_iot_sensor_applications_ranking
12. I.F. Akyildiz, W. Su, Sankarasubramanian, E. Cayirci, Wireless sensor networks: a survey, *Computer Networks* 38 (2002) 393–422.
13. A. Menon¹, et al. "Implementation of internet of things in bus transport system of singapore" *Asian Journal of Engineering Research* (2013).
14. Shao-Lei Zhai et al. "Research of Communication Technology on IOT for High-Voltage Transmission Line " *International Journal of Smart Grid and Clean Energy* (2012)

IOT AND SMART HOME BASED ON LI-FI TECHNOLOGY

R.Subramanyam,
Narayana Engineering College,Gudur,AP.

Abstract

Smart homes are among the most interesting applications of Internet of Things that enhance the quality of human life and brings more comfort, savings, convenience and peace of mind. In this paper, we propose a smart home system that uses Li-Fi technology as medium of communication between all the connected devices and uses a video surveillance system based on Wireless Visual Sensor Network. Li-Fi is a high-speed bi-directional fully connected technology that provides transmission of data through illumination using LED light bulb. The use of such technology in our proposed system ensures a high level of security, high speed of data transmission, low energy consumption and more convenience

INTRODUCTION

Recent advances in wireless communications, cloud computing, Big Data and the availability of inexpensive wireless sensors, have led to the rapid development of the Internet of Things (IoT). Smart home, also referred to as home automation or eHome, is among the most interesting applications of IoT that makes human life more comfortable by providing connectivity and control of every digital devices in home, such as air conditioning, heating, ventilation, lighting and security systems, regardless of time and place. Smart home brings significant advantages to the human daily life such as:

- Comfort and convenience: by controlling the home from anywhere and at any time, and by setting the home devices to specific needs such as adapting rooms heating and air conditioner to the preferences of users or according to the weather changes.
- Accessibility: by enabling people with disabilities or special needs to live more independently by using assistive or adaptive technologies, for example, a person who cannot see can use voice activated interface to control his connected devices.
- Communication between the devices at home: by connecting devices, appliances and sensors in the network, they are able to communicate with each other, and can be controlled remotely.
- Energy efficiency: By providing the flexibility of monitoring electricity consumption to consumers[1] and permitting the system to switch on or off appliances when they are not

needed in order to save electricity. Otherwise, by using renewable energy to produce its own energy, such as installing photovoltaic systems on the home roof

- Security: By installing cameras, motion detectors, smoke and fire detectors, locks, etc., the user can monitor and view his home no matter where he is and can be notified immediately if something is out of the ordinary regarding the home's condition.

However, smart home application is facing a few challenges, which need to be overcome such as interoperability, cyber security / privacy and the interference issue caused by the radio waves of Radio Frequency (RF) technologies (such as WiFi) that interfere with other electronics in the home, leading to slowing its functionality or even stop it.

Many applications for smart home have been envisaged like smart lighting, security systems, smart appliances, elderly care and kindergartens. Most of these applications use standards of communication such as Bluetooth, Zigbee, Z-wave and WiFi to transmit the collected data from the home environment to a Base Station (BS) called home sink or home hub.

In this paper, we propose a smart home system that uses Light-Fidelity (Li-Fi) technology as medium of communication between all the connected devices and uses a video surveillance system based on Wireless Visual Sensor Network (WVSN).

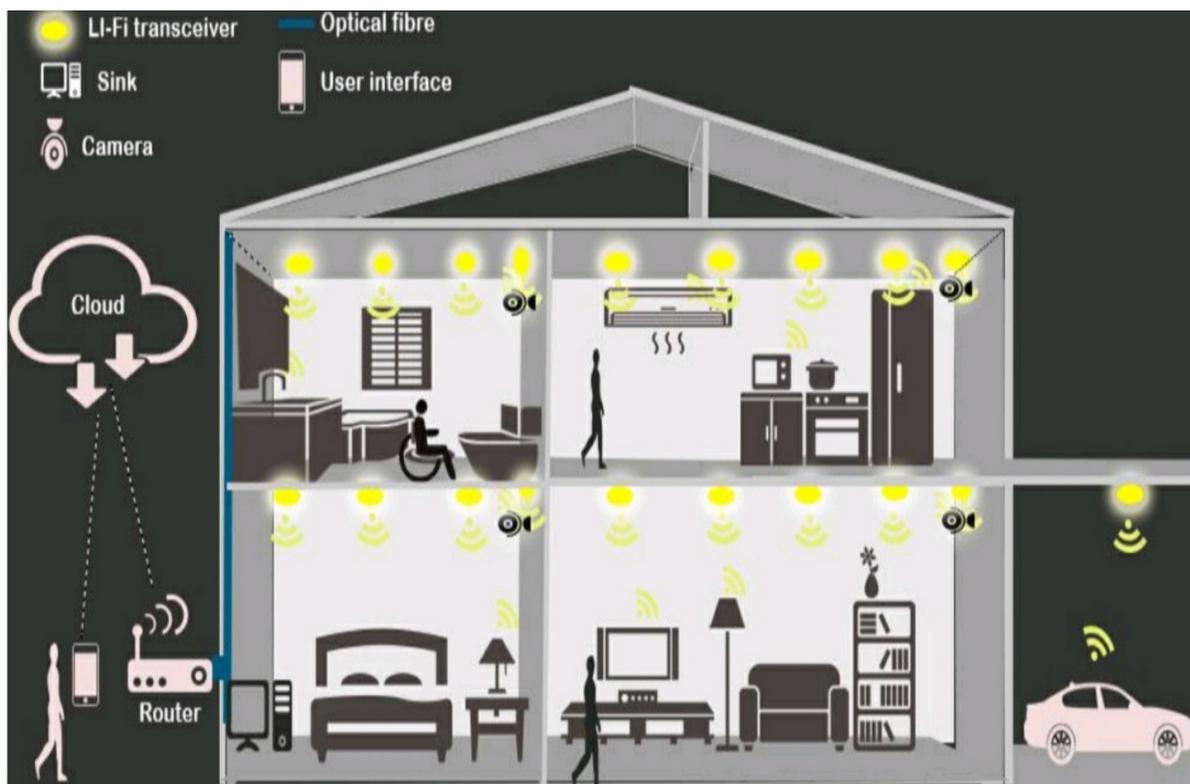


Fig. 1. Our Smart home system based on Li-Fi technology

Our Proposed System

In this section, we will present our proposed smart home system that uses Li-Fi technology for communication between the connected devices, and uses WWSN as system for surveillance.

Our proposed system concept is shown in Fig.1, where all devices are connected to each other and to the internet via Li-Fi and controlled by a user interface from anywhere and at any time using cloud computing.

Li-Fi is a high-speed bi-directional fully connected Visual Light Communication (VLC) technology that was proposed by the German physicist Harald Haas. It uses visible light with wavelength in the range of 380 nm– 750 nm [2] in order to provide transmission of data through a very high switching ON and OFF speed of LED light bulb illumination that cannot be tracked by the human eye. The idea behind choosing Li-Fi technology rather than traditional RF technologies is due to its several advantages [3] [4] [5] that are summarized in Fig.2. One of the best advantages of Li-Fi is that it can be used to provide both lighting and information in the same time. Li-Fi is also characterized by its higher bandwidth, which makes it more than sufficient for transmitting all types of data (scalar, video, etc.) in a very short time.

Unlike Wi-Fi technology, Li-Fi cannot travel through non-transparent material such as walls, which will provide more secure data transfer as it confines the data transmission to one area and it do not have any interference issue. The use of LED light in Li-Fi, makes it more suitable for indoor applications because it is cheaper and safer for eyes.

As a security system in our proposed smart home, we suggest the use of a WVSAN that consist of a large number of tiny visual sensor nodes called camera nodes, which integrate an image sensor, an embedded processor, and a wireless transceiver[6]. These nodes can collect image / video data from a region of interest, process it collaboratively, and transmit the useful information to the BS.

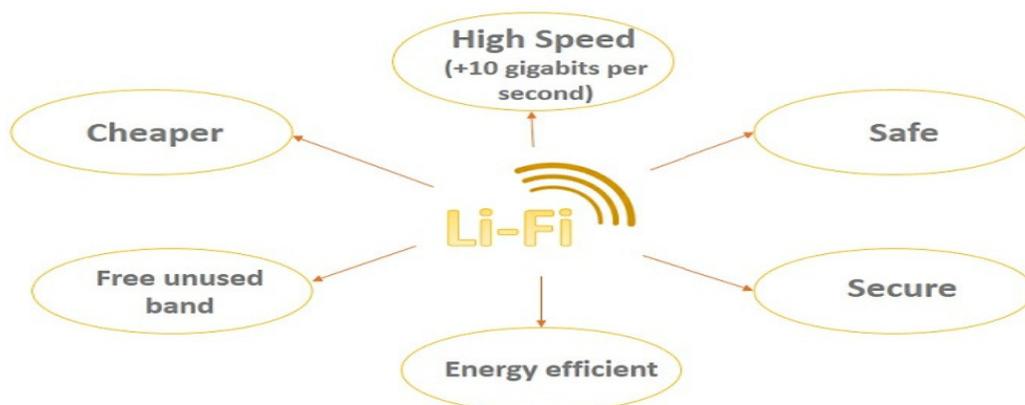


Fig. 2. Li-Fi advantages

The extension or the reparation of this system is easier because new cables installation are not needed to add new cameras in the network[7]. In this system, camera nodes will be able to communicate with other devices to complete a specific task. As a use case, we take elderly care.

USE CASE: Elderly Care

Due to chronic illness and declining health that affect most of older people, performing everyday tasks becomes more challenging and home environment becomes a place with a high-risk for falls[8].

Thus, smart home technology could be an alternative to minimize risks resulting from aging or disabilities and to facilitate old people independence and activity[9].

In our proposed system, camera nodes will be able to detect and monitor older person activities and collect other information such as walking speed, posture, balance, etc., and send the collected data to a repository in the BS, which will be then stored in the cloud in order to be accessed by his doctor or therapist and by the family. If something is out of the ordinary regarding the older person activity, a notification will be sent automatically to the concerned persons.

- Other sensors and devices can communicate with the camera nodes to facilitate older people daily life, by turning on/off water or lights automatically, adjusting room temperature according to the preferences of the old person, reminding him to take medicine in time and if he forgets, his family will be alerted, etc
- In addition, the older person can have real-time interactions with his doctor or his rehabilitation center in order to have a consultation or a training session remotely by using interactive telemedicine services
- Due to Li-Fi advantages presented in Fig.2 and the use of WWSN, our system will provide a cheap, fast, secure and energy efficient connected system, that enhances the quality of human life in a more comfortable home.

Conclusion:

In this paper, a cheap, fast, secure, and energy efficient smart home system based on Li-Fi technology as medium of communication and wireless visual sensor network as video surveillance system is presented. To evaluate the utility and the advantage of this system, an elderly care use case is discussed.

References

- [1] A. Bhati, M. Hansen, and C. M. Chan, "Energy conservation through smart homes in a smart city: A lesson for Singapore households," *Energy Policy*, vol. 104, no. February, pp. 230–239, 2017.
- [2] I. Meena and D. Kumar, "A Review Paper on Li - Fi," in *National Conference on Innovations in Micro-electronics, Signal Processing and Communication Technologies*, 2016, no. February, pp. 9–11.
- [3] P. Mishra, J. Poddar, and S. Priya, "A Review On LiFi : The Green WiFi," *Int. Res. J. Eng. Technol.*, vol. 3, no. 3, pp. 99–103, 2016.
- [3] P. Mishra, J. Poddar, and S. Priya, "A Review On LiFi : The Green WiFi," *Int. Res. J. Eng. Technol.*, vol. 3, no. 3, pp. 99–103, 2016.
- [5] S. Wu, H. Wang, and C. Youn, "Visible Light Communications for 5G Wireless Networking Systems: From Fixed to Mobile Communications," *Ieee Netw.*, vol. 28, no. 6, pp. 41–45, 2014.
- [6] S. Soro and W. Heinzelman, "A Survey of Visual Sensor Networks," *Adv. Multimed.*, vol. 2009, p. 21, 2009.

A Quick Review of Internet Of Things and its latest applications

Selam.Bhanu Prakash¹, Saney Dinesh Kumar Reddy²

B.Tech, CSE, Narayana Engineering College, Nellore Dt, India¹

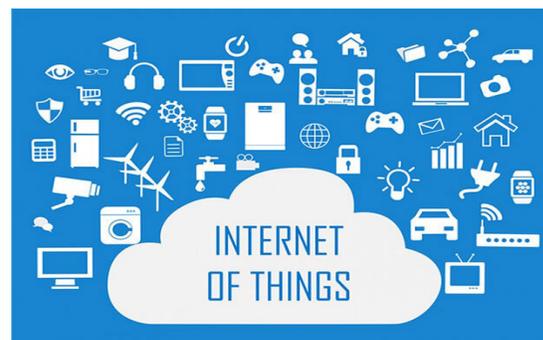
B.Tech, CSE, Narayana Engineering College, Nellore Dt, India²

Abstract: Internet of Things(IOT)is an globally emerging technology in the modern era from the support of internet. It is the concept through which we can control object through internet. It aims at making the internet even more immersive and pervasive.We can control the things around us with our finger tips with the help of electronic devices that has access to internet. Not only to do our work in a simple way ,it also sends us alert notifications about a un-common situations in the business point of view which may help us to prevent a huge loss, or to send alert notifications to the near by people who might be effected by the natural calamities by detecting the unusual environmental conditions.

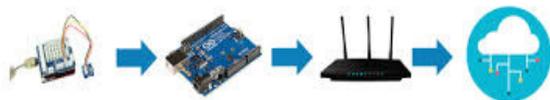
Keywords: IOT Devices, Sensors, Collected data in the database,Arduino and Raspberry Pi

I. INTRODUCTION

In 1964,Karl Steinbuch(a German Computer scientist) said ,“In a few decades of time computers will be interwoven into almost every industrial project”. The first IOT product was a roster that can be switched on/off using the internet and in the same year, some students also designed a coco-cola dispensing machine that is controlled using the internet. The actual growth of the Internet Of Things was started in 1990,and was coined by Kevin Aston. From the day of applying the technology on the simple items has been raised to applicability in most of the devices that are present in the contemporary days. Still it is being embedded into many things for many purposes. Lets see some of them here in the research paper



II. REQUIREMENTS OF THE IOT DEVICES



IOT devices are the electronic devices that has access to the internet and we need various sensors for various purposes like temperature sensor(for collecting temperature data), light sensors(for collecting the data on the intensity of light falling on an object) and ultrasonic sensor(for collecting the sound intensity that collects the corresponding data and sends that data into the database using some IOT platform by Arduino and Raspberry pi programming. And a buzzer is required when we need to provide sound alerts or else we can use LED

lights to give visual alerts.

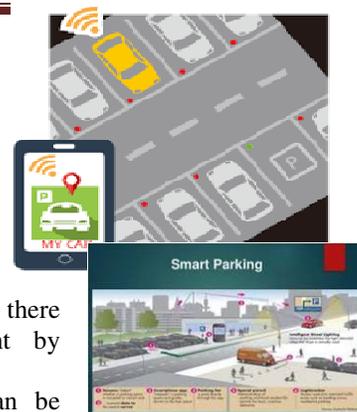
III.APPLICATION SENARIOS

IOT devices are used in many scenarios. Most of the people see that IOT is only to switch on/off an electronic device with our hands using a smart device like a smart phone or tab-let. But apart from these usage of IOT technology ,it can be used to a extinct level for many wonderful scenarios. For example,



DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

1) In **Smart industries**, IOT devices are deployed by the managers in a factory which needs an optimal temperature to be maintained then if something goes wrong, the data collected by the sensors is analysed with in the IOT platforms using the programming code of arduino or raspberry pi then it sends an alert notification to the respective person that there is an problem in the factory so that the manager can take an immediate action and can prevent the company from great loss.



2) In **Smart Plantation**, IOT devices can also be used to determine whether there is an adequate sun light falling on the plant by using light sensors.



3) In **Smart Agriculture**, IOT devices can be installed in the fields to determine the level of water. If the water level reaches maximum range then water pumps can automatically switched off. Similarly if the water level falls below the dead level then water pumps can be switched on automatically by using the Internet of things technology.

4) These IOT devices can also be used to detect unusual changes in the environment and might predict the probability of occurring a natural calamities (tsunami or earthquake or cyclone or flood) and warn the people

who are very close to that region.

5) In **Smart Parking** IOT can be used to provide the driver a best available parking slot to park the vehicle.

6) In **Smart lighting system** using the IOT technology, the intensity of the lights is adjusted based on the crowd on the streets or else on roads that reduces the power consumption.

7) Proper **maintenance of the historical buildings** of a city. IOT may provide a distributed database of building structural integrity measurements, collected by suitable sensors located in the buildings, such as vibration and deformation sensors to monitor the building stress, atmospheric agent sensors. It includes vibrations, temperature and humidity sensors.

8) In **Smart Traffic**, By using the sensing capabilities and GPS installed on modern vehicles, and also adopting a combination of air quality and acoustic sensors along a given road. To discipline traffic and to send officers where needed and for the latter to plan in advance the route to reach the office when needed

III. CONCLUSION

With the day progress of internet, IOT technology is also getting popular with many future innovations. IOT Product developers are now concentrating much on the development of Smart cities, Growing security concerns, Connected wearables, Smart stores, The rise of 5G, Edge computing, and many more. In a few years, everything will be digitalized and with the fastest internet speed, we can easily control the devices around us and make accurate prediction on unusual situations and take precautions. In view of all these applications, we can Internet Of Things (IOT) technology is a boon in contemporary world.

ACKNOWLEDGMENT

REFERENCES

[1] Design Spark, '11 Internet of Things (IOT) Protocols you need to know about'. Accessed December 10, 2016.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- [2] Young Hua Ling, Jiabin Tang, Qing Yang, Chao Zui , "Wireless Communication for IOT(Internet of Things)", IBM Research. Accessed December 17, 2016.
- [3] Pavana NR and Dr. M.C. Padma, "Design of Low Cost System for Real Time Monitoring of Water Quality Parameters in IOT Environment", International Journal of Advanced Research in Computer Science and Application Volume 4, Issue 5, May 2016.
- [4] A.N. Prasad, K.A. Mamun, F.R. Islam, H. Haqva," Smart Water Quality Monitoring System", IEEE, 2015.
- [5] N Vijayakumar, R Ramya,"The Real Time Monitoring of Water Quality in IOT Environment", International Conference on Circuit, Power and Computing Technologies, IEEE, 2015.
- [6]. R.M. Bhardwaj, "Overview of Ganga River Pollution", Report: Central Pollution Control Board, Delhi, 2011
- [7]. NivitYadav, "CPCB Real time Water Quality Monitoring", Report: Center for Science and Environment, 2012
- [8]. Tuan Le Dinh, Wen Hu, PavanSikka, Peter Corke, L. Overs, Stephen Brosman, "Design and Deployment of a Remote Robust Sensor Network: Experiences from Outdoor Water", 32nd IEEE Conf. on Local Computers, pp 799-806, Feb., 2007
- [9]. Quio Tie-Zhn, Song Le, "The Design of Multiparameter On line Monitoring System of Water Quality based on GPRS", Report: Advanced Transducers and intelligent Control System Lab, Taiyuan Technical University, Taiyuan, China, 2010
- [10]. Steven Silva, Hoang N Ghia Nguyen, Valentina, Tiporlini, Kamal Alameh, "Web based Water Quality Monitoring with Sensor Network: Employing ZigBee and WiMAX Technology", 36th IEEE Conf. on Local Computer Networks, 2011
- [11]. Donge He, Li-Xin Zhang, "The Water Quality Monitoring System based on Wireless Sensor Network" Report: Mechanical and Electronic Information Institute, China University of GeoScience, Wu Hen, China, 2012
- [12]. Pavlos Papageorgiou, "Literature Survey on Wireless Sensor Networks", Report: University of Maryland, 16 July 2003
- [13]. SatishTurken, Amruta Kulkarni, "Solar Powered Water Quality Monitoring System using Wireless Sensor Network", IEEE Conf. on Automation, Computing, communication, control, and compressed sensing, pp281-285, 2013.

Cyber – Physical Systems and Applications

D. Suhasini

Asst. Professor, HOD of Computer Science

AV College of Arts, Science & Commerce

Gagan Mahal, Hyderabad-029, TS, India.

dabbara.suhasini@gmail.com

ABSTRACT

Cyber-Physical System is integrations of computation and physical processes. It is similar to Internet of Things. A CPS is composed of highly – integrated Computation, Communication Control and Physical elements. CPS is currently of interest in academia, industry and government. Cyber space is being linked to versatile individuals in physical space and social space – CPS. With the rapid progresses in ICT (Information Communication and Technology), multidisciplinary research fields such as IoT (Internet of Things), Cyber-Physical System (CPS) and Social Computing have been widely explored in recent years. CPS are characterized by tight coupling and interaction between computation, communication and control elements and physical processes such as motion, heating, vibration. The significance of CPS is to connect the physical device to the internet, allowing the physical device to have some functions of computing communication, autonomy, remote coordination, precise control. In order to understand CPS and its future potential in manufacturing, definitions and characteristics of CPS are explained and compared with cloud manufacturing concept.

I. Introduction:

The recent technology convergence of sensor devices, networks, and communications has been creating new demands and opportunities for cyber-physical systems and their applications. In particular, the growing number of the heterogeneous sensor and actuator devices and the demanding networking systems of wired/wireless communications in cyber-physical systems has imposed several challenges in both traditional and new research issues. Moreover, human and nonhuman systems interact consistently and allow each to achieve a set of nontrivial technical goals. This Special Issue aims to gather recent advanced technologies and applications that address network protocol design, low latency networking, context-aware interaction, energy efficiency, resource management, security, human–robot interaction, assistive technology and robots, application development, and integration of multiple systems that support cyber-physical systems and smart interaction. Cyber-physical systems (CPS) have been at the core of critical infrastructures and industrial control systems for many decades, and yet, there have been few confirmed cases of computer-based attacks. Cyber-Physical Systems (CPSs) integrate the dynamics of the physical processes with those of the software and communication, providing abstractions and modeling, design, and analysis techniques for the integrated whole. The technology

depends on the multi-disciplines such as embedded systems, computers, communications, etc. and the software is embedded in devices whose principle mission is not computation alone, e.g. cars, medical devices, scientific instruments, and intelligent transportation systems. Cyber Physical System is a system featuring a tight combination of, and coordination between, the system's computational and physical elements. CPS uses computations and communication deeply embedded in and interacting with physical processes to add new capabilities to physical system. Convergence of computation, communication, and control. Cyber-Physical Systems (CPS) comprises interacting digital, analog, physical, and human components engineered for function through integrated physics and logic.

II. Applications of Cyber-Physical Systems

1. Modern manufacturing

CPS is technical systems where networked computers and robots interact with the physical world. Found in a wide range of services and applications, CPS are quickly becoming a part of modern manufacturing processes, making it necessary to examine the impacts they will have in this area. The [Science and Technology Options Assessment](#) (STOA) Panel recently published a study on the '[Ethics of Cyber-Physical Systems](#)' (CPS).

1.1 What changes are we going to see in manufacturing?

Use of CPS in manufacturing could result in massive changes to the way the manufacturing process is currently conducted, leading to a fourth industrial revolution, '[Industry 4.0](#)'. CPS could help manufacturing through the [continuous miniaturization of sensors and actuators](#), driven through current developments and advances in nanotechnology. As smart manufacturing requires massive amounts of real-time data, gathered through sensors, in order to function, the miniaturization of sensors will help pave the way for Industry 4.0. The development of smaller sensors, combined with the new internet protocol, [IPv6](#) developed in 2012, allows the sensors to become part of the '[internet of things](#)', where everything is interconnected online – a defining feature of Industry 4.0. These changes will lead to radical new manufacturing business models, with [data becoming a competitive asset](#), much as it is for internet firms like Google or Facebook.

2. Humanoid Robots

Robots and other complex cyber-physical systems (CPS) sense, process, and react to information from the physical world. They must operate safely even in the presence of uncertainties and resource constraints. To enable advanced robotics and CPS applications, research in this area tackles a wide range of issues including visual perception, inference from empirical data, motor learning and control, and the design, implementation, and verification of safe and performant CPS. The use of robots is moving rapidly beyond

controlled environments such as factories to complex environments in the midst of human activity, demanding a nimble cross-disciplinary approach. Princeton engineers are advancing the productive, safe, and ethical use of robotics in society by building and connecting expertise in sensing, artificial intelligence, neuroscience, public policy, and other fields.

3. Intelligent Transportation System

A cyber-physical system (CPS) is composed of a physical system and its corresponding cyber systems that are tightly fused at all scales and levels. CPS is helpful to improve the controllability, efficiency and reliability of a physical system, such as vehicle collision avoidance and zero-net energy buildings systems. It has become a hot R&D and practical area from US to EU and other countries. In fact, most of physical systems and their cyber systems are designed, built and used by human beings in the social and natural environments. So, social systems must be of the same importance as their CPSs. The indivisible cyber, physical and social parts constitute the cyber-physical-social system (CPSS), a typical complex system and it's a challengeable problem to control and manage it under traditional theories and methods. An artificial systems, computational experiments and parallel execution (ACP) methodology is introduced based on which data-driven models are applied to social system. Artificial systems, i.e., cyber systems, are applied for the equivalent description of physical-social system (PSS). Computational experiments are applied for control plan validation. And parallel execution finally realizes the stepwise control and management of CPSS. Finally, a CPSS-based intelligent transportation system (ITS) is discussed as a case study, and its architecture, three parts, and application are described in detail.

4. Medical CPS

With the rapid transformation for various medical systems, there is a strong requirement for new devices with increased functionalities. The term Medical Cyber Physical system refers to a system that has combination of embedded devices, software for controlling these devices and communication channel for interaction. For developing safe and effective MCPS requires new design, verification and evaluation techniques due to increase in size and complexity. And the challenges for developing these kinds of systems include executable clinical workflow, model based development, physiological close-loop control, adaptive patient specific algorithms, smart alarms and user centered design and infrastructure for medical integration and interoperations. The application scenario for MCPS varies from patient monitoring, analgesic infusion pumps to implant sensors devices. Cyber physical medical system modeling and analysis is a framework proposed by for safety verification of different applications. The scenario considered for the experiment was analgesic infusion pump control algorithm for keeping the drug concentration in the blood for a fixed level. These systems are example for typical closed loop systems. Any change in the physical world can directly influence it in cyber world

and an action will be taken at the physical world based on the instruction given from the cyber space.

5. Green Buildings

Greenhouse effect is one of the major problems in today's world. The old buildings consume 70% of the electricity produced and generate the green house gases which in turn increase green house effect. By using the integrated Wireless Sensor Network, cognition manager and control systems we can achieve Zero Net Energy goal.

6. Aeronautic Applications

CPS are of great importance to Boeing, the Aerospace Industry• “The challenges today are far greater than those faced in even the recent past and continue to grow as individual systems evolve, operate with greater autonomy and intelligence, and operate as part of a networked system of systems,”• “Requirements for cyber-physical systems and software are far more stringent than those for typical office automation applications. Our systems must support real-time behavior. We require ultra-high reliability and many of our systems are safety critical and require certification.”

6.1 Today's Aerospace Systems are Increasingly CPS- Intensive

- Aerospace systems for today and beyond
- New capabilities
- Agile behavior in highly dynamic operating environments
- Operation in a SoS Network
- Avionics S/W challenges – 100M – >1B SLOC
- Software Intensive Systems
- Multiple levels of criticality

7. Metamodeling approaches relevant to modeling cyber-physical systems in civil engineering

For modeling cyber-physical systems, a variety of metamodeling approaches can be applied. In this section, basic concepts of metamodeling are illuminated and three common metamodeling approaches are analyzed for describing information related to cyber-physical systems, (i) the Unified Modeling language (UML) and additional modeling languages published by the Object Management Group (OMG), (ii) the seven standards forming the Sensor Web Enablement (SWE) framework, and (iii) the data modeling language EXPRESS, all following the object-oriented paradigm, are metamodeling approaches frequently used in computing in civil engineering.

In software engineering and systems engineering, models are used to capture real-world aspects of problem domains with different levels of abstraction. The term “modeling” describes design techniques and development processes that require technical frameworks for information integration and for tool interoperability. Software and systems engineering approaches based on models are described in technical frameworks, referred to as model-

driven development (MDD). The MDA framework has been developed to separate specifications of system functionalities from platform-specific system implementations. Therefore, the MDA design process starts with a platform-independent model (PIM) describing functionalities and behavior of a system.

III. Cyber Physical Systems: Advantages and Disadvantages

Cyber Physical systems are computers that perform functions in real time as a result of stimuli in the physical world. They are prevalent in various parts of everyday life. Please comment about any uses you have made with Cyber Physical Systems.

Advantages of Cyber Physical Systems

- Fast way to ensure safety in various real world processes
- Ensures efficiency in various real world processes
- Improvement in life quality for countless people
- Potential to bring a positive revolution to the world
- Can perform countless calculations instantaneously
- **Disadvantages of Cyber Physical Systems**
- Possible Terminator type scenario
- Unemployment
- Unpredictability
- Loss of purpose in life
- Computers gaining self awareness

IV. Conclusions

This paper discussed the advantages and disadvantages and applications of Cyber physical systems in different domains briefly. CPS provides better solutions to some of the real time problems facing in today's world. Cyber Physical Systems change the way how humans interact with the physical world.

V. References:

- [1]https://www.researchgate.net/profile/Jiafu_Wan2/publication/228934884_A_Survey_of_Cyber_Physical_Systems/links/550b9aa0cf2855640971ce5/A-Survey-of-Cyber-Physical-Systems.pdf
- [2] <https://pdfs.semanticscholar.org/d514/97e5827cc00d9d00c26e27a769d42284cfba.pdf>
- [3] https://www.mdpi.com/journal/sensors/special_issues/smart_interactive_cyber-physical_systems

- [4] Lee Insup and Oleg Sokolsky, Medical Cyber Physical Systems, in Proc. of DAC, USA 2012.
- [5] Banerjee, Ayan and Gupta, Sandeep K. S. and Fainekos, Georgios and Varsamopoulos, Georgios, Towards modeling and analysis of cyberphysical medical systems, Proc. ISABEL'11, pp:154-158, 2011
- [6] Arney, David and Pajic, Miroslav and Goldman, Julian M. and Lee, Insup and Mangharam, Rahul and Sokolsky, Oleg, Toward patient safety in closed-loop medical device systems, ICCPS'10, pp:139-148, 2010.
- [7] <https://epthinktank.eu/2016/08/23/cyber-physical-systems-and-their-application-in-modern-manufacturing/>
- [8] <https://www.cis.mpg.de/robotics/>
- [9] <https://www.semanticscholar.org/paper/Cyber-physical-social-system-in-intelligent-Xiong-Zhu/c7cda5060d8c5d074889a6c298f868993b260432>
- [10] <https://smarsly.files.wordpress.com/2018/12/smarsly2019d.pdf>

DRUNK DETECTION FOR LOCKING IGNITION

K.Ravali

Asst.professor

Department of Computer Science & Engineering

Abstract

A new approach towards automobile safety and security to decrease the number of accidents caused due to the drunken drivers. It has a smart electronic system which continuously monitors the alcohol content in the air surrounding by the body of the protagonist. Speed of the vehicle varies on the content of alcohol detected. Vehicle based counter measure system progressively monitors the speed locking system of the vehicle running from high (100kmph), medium (60-80kmph) and low (40kmph) which helps the driver to reach destination safely. In extreme situations the system disables the vehicle by switching off the ignition. The Global Positioning System (GPS) captures the location. Only after the password is entered the vehicle can be restarted. In the back end, all the riding circumstances, amount of alcohol detected, vehicle speed and the location is uploaded into web server automatically and accurately as soon as the sign of alcohol detected for investigation purpose.

1. INTRODUCTION

Drunk driving is the reason behind most of the deaths, so the Drunk Driving Detection with Car Ignition Locking Using Raspberry Pi aims to change that with automated, transparent, noninvasive alcohol safety check in vehicles. The system uses raspberry pi with alcohol sensors, dc motor, and LCD display circuit to achieve this purpose. System uses alcohol sensor with, raspberry pi with dc motor to demonstrate as vehicle engine. System constantly monitors the sensitivity of alcohol sensor for drunk driver detection. If driver is drunk, the processor instantly stops the system ignition by stopping the motor. If alcohol sensor is not giving high alcohol intensity signals, system lets engine run. The raspberry pi processor constantly processes the alcohol sensor data to check drunk driving and operates a lock on the vehicle engine accordingly. At the same time it is connected to a network from where the person who is driving id is being monitored and if any necessity of help will be provided to him by the caretaker who will get the alert on the monitoring webpage automatically.

1.1 STATEMENT OF PROBLEM

Drunken driving is considered as one of the major reason of accidents in worldwide. Drivers under the influence of alcohol show a clear failure of perception recognition and vehicle control. So, bythis accident tends to make certain types of dangerous actions. Iris recognition technology is continuously growing over years which could resolve drunken driving accidents worldwide.The idea behind this is to avoid accidents by using alcohol sensor with, raspberry pi with dc motor to demonstrate as vehicle engine. System constantly monitors the sensitivity of alcohol sensor for drunk driver detection. If driver is drunk, the processor instantly stops the system ignition by stopping the motor. In this system we were able to lock the engine if the alcohol content is high. Hence in future we may even predict the accidents occurred, so that the injured person is rescued earlier without much delay. Automatic alcohol detection system which is also connected to the server from where the status of the person will be known at the same time if the alcohol is detected automatically the ignition will be off.

2. MODULES AND THEIR FUNCTIONALITIES

ALCOHOL GAS SENSOR MQ-3

Description: This alcohol sensor is suitable for detecting alcohol concentration on your breath, just like your common breathalyzer. It has a high sensitivity and fast response time. Sensor provides an analog resistive

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

output based on alcohol concentration. The drive circuit is very simple, all it needs is one resistor. A simple interface could be a 0-3.3V ADC.

GPS MODULE

The Global Positioning System (GPS) comprises three segments: The space segment (all functional satellites).The control segment (all ground stations involved in the monitoring of the system master control station, Monitor stations, and ground control stations).The user segment (all civil and military GPS users).

DC MOTOR

- The DC motor is an electric DC motor used to demonstrate the concept of engine locking.
- Here in this work, the DC motor will be connected to pin 9 on the microcontroller, when alcohol is detected the DC motor stops in other to indicate that alcohol is detected and continue running when there is no alcohol detected.

ALARM SECTION

The alarm unit used is a buzzer which indicates when alcohol is detected. The buzzer used belongs to the PS series. The PS series are high-performance buzzers that employ Unimorph piezoelectric elements and are designed for easy incorporation into various circuits. They have very low power consumption

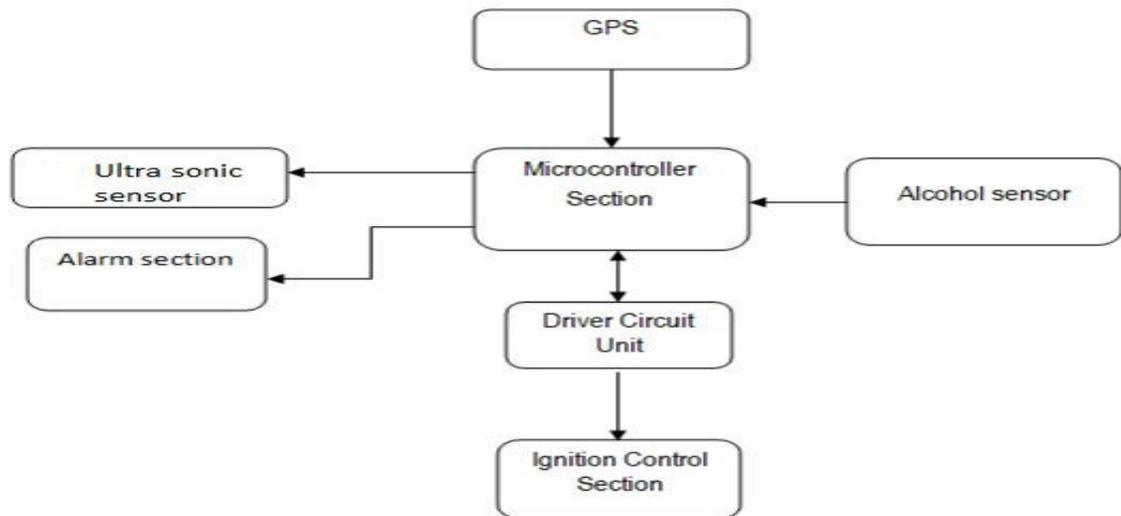
TEMPERATURE SENSOR

Temperature sensors are typically silicon based integrated circuits. Most contain the temperature sensor, an analog to digital converter (ADC), memory to temporarily store the temperature readings, and an interface that allows communication between the sensor and a microcontroller. Unlike analog temperature sensors, calculations are performed by the sensor, and the output is an actual temperature value.It used to sense the temperature of the car engine.

ULTRASONIC SENSOR

A basic ultrasonic sensor consists of one or more ultrasonic transmitters (basically speakers), a receiver, and a control circuit. The transmitters emit a high frequency ultrasonic sound, which bounce off any nearby solid objects. Some of that ultrasonic noise is reflected and detected by the receiver on the sensor. That return signal is then processed by the control circuit to calculate the time difference between the signal being transmitted and received. This time can subsequently be used, along with some clever math, to calculate the distance between the sensor and the reflecting object. It sense the distance between our vehicle to other front vehicle.

SYSTEM ARCHITECTURE/ BLOCK DIAGRAM.



In this MCP3008 is used, so connect 3.3v pin from raspberry to the sensor.

- Similarly MCP3008 and all sensor's ground pins should be grounded
- Now connect MQ3 Alcohol output pins to first channel of MCP3008 IC.
- Connect power supply for Raspberry pi
- Plug the HDMI cable in Raspberry pi from the monitor using VGA to HDMI converter cable
- Connect USB Mouse and USB keyboard to the Raspberry pi

TECHNOLOGIES

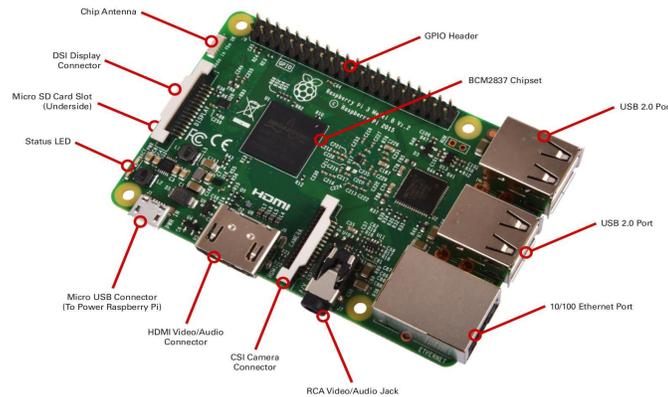
RASPBERRY PI

The Raspberry Pi foundation is working on yet another model of the popular Raspberry Pi boards, as the Raspberry Pi 3 model B board. The new board looks very similar to Raspberry Pi 2 model B, but adds on-board WiFi 802.11 b/g/n (2.4GHz only) and Bluetooth 4.0. Let's play "spot the difference" with Raspberry Pi 2 at the top and Raspberry Pi 3 under. We'll find the WiFi/BT chip antenna on the top left corner, and two through holes on the right of the 40-pin connectors, likely the RUN header for reset that can be found on the RPi2 where the chip antenna is now placed on RPi 3. So the through holes are not new, they've just moved it. All connectors have the exact same placement between the two versions.

The wireless module (likely Broadcom based) can be found just above the micro SD slot, and J5 connector is soldered. J5 is the JTAG connector, so it will probably not be soldered with the version that ships. The picture is not very clear but it looks like they've used the same Elpida B8132B4PB-8D-F RAM chip (1GB) as on Raspberry Pi 2. So although we can't be 100% certain right now, the RAM appears to be the same, and the processor is still connected to a similar USB to Ethernet chip, so they've probably kept the same architecture, expect possibly for the CPU core. So the only major

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

changes on Raspberry Pi 3 appears to be built-in Wi-Fi and Bluetooth, and 64-bit ARM cores (likely Cortex A53).



- SoC – Broadcom BCM2837 64bit ARMv8 quad core Cortex A53 processor @ 1.2GHz with dual core VideoCore IV GPU @ 400 MHz supporting OpenGL ES 2.0, hardware-accelerated OpenVG, and 1080p30 H.264 high-profile decode. Capable of 1Gpixel/s, 1.5Gtexel/s or 24GFLOPs with texture filtering and DMA infrastructure
 - System Memory – 1GB LPDDR2
 - Storage – micro SD slot
 - Video & Audio Output – HDMI 1.4 and 4-pole stereo audio and composite video port
 - Connectivity – 10/100M Ethernet, WiFi 802.11 b/g/n up to 150Mbps and Bluetooth 4.1 LE (BCM43438 module)
- USB – 4x USB 2.0 host ports (with better power management, allowing higher power peripherals), 1x micro USB port for power Expansion
 - 40-pin GPIO header
 - MIPI DSI for Raspberry Pi touch screen display
 - MIPI CSI for Raspberry Pi camera
 - Power Supply – 5V up to 2.4A via micro USB port
 - Dimensions – 85 x 56 x 17 mm

So the specifications were more or less as expected from the leaked information we've received before the official release, and people hoping for 4K or H.265 video decoding will be disappointed. The new processor is said to be 10 times faster than BCM2835 processor found in the first Raspberry Pi Model B board, and it's likely it can handle 1080p H.264 @ 60 fps using software decoding. The Video Core IV GPU's subsystem is now clocked at 400MHz and the 3D core at 300MHz against 250MHz for previous "Raspberry Pi processors".

Since they've basically kept the same features as Raspberry Pi 2, beside changing the Cortex A7 cores to 64-bit Cortex A53 ones, and adding built-in WiFi and Bluetooth via a BCM43438 module, firmware support is basically the same with various Linux distributions Raspbian being the recommended distro and Windows 10 IoT.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

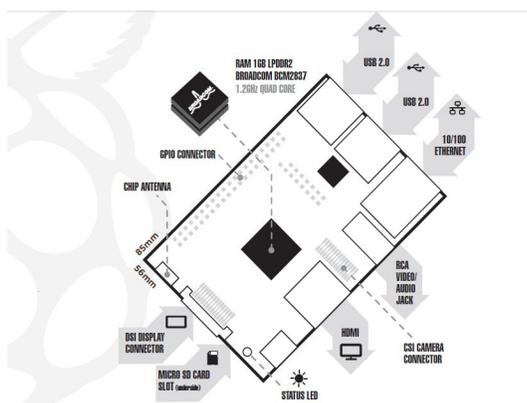
BCM2837:

For Raspberry Pi 3, Broadcom have supported us with a new SoC, BCM2837. This retains the same basic architecture as its predecessors BCM2835 and BCM2836, so all those projects and tutorials which rely on the precise details of the Raspberry Pi hardware will continue to work. The 900MHz 32-bit quad-core ARM Cortex-A7 CPU complex has been replaced by a custom-hardened 1.2GHz 64-bit quad-core ARM Cortex-A53. Combining a 33% increase in clock speed with various architectural enhancements, this provides a 50- 60% increase in performance in 32-bit mode versus Raspberry Pi 2, or roughly a factor of ten over the original Raspberry Pi.

Cortex-A53 Processor

The ARM® Cortex®-A53 processor offers a balance between performance and power-efficiency. Cortex-A53 is capable of seamlessly supporting 32-bit and 64-bit instruction sets. It makes use of a highly efficient 8-stage in-order pipeline enhanced with advanced fetch and data access techniques for performance. It fits in a power and area footprint suitable for entry-level smart phones. It also can deliver high aggregate performance in scalable enterprise systems via high core density, which accounts for its popularity in base station and networking designs.

ARM vs. x86



The processor at the heart of the Raspberry Pi system is a Broadcom BCM2837 system-on-chip (SoC) multimedia processor. This means that the vast majority of the system's components, including its central and graphics processing units along with the audio and communications hardware, are built onto that single component hidden beneath the 256 MB memory chip at the centre of the board.

It's not just this SoC design that makes the BCM2837 different to the processor found in your desktop or laptop, however. It also uses a different instruction set architecture (ISA), known as ARM. The BCM2837 SoC, located beneath a Hynix memory chip Developed by Acorn Computers back in the late 1980s, the ARM architecture is a relatively uncommon sight in the desktop world. Where it excels, however, is in mobile devices: the phone in your pocket almost certainly has at least one ARM-based processing core hidden away inside. Its combination of a simple reduced instruction set (RISC) architecture and low power draw make it the perfect choice over desktop

chips with high power demands and complex instruction set (CISC) architectures.

The ARM-based BCM2837 is the secret of how the Raspberry Pi is able to operate on just the 5V 1A power supply provided by the onboard micro-USB port. It's also the reason why you won't find any heat-sinks on the device: the chip's low power draw directly translates into very little waste heat, even during complicated processing tasks. It does, however, mean that the Raspberry Pi isn't compatible with traditional PC software. The majority of software for desktops and laptops is built with the x86 instruction set architecture in mind, as found in processors from the likes of AMD, Intel and VIA. As a result, it won't run on the ARM-based Raspberry Pi. The BCM2837 uses a generation of ARM's processor design known as ARM11, which in turn is designed around a version of the instruction set architecture known as ARMv6. This is worth remembering: ARMv6 is a lightweight and powerful architecture, but has a rival in the more advanced ARMv7 architecture used by the ARM Cortex family of processors. Software developed for ARMv7, like software developed for x86, is sadly not compatible with the Raspberry Pi's BCM2837—although developers can usually convert the software to make it suitable. That's not to say you're going to be restricted in your choices. As you'll discover later in the book, there is plenty of software available for the ARMv6 instruction set, and as the Raspberry Pi's popularity continues to grow, that will only increase. In this book, you'll also learn how to create your own software for the Pi even if you have no experience with programming.

Windows vs. Linux

Another important difference between the Raspberry Pi and your desktop or laptop, other than the size and price, is the operating system the software that allows you to control the computer. The majority of desktop and laptop computers available today run one of two operating systems Microsoft Windows or Apple OS X. Both platforms are closed source, created in a secretive environment using proprietary techniques. These operating systems are known as closed source for the nature of their source code, the computer-language recipe that tells the system what to do. In closed-source software, this recipe is kept a closely-guarded secret. Users are able to obtain the finished software, but never to see how it's made. The Raspberry Pi, by contrast, is designed to run an operating system called GNU/Linux— hereafter referred to simply as Linux. Unlike Windows or OS X, Linux is open source: it's possible to download the source code for the entire operating system and make whatever changes you desire. Nothing is hidden, and all changes are made in full view of the public. This open source development ethos has allowed Linux to be quickly altered to run on the Raspberry Pi, a process known as porting. At the time of this writing, several versions of Linux known as distributions have been ported to the Raspberry Pi's BCM2837 chip, including Debian, Fedora Remix and Arch Linux. The different distributions cater to different needs, but they all have something in common: they're all open source. They're also all, by and large, compatible with each other: software written on a Debian system will operate perfectly well on Arch Linux and vice versa.

Linux isn't exclusive to the Raspberry Pi. Hundreds of different distributions are available for desktops, laptops and even mobile devices; and Google's popular Android platform is developed on top of a Linux core. If you find that you enjoy the experience of using Linux on the Raspberry Pi, you could consider adding it to other computing devices you use as well. It will happily coexist with your current operating system, allowing you to enjoy the benefits of both while giving you a familiar environment when your Pi is unavailable. As with the difference between ARM and x86, there's a key point to make about the practical difference between Windows, OS X and Linux:

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

software written for Windows or OS X won't run on Linux. Thankfully, there are plenty of compatible alternatives for the overwhelming majority of common software products better still, the majority are free to use and as open source as the operating system itself.

Getting Started with the Raspberry Pi:

Now that you have a basic understanding of how the Pi differs from other computing devices, it's time to get started. If you've just received your Pi, take it out of its protective anti-static bag and place it on a flat, non-conductive surface before continuing with this chapter.

Connecting a Display:

Before you can start using your Raspberry Pi, you're going to need to connect a display. The Pi supports three different video outputs: composite video, HDMI video and DSI video. Composite video and HDMI video are readily accessible to the end user, as described in this section, while DSI video requires some specialized hardware.

Composite Video:

Composite video, available via the yellow-and-silver port at the top of the Pi known as an RCA phono connector is designed for connecting the Raspberry Pi to older display devices. As the name suggests, the connector creates a composite of the colours found within an image—red, green and blue—and sends it down a single wire to the display device, typically an old cathode-ray tube (CRT) TV. The yellow RCA phono connector, for composite video output When no other display device is available, a composite video connection will get you started with the Pi. The quality, however, isn't great. Composite video connections are significantly more prone to interference, lack clarity and run at a limited resolution, meaning that you can fit fewer icons and lines of text on the screen at once.

HDMI Video:

A better-quality picture can be obtained using the HDMI (High Definition Multimedia Interface) connector, the only port found on the bottom of the Pi. Unlike the analogue composite connection, the HDMI port provides a high-speed digital connection for pixel-perfect pictures on both computer monitors and high-definition TV sets. Using the HDMI port, a Pi can display images at the Full HD 1920x1080 resolution of most modern HDTV sets. At this resolution, significantly more detail is available on the screen. If you're hoping to use the Pi with an existing computer monitor, you may find that your display doesn't have an HDMI input. That's not a disaster: the digital signals present on the HDMI cable map to a common computer monitor standard called DVI (Digital Video Interconnect). By purchasing an HDMI-to-DVI cable, you'll be able to connect the Pi's HDMI port to a monitor with DVI-D connectivity. If your monitor has a VGA input a D-shaped connector with 15 pins, typically coloured silver and blue. The raspberry pi can't connect to it. Adapters are available that will take in a digital DVI signal and convert it to an analogue VGA signal, but these are expensive and bulky. The best option here is simply to buy a more-modern monitor with a DVI or HDMI input.

DSI Video:

The final video output on the Pi can be found above the SD card slot on the top of the printed circuit board—it's a small ribbon connector protected by a layer of plastic. This is for a video standard known as Display Serial Interface (DSI), which is used in the flat-panel displays of tablets

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

and smart phones. Displays with a DSI connector are rarely available for retail purchase, and are typically reserved for engineers looking to create a compact, self-contained system. A DSI display can be connected by inserting a ribbon cable into the matched connector on the Pi, but for beginners, the use of a composite or HDMI display is recommended.

Connecting Audio:

If you're using the Raspberry Pi's HDMI port, audio is simple: when properly configured, the HDMI port carries both the video signal and a digital audio signal. This means that you can connect a single cable to your display device to enjoy both sound and pictures. Assuming you're connecting the Pi to a standard HDMI display, there's very little to do at this point. For now, it's enough to simply connect the cable. If you're using the Pi with a DVI-D monitor via an adapter or cable, audio will not be included. This highlights the main difference between HDMI and DVI: while HDMI can carry audio signals, DVI cannot. For those with DVI-D monitors or those using the composite video output, a black 3.5 mm audio jack located on the top edge of the Pi next to the yellow phono connector provides analogue audio. This is the same connector used for headphones and microphones on consumer audio equipment and it's wired in exactly the same way. If you want, you can simply connect a pair of headphones to this port for quick access to audio. While headphones can be connected directly to the Raspberry Pi, you may find the volume a little lacking. If possible, connect a pair of powered speakers instead. The amplifier inside will help boost the signal to a more audible level. If you're looking for something more permanent, you can either use standard PC speakers that have a 3.5 mm connector or you can buy some adapter cables. For composite video users, a 3.5 mm to RCA phono cable is useful. This provides the two white and red RCA phono connections that sit alongside the video connection, each carrying a channel of the stereo audio signal to the TV. For those connecting the Pi to an amplifier or stereo system, you'll either need a 3.5 mm to RCA phono cable or a 3.5 mm to 3.5 mm cable, depending on what spare connections you have on your system. Both cable types are readily and cheaply available at consumer electronics shops, or can be purchased even cheaper at online retailers such as Amazon.

Connecting a Keyboard and Mouse

Now that you've got your Raspberry Pi's output devices sorted, it's time to think about input. As a bare minimum, you're going to need a keyboard, and for the majority of users, a mouse or trackball is a necessity too. First, some bad news: if you've got a keyboard and mouse with a PS/2 connector—a round plug with a horseshoe-shaped array of pins—then you're going to have to go out and buy a replacement. The old PS/2 connection has been superseded, and the Pi expects your peripherals to be connected over the Universal Serial Bus (USB) port. Depending on whether you purchased the Model A or Model B, you'll have either one or two USB ports available on the right side of the Pi (see Figure 1-4). If you're using Model B, you can connect the keyboard and mouse directly to these ports. If you're using Model A, you'll need to purchase a USB hub in order to connect two USB devices simultaneously. Figure 1-4: Model B's two USB ports A USB hub is a good investment for any Pi user: even if you've got a Model B, you'll use up both your available ports just connecting your keyboard and mouse, leaving nothing free for additional devices such as an external optical drive, storage device or joystick. Make sure you buy a powered USB hub: passive models are cheaper and smaller, but lack the ability to run current hungry devices like CD drives and external hard drives. As you've probably noticed, the Raspberry Pi doesn't have a traditional hard drive. Instead it uses a Secure Digital (SD) memory card, a solid-state storage system typically used in digital cameras. Almost any SD card will work with the Raspberry Pi, but

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

because it holds the entire operating system, it is necessary for the card to be at least 2 GB in capacity to store all the required files. SD cards with the operating system preloaded are available from the official Raspberry Pi Store along with numerous other sites on the Internet. If you've purchased one of these, or received it in a bundle with your Pi, you can simply plug it in to the SD card slot on the bottom side of the left-hand edge. If not, you'll need to install an operating system known as flashing onto the card before it's ready to go. Some SD cards work better than others, with some models refusing to work at all with the Raspberry Pi. For an up-to-date list of SD card models known to work with the Pi, visit the eLinux

Connecting External Storage

While the Raspberry Pi uses an SD card for its main storage device—known as a boot device you may find that you run into space limitations quite quickly. Although large SD cards holding 32 GB, 64 GB or more are available, they are often prohibitively expensive. Thankfully, there are devices that provide an additional hard drive to any computer when connected via a USB cable. Known as USB Mass Storage (UMS) devices, these can be physical hard drives, solid-state drives (SSDs) or even portable pocket-sized flash drives. Two USB Mass Storage devices: a pen drive and an external hard drive The majority of USB Mass Storage devices can be read by the Pi, whether or not they have existing content. In order for the Pi to be able to access these devices, their drives must be mounted a process “Linux System Administration”. For now, it's enough to connect the drives to the Pi in readiness.

Connecting the Network

While the majority of these setup instructions are equally applicable to both the Raspberry Pi Model A and the Model B, networking is a special exception. To keep the component count and therefore the cost as low as possible, the Model A doesn't feature any onboard networking. Thankfully, that doesn't mean you can't network the Model A; only that you'll need some additional equipment to do so.

Python programming language

For Windows users (and also those of you with Mac OS X) the tool of choice for writing a Raspberry Pi OS image to SD card is SD Formatter, from the SD Association. With the card inserted into your computer's card reader, and ensuring you have the correct Drive letter selected in the drop down menu, open the Option menu and select Full (Erase) and On. This ensures that the full capacity of the storage card will be available. Click OK, then Format to begin. To write the disk image, use Win32DiskImager, available from Source forge. You may need to run with administrator privileges. Select the correct drive letter for your SD card, browse to the image file and click Write to commence the process. Win32DiskImager will inform you when the data has been written. If writing the disk image seems too much hassle or is beyond your abilities, it is possible to purchase SD cards with Raspbian pre-installed. Booting Raspbian for the First Time with Raspbian installed, you'll need to login with the following credentials: Username: pi Password: raspberry.

CONCLUSION

In this we proposed a method to sense the presence of alcohol from the breath of drivers and curtail the catastrophic effects it can have on peoples' lives. The system was designed and

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

implemented successfully via the use of Raspberry pi microcontroller and MQ- 3 sensor. Experimental evaluation of the system showed that the alcohol sensor was able to deliver fast response when alcohol is detected. Also, the ability of the alcohol sensor to operate over a long time is a feature of this system.

Future enhancement:

In Future work, Government must authorize laws to introduce such circuit in each car and must manage all car organizations to preinstall such systems while manufacturing the car itself. If it is achieved the death rate because of drunken drivers can be brought to least level.

In this kind of system, securely landing of car aside without disturbing other vehicles can also be added as a future extension

In future, even we can add seat belt feature in this project so that if the driver forgets to keep seat belt it indicates to driver and engine won't start until the seat belt is kept

With this system the driver becomes more alert and he stops drinking on that day if he or she want to drive a car.

REFERENCES

[1] Mandalkar Rahul B,Pandore Rahul N,Shinde Manoj B,Godse Valmik D, “Alcohol Detection and Accident Avoidance Using Locking With Tracking”

[2] Abhi R. Varma, Seema V. Arote, Chetna Bharti, Kuldeep Singh, “Accident Prevention Using Eye Blinking and Head Movement”.

[3] Dhivya M and Kathiravan S, “Driver Authentication and Accident Avoidance System for Vehicles”

[4] Road accidents in India [online] 2007 June 25.
Available from: URL: <http://www.easydriveforum.com/f44-share-your-road-experience/road-accidents-in-india-834.html>

[5] Killoran, A., Canning, U., Doyle, N., & Sheppard, L, “Review of effectiveness of laws limiting blood alcohol concentration levels to reduce alcohol-related road injuries and deaths” Final Report. London: Centre for Public Health Excellence (NICE), 2010.

[6] Adamson, S., & Enright, S. Alcohol Gas Detector “Breathalyzer”.

E-Commerce: An Analysis of present Trends, Challenges & Opportunities

Jogannagari Malla Reddy
Prof. & Head, Dept. of CSE
Indur Institute of Engineering & Technology
Siddipet, Telangana State, India
Email: jmrsdpt06@gmail.com

T. R. Srinivas
Prof. & Head, Dept. of CSE
AAR Mahaveer Engineering College,
Hyderabad, Telangana State, India
Email: stelkar@gmail.com

Abstract: The concept of e-commerce started in 1960 with development of Electronic Data Interchange (EDI) for transferring the documents. The technology continues to grow rapidly, in 1990 the online business expanded with new features. E-commerce is performs the vital role in the business phenomena and facilitate to explore in the future. E-commerce is the paradigm provides the trading in the business world. It provides the extraordinary features in the business growth of the country economy. The customers trading the products using the mobile phone applications and Internet of Things. The business sector is grown up to the greater heights with expansion of e-commerce. However, it may contains come critical challenges apart from the opportunities. This paper highlights the various software testing challenges, difficulties, strategies and techniques in perspective of Component Based Software Development

Key words: Modularity, Reusability, Component Based Software Development, Black box testing, White box testing, Regression Testing.

1. INTRODUCTION

The invention of Internet technologies the huge revolution in the business. Internet reshaped the information systems in the business area and make into globalized. The

utility of information systems are expended and new business models have become possible only because of the internet. Internet eliminated technical, geographic and business, cost barriers with global flow of information. There is instrumental growth in business sector with invention of internet make possible of information exchange between buyers and sellers in the form of e-commerce. Tremendous change happened in the business sector with the concept of e-Commerce.

Initially the e-Commerce started the advertisement publishing of major organizations in 1995 by Netscape.com through its web portal. The phenomenal growth of e-Commerce is much more than the other technological inventions such as telephone, radio and television. The major setback in the e-Commerce movement with “Dot.com” companies in 2001. As the outcome most of the dot-com companies failed. They learned a lot about the practical limitations of e-Commerce. The few of the companies like Amazon, Google, e-bay not only survived and continued e-commerce movement with better growth. E-commerce achieved the significance in 2006 with the retail market in the developed countries of US, Asia and Europe.

Today, good number of people offered the services such as buying and selling products through e-Commerce. As per E-Commerce times report, The Amazon got 28% of sales increase in 2003, due to the more than 20% Americans access the online shopping from their homes which are connected with broadband services. Year 2010 onwards development in of the e-commerce reached into people hands with smart phones. With Google provides the services at their fingertips, shopper will now turn to the internet to refer reviews and online offers and browsing websites at home. Social media has also changed how consumers shop their chosen ecommerce retailers. Social platforms have made brands much more accessible for their customers and has changed the way they communicate with business.

This paper focuses the technical aspects of e-commerce system with innovative direction. The rest of paper is organized as follows. The Section 2 explains the review of literature. Section 3 explains the taxonomy and overview of E-Commerce spectrum. The Section 4 states the various phases of the trade life cycle model. The Opportunities and

Challenges of E-Commerce discussed in the Section 5. Finally, concluded with discussion in section 6.

2. REVIEW OF LITERATURE

Over the year many researchers and theorist identified the vile role of information technology in the commerce. The e-commerce in the early years is dominated with idea both in terms of product and process. A numbers of theorists and researchers have worked on various issues in the domain of e-commerce applications. The developed taxonomy can be benefit to extend the knowledge of ecommerce application for discussion, setting up and evaluating the more quality service.

- Dr. P. Devaraju[1] discussed the various Challenges and opportunities of E-commerce in the India. The author visualized the demographic figures with statistical graphs
- Dr. Subhash Masanappa Suryawamshi. [2] focused on the recent trends of e-commerce industry in India . In his study undertaken to analyze the present trends and examine the challenges, opportunities of e-commerce in India.
- DP Goyal [3] differentiated the e-commerce and e-business. The author identified the advantages and challenges of e-commerce application. The key components infrastructure and its functionality is more understandable for e-commerce users. He highlighted the Business opportunities and challenges of ecommerce in taxonomy.
- Abdul Gaffer [4] Khan discussed various benefits of e-commerce and competitive advantage over the other competitors. This study predicts some challenges in an emerging economy.
- Dr. Rajeshwari et.al [5], highlighted the different challenges faced by the Indian ecommerce industry in the empirical study. The study found that, the e-commerce provides the various opportunities to the retailers, wholesalers and people.
- Shahriari et al [6], discussed the various benefits of ecommerce and its impact on the market. The author classified the various benefits on user perspective of e-commerce.

3. THE TAXONOMY AND OVERVIEW OF E-COMMERCE SPECTRUM

The term of e-commerce has become popular because of it is one of the brain child of information technology business application. E-commerce is is the process. E-commerce

is the process of buying the selling the products and services electronically. Another term e-Business is similar to e-commerce, whereas e-business is broader and both are interchangeably. The e-business includes e-commerce with front and back office applications.

Internet has been the major driven force for the wide expansion of computer applications in the business area. Many Commercial organization benefited with e-commerce making the transactions electronically. Today, many companies irrespective of size using the e-commerce applications in some way. Worldwide the many stakeholders gain the access of commerce applications which provides the feature such as universal standards, customization, global reach and social networking for the rapid growth of e-Commerce.

The other part of trading products through e-commerce, it encompasses the other activities such as developing, marketing, selling, delivering, paying and servicing the products and services through internet service. The e-Commerce broadly includes the following functions.

- Provide the product description though catalog.
- Defining the customer requirement through the search option.
- Perform the purchase transition though electronic payment systems.
- Delivery of the products with various methods such as couriers, road, air etc.
- Provide the Customer service after the sales.

Apart from the above, ecommerce reduces the transaction costs, fast flow of information, improved customer service and proper coordination between stakeholders such as manufacturers, suppliers and customers.

A. *Classification of e-Commerce*

- a. Business to Business (B2B): This type of e-commerce application between the commercial organizations. All the participants are organizations ex: Neoforma.com

- b. Business to Consumer (B2C): The B2C e-commerce deals the transaction in between the organization and customers directly without involvement of mediators. Ex: Amazon, Flipchart and indiatimes.com. These portal trade the products directly to the customers.
- c. Consumer to Consumer (C2C): It provides the trade transactions between customers to customer. Ex: ebay.com.
- d. Business to Government (B2G): The B2G e-Commerce is generally defined transactions between the Companies and the public sector. The marketing products and services to various government levels.

B. Area of Applicability of E-Commerce:

The e-commerce application used in the businesses such as retail and wholesale business, finance, manufacturing and marketing etc.

Retail and Wholesale business: Most number of c-commerce applications are in retail and wholesale in the online mode. The electronic storefronts sale the products directly to consumers. The e-commerce software facilitates selling, cataloging and shipping the products to the consumer site. The cybermall provides virtual space for multiple buyers and sellers through web browser. E-Commerce sites are used in the wholesale trading of products by various companies.

Finance: Most of financial institutions are using the e-commerce applications for smooth functioning of financial services. E-Commerce of finance provides the various operations to customers such as depositing, withdrawal, transfer the money to other accounts, order for checkbooks, demand drafts, pay the bills through e-banking etc. Stack trading is another part of the e-Commerce application which provide the news, analytical Charts, company profiles and analysis of trading at the online.

Manufacturing: Supply chain operations of the company provided through e-commerce applications. Some companies can form an electronic exchange of trading of market

information and run back office operations such as inventory control [8]. This type of applications can speed up the flow of raw material and finished products among the business community which reduces the inventory cost.

Marketing: E-Commerce applications used in the area of marketing for the collection of customer information relevant with their behavior, preferences, buying patterns, needs through the web [7]. This information used in marketing for price fixation, product enhancement, negotiations and promotions.

Bidding and Auctions: C2C is direct selling the products among the customers through electronic auctions. Bidding allows buyers to place a bid for a product or service. Ex: quote the price for seat booking in airline & premier tatkal of train

C. Infrastructure of E-Commerce

The E-commerce infrastructure classified into two types such as Hardware and Software.

Hardware: E-Commerce hardware relevant with web server configuration like storage capacity and processor computing capability. The configuration required based on the used software as well as e-Commerce transactions performed. Sometimes the companies acquired the services from the third parties on lease base. There must be adequate hardware backup to avoid the transaction failures and hazards.

Software: The e-Commerce software classified into following parts:

.

Web-server Software: In server software. The Web enabled operating system, which provides the functions like security, retrieval of web pages, sending of web pages, tracking, web site development and webpage development.

E-Commerce Software: It is host software of e-Commerce. It contains various process about the e-commerce product transactions. Ex: Catalog Management, Product Configuration, Transport shipping Cart, Transaction Processing and Web traffic Data Analysis.

D. E-Commerce Payment Systems & Security

E-Commerce facing the main threat with electronic payment systems through the internet. The computer criminal capture the data about credit information through online. The consumers become suspicious about money transitions. Today e-Commerce provides the enhanced security provisions to electronic payment systems over the web. But, clearly understood that there is no absolute security on the internet. The e-Commerce applications provides various security mechanisms such as user identification, password, encryption technologies and digital certifications. Internet provides another security mechanism by Secure Socket Layer (SSL) communication protocol to safeguard the payment transactions. This protocol work above the TCP layer of OSI model and other protocols like Telnet and HTTP. Some of the Electronic payment is as follows.

Electronic Cash: Electronic cash is like hard cash that can be used for online payments. The financial institutions provide the net banking facility to the customers for online payments.

Electronic Wallets: Electronic Wallet is computerized stored value that holds the credit card information. It is most convenient approach to purchase the products at online.

Cards: Cards are more flexible for the online buyers. The credit card, such as visa, master card has predefined spending limit. Debit card is another form of payment on the internet [9]. Debit cards work as cash or personal cheque, these cards have magnetic strip to withdraw the amount by swapping process.

4. THE VARIOUS PHASES OF TRADE LIFE CYCLE MODEL

E-Commerce Trade life Cycle model have the various phases. The e-commerce system supports each of the phase. The phases are described as following

Searching for the item: customer will search for the required product at supplier's home page. The customer verify the product catalog for description and search the required product.

Product Selection and Negotiation: After the searching required product, the customer fulfill the quotation formats with entering product code and no. of items required. After the quotation is received, the consumer examines them and selects the items by clicking on the request for quotation form.

Product Purchasing: The customer complete the purchase order by sending a completed electronic form to the supplier. In this phase the customer can choose the mode of payment [10]. The various security procedure can be incorporated on the internet for safeguard the money transaction.

Product Delivery: The logical products such as software and multimedia products can be downloaded through the internet after the payment. T However, the physical products cannot be delivered that be possible with traditional methods like road, air and courier. Product delivery either by company or it may be outsourced with third party.

After the Sales Service: This phase belong to service and maintenance of the product relevant with product usage, repair service under warranty can be obtained from the websites.

5. BUSINESS OPPORTUNITIES AND CHALLENGES OF E-COMMERCE

Internet Technology has provided significant opportunities for the business area. Apart from the opportunities there will be certain challenges and threats to the e-Commerce.

Upcoming Business Models: The e-Commerce can motivate the new business models with innovative features.

Channel Conflicts: The sales force and distribution may feel the loss and fear about their financial benefits a result of direct buying the product by the buyers.

Security and Privacy: Security and Privacy is another major challenge for the e-commerce. There is choice of threat from the intruders on customer transactions relevant with security and confidentiality of credit card information.

Reorganization of Business Process: In order to implement e-Commerce applications, business firms required to redesign the business processes and functional scope. The Business firms are defined with well-defined policies and procedures in transparent manner for sharing the data with other business firms.

Legal Problems for e-Commerce: Biggest challenge for e-Commerce is the handling of legal issues relevant with email contract, the role of electronic signatures. Copyright laws etc. The internet is wide area network which connect the heterogeneous countries with different legal systems. It is a complex situation in terms of its legal implications.

Managerial Opportunities: The e-commerce provides many managerial opportunities such as to reduction of transaction costs; the customers and suppliers can exchange business communication without involvement of intermediaries and proper communication and coordination established in between the business communities.

6. CONCLUSIONS

The E-Commerce reduced the gap in between manufacturer and consumer with the innovation of ecommerce applications. One of the main threat to E-commerce, that the computer criminals capture the crucial data of payment transaction system. The intruders escape from the crimes with weakness of cyber laws. The role of constitution is to frame the proper legal framework for e-commerce applications to protect the basic rights such as intellectual property, privacy, and consumer protection etc. There is need of extensive research in the area security and privacy of e-commerce. On other hand the government should frame the legal system and enforce the law and order across the cyber criminals.

References:

- [1] Dr. P. Devaraju, "The Challenges and Opportunities of E-Commerce in India: Future Prospective", International Journal of Engineering and Computer Science, Vol.5, Issue 11, Nov, 2016.
- [2] Dr. Subhash Masanappa Suryawanshi, "E-Commerce in India – Challenges and Opportunities E-Commerce", International Research Journal of Multidisciplinary Studies, Vol. 3, Issue 3, March, 2017.
- [3] D P Goyal, "Management Information Systems – Managerial Perspectives", Macmillan Publishers- 3rd Edition, 2010.
- [4] Abdul Gaffar Khan, "Electronic Commerce: Study on Benefits and Challenges in an Emerging Economy", Global Journal of Management and Business Research: Economics and Commerce, Vol.16, Issue. 1, 2016.
- [5] Dr. Rajeshwari and M. Shettar, "Emerging Trends of E-Commerce in India: An Empirical Study", International Journal of Business and Management Invention, Vol.5, Issue. 9, Sept. 2016.
- [6] Shahriari et al, " E-Commerce and it Impacts on Global Trend and Market", International Journal of Research - Granthaalayah, Vol.3, Issue.4, April,2015
- [7] Dr. Akshat Dubey et al, " E-Commerce Study of Privacy Trust and Security from Consumer's Perspective", International Journal of Computer Science and Mobile Computing, Vol.5, Issue.6, June, 2016
- [8] Dr. Naveen Kumar, "E-Commerce in India: An Analysis of Present Status, Challenges and Opportunities", International Journal of

- [9] Management Studies, Vol.5, Issue.2, April, 2018.
Kumar Anuj et al, "Impact of E-Commerce in Indian Economy", IOSR
Journal of Business and Management, Vol.20, Issue.5, May, 2018
- [10] Pratima Bhalekar et al, " The Study of E-Commerce", Asian Journal
of Computer Science

Internet of Things – Applications and Hurdles

T. R. Srinivas

Head, Department of Computer Science and Engineering
AAR Mahaveer Engineering College

Abstract

Internet of Things (IoT) refers to physical and virtual objects that have unique identities and are connected to the internet to facilitate intelligent applications that make energy, logistics, industrial control, retail, agriculture and many other domains “smarter”. Continuous progress in microelectronics and network techniques now make it possible to anticipate networks. The interconnection of objects with advanced processing and connection possibilities will lead to a revolution in terms of creation and availability of services. In short, the physical world will merge with the digital world. The main contribution of this paper is to summarize uses of internet of things (IoT) in different domain and highlight the main hurdles in the usage of internet of things.

Keywords: Internet, Internet of Things, sensors, microelectronics.

1. INTRODUCTION

Internet of Things is a new revolution of the internet that is rapidly gathering momentum driven by the advancements in sensor networks, mobile devices, wireless communications, networking and cloud technologies. Experts forecast that by the year 2020 there will be a total of 50 billion devices/things connected to the internet [1].

While many existing devices, such as networked computers or 4G enabled mobile phones already have some form of unique identities and are also connected to the internet, the focus on IoT is in the configuration, control and networking via internet of devices or “things” that are traditionally not associated with the internet. These include devices such as thermostats, utility meters, a bluetooth-connected headset, irrigation pumps and sensors or control circuits for an electric cars engine. The majority industry players are excited by the prospects of new markets for their products. The products include hardware and software components for IoT endpoints, hubs or control centers of the IoT universe.

The scope of IoT is not limited to just connecting things (devices, appliances, machines) to the internet. IoT allows these things to communicate and exchange data (control and information that could include data associated with users) while executing meaningful application towards a common user or machine goal.

This paper focuses on the different applications domains of IoT and hurdles in the usage of Internet of Things. The rest of paper is organized as follows. Section 2 provides an overview of IoT applications. Section 3 explains the use of IoT in retail. The Section 4 states the usage of IoT

in insurance. The hurdles in the use of IoT is discussed in the Section 5. Section 6 deals with the Technology in the development of IoT and Section 7 highlights the vision of IoT. Finally, the paper completes with conclusion in section 8 and references in section 9.

2. IoT APPLICATIONS

Application on IoT networks extract and create information from lower level data by filtering, processing, categorizing, condensing and contextualizing the data. This information obtained is then organized and structured to infer knowledge about the system and/or its users, its environment and its operations and progress towards its objectives, allowing a smarter performance.

The applications of Internet of Things (see Figure 1.1) span a wide range of domains including (but not limited to) homes, cities, environment, energy systems, retail, logistics, industry, agriculture and health as listed in Table 1.1.

Domain	Description
Homes	IoT has several applications such as smart lighting that adapt the lighting to suit the ambient conditions, smart appliances that can be remotely monitored and controlled, intrusion detection systems, smart smoke detectors, etc.
Cities	IoT has applications such as smart parking systems that provide status update on available slots, smart lighting that helps in saving energy, smart roads that provide information on driving conditions and structural health monitoring systems
Environment	IoT has applications such as weather monitoring, air and noise pollution, forest fire detection and river flood detection system.
Energy systems	IoT has applications such as including smart grids, grid integration of renewable energy sources and prognostic health management systems
Retail domain	IoT has applications such as inventory management, smart payments and smart vending machines.
Agriculture domain	IoT has application such as smart irrigation systems that help in saving water while enhancing productivity and green house control systems.
Industrial applications	IoT include machine diagnostics and prognosis systems that help in predicting faults and determining the cause of faults and indoor air quality systems.
Health and lifestyle	IoT has application such as health and fitness monitoring systems and wearable electronics

Table 1.1 IoT application domains

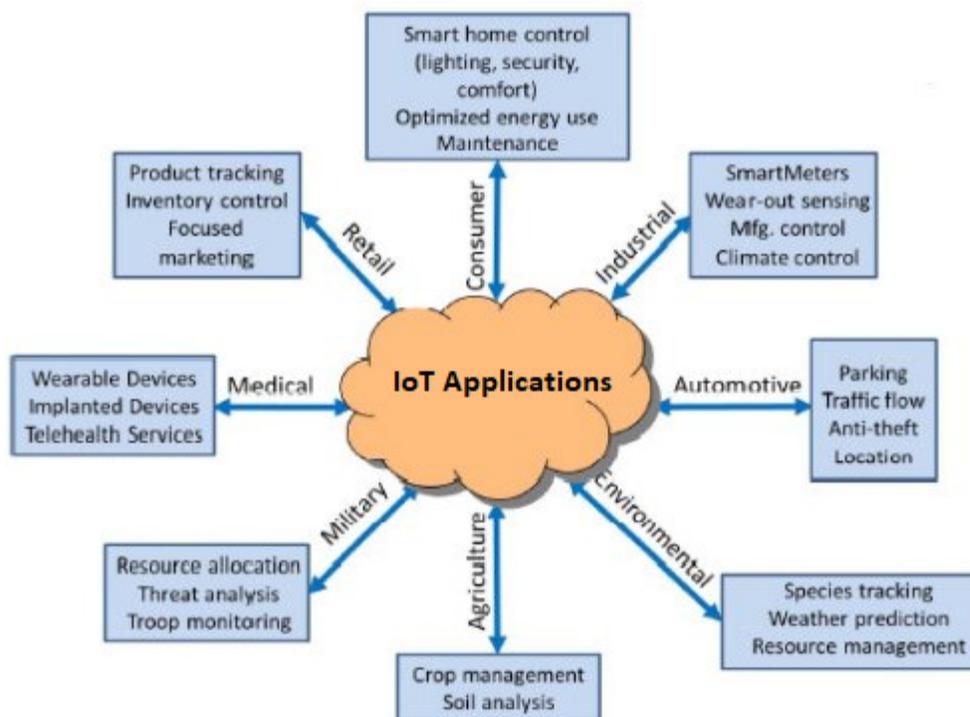


Figure 1.1 Application of Internet of Things (IoT)

Currently it is the human interaction with computers, in the IoT scenario, the machines will interact directly with the machines. According to Garter, "Internet of everything = Internet of information + Internet of people + Internet of things + Internet of places. The greatest innovation the current generation is taking full advantage is smartphones and the Internet. Internet has really made the world a global village. Many of the best services are turning the social network into a place where the flow of information is instantaneous and the Internet is also helping government agencies and the entire manufacturing sector do business.

The real time data helps the planners and strategists to make necessary changes in real time getting the best results. Internet of things of Internet of everything is one innovation which is extension of internet from humans to devices [2]. IoT is still in its infancy and all major telecom and IT companies are putting money and manpower to make it happen in most economical way. The emerging technologies like RFID's will help development of IoT in smoother way (see Table 1.2). For meeting growing requirement of addressable objects migration from IPV4 to IPV6 will go a long way. GPS, AI is also helping in the development of IoT.

	2010	2020
Internet users	1.7 billions	5 billions

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

IP addresses	4 millions	6 billions
Average speed	1.7 Mbps	10 Mbps
% of connections	21%	50%

Table 1.2 Global state of Internet

Internet of Things is referred as “Extending the current internet and providing connection, communication and inter-networking between devices and physical objects”.

The technologies and solutions that enable integration of real world data and services with current networking technologies are often described under the umbrella term Internet of Things (IoT).

In 1999, four universities including MIT encapsulated the concept of IoT. These universities started the development of sensors. In 2010, the number of devices connected on internet was more than human population [3]. World population is about 6.8 billions and the number of devices connected is 12.5 billion that works out to 1.84 per person (Table 1.3). There are about 2.5 billion smart phones. It is presumed that all persons have internet, but it was only those having internet this ratio comes to 6.25. By 2020, it is presumed that number of devices will be around 50 billion, which gives a ratio of about 7 per person, which translates to about 20 devices per internet user.

	2010	2015	2020
World Population	6.8 billions	7.2 billions	7.6 billions
Devices connected	12.5 millions	25 billions	50 billions
Devices per person	1.84	3.47	6.58

Table 1.3 Internet usage and world population

This is the power of Internet which doubles every 5.32 years. The data generated by things connected on Internet may be small in volume but the number of devices being in billions, the volume of data generated will be enormous due to cumulative effect.

Internet has already affected our lives. Health, Education, Businesses, Government doing e-governance are all into exploiting the use of internet. Internet is the most powerful creation of the human history. Internet is glue to IoT. Huge amount of data is being generated on Internet and this becomes knowledge.

Applications of IoT in real world are (see Figure 1.2):

- Aerospace and Aviation (systems status monitoring, green operations)
- Telecommunications
- Intelligent Buildings (automatic energy metering / home automation / wireless monitoring)
- Medical technologies, Health care
- Independent living (wellness, mobility, monitoring of an aging population)
- Pharmaceuticals
- Manufacturing, Product life cycle (from cradle to grave)
- Retail, Logistics, Supply chain Management
- Processing industries (oil and gas)
- Safety, Security and Privacy
- Environment Monitoring
- People and goods transportation
- Food traceability
- Agriculture and Breeding
- Media, environment and ticketing
- Insurance

It is estimated that by 2032, each person on earth will be connected to about 2000 devices [4].

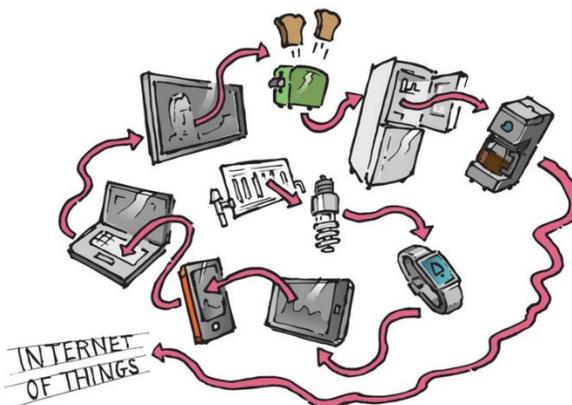


Figure 1.2 Internet of things devices

3. USE OF IoT IN RETAIL

Retail was digitally transformed when it went online. Today it is seeing another disruption on its horizon – Internet of Things (IoT). The writing on the wall is clear for retailers: *Adapt or perish!* Social, Mobility, cloud and analytics have already ushered in a revolution in the way retailers conduct their business. In conjunction with IoT, these technologies are going to unlock new business potential for retail businesses. And that's why they are excited. According to a recent

paper published, IoT will change the retail experience for the business owner and the customer in the following ways: **Connected shopping experience:** customers today rely on social media for a lot of things and that includes product reviews. In an IoT enabled store, they can scan the product using an appropriate device and look up for relevant reviews and ratings from their social network or online store network, this would help them take more informed decisions. It also means fewer returns and higher customer satisfaction for the retail business.

Delighting customers in real time: When customers walk into a store and make a purchase, the IoT activates digital coupons that they can use to save on their transactions. There is nothing new about offers and coupons, except that with IoT, they are given in real time. For instance, when the customer purchases for the third consecutive time, the store can activate a coupon for them right during their shopping.

Browsing assistance: Customers can save time during their retail walk by browsing for a desired product on a touchscreen monitor. They can also setup alerts when a certain product is out of stock and want to purchase it the moment it is back in the store's inventory.

Inventory Management: Smart store shelves can detect the shortage of a certain product or detect that they have attained their shelf-life. Either way, this can be used to alert their supplier, which gets enough time to supply the next batch of stock items. Artificially intelligent systems like robots can be roped in to populate the shelf stores. This system allows the business owner to reduce wastage while seeing that customer demand is always met in time. This increases customer loyalty.

Pricing: The pricing on products can be changed in real time detecting movements in supply and demand. This will ensure that the retail business runs profitably, while offering the benefit of potentially lowered prices for the customer.

Store operations: IoT enabled store will be able to reduce spend on lighting and air-conditioning. IoT enabled cameras, sensors and beacons can deliver real-time store data such as customer buying behaviour which is used to build customer history. This can provide useful insights.

Amazon has pioneered in bleeding-edge technology powered stores such as its Amazon Go. The technology that powers Amazon Go is the same technology that powers its self-driving cars. While not every retail store may come up with their own version of Amazon Go, they lie somewhere on the spectrum of technology adoption. What's a cool thing to have may soon become the only smart thing to have.

4. USE OF IoT IN INSURANCE

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

As per Business Insider intelligence, the total number of IoT devices is set to explode to 9 billion all over the world. They have the power to influence every facet of our life and insurance is no exception. Technology has created a new trend in insurance, namely *usage based insurance*.

IoT, key in Insure Tech: Wikipedia says that the Internet of Things (IoT) is the network of physical devices, vehicles and other items embedded with electronics, software, sensors, actuators and network connectivity which enable these objects to collect and exchange data. Notably, India is set to see the world's biggest IoT network in the coming years. According to Fin Tech weekly, InsureTech is the technology & platform that helps optimize any of the principles for success or requirements of insurance. IoT is one such technology.

IoT helping the Insurance Providers: IoT network is so closely tied to the habits, life style choices and behaviour patterns of the policy holder that it leaves nothing to imagination. For instance, using the inputs from the vehicle the customer uses, the vehicle insurance company can adjust the premium amount, rewarding the good driver.

Similarly, a policy holder who takes conscious steps to protect his health by say, regularly exercising will be rewarded with a discount on premium to be paid. This will reduce the customer's costs and improve the perception on the insurance company. Welcome to usage based insurance! Also, insurance companies can use the inputs from the IoT to send timely alerts to a customers and thus personalize their customer's journey.

IoT helping policy holders: Customers doesn't have to put a lot of effort to get his claim processed. Data will do all the taking, especially because it is instant and no time is lost. This directly creates a positive branding experience for the insurance company. Insurance companies have been known conventionally to be distant, aloof and impersonal. Thanks to technology, that is about to change. Customers will get timely and relevant advice during calamities which they can use to minimize the harm done, something they would appreciate a lot.

When combined with data analytics and Artificial Intelligence (AI), Internet of Things can become extremely useful to the policy holder as well as the insurance company. An insurer who deploys these technologies will have a clear edge over the insurer who doesn't. With the sweeping changes being made to the industry, it is question of when and not a question of if these technologies will prove to usher in the next wave of growth.

5. HURDLES IN IoT USE

According to Gartner, by 2020, there would be 26 billion devices connected to the IoT world over, exploding the market size to a massive \$7 trillion. India is expected to grow into a market of \$15 billion by the same year. Though that is the expectation, the reality is less rosier, with India

having to cross many hurdles to become a significant world player. India must overcome the following challenges in IoT adoption to be able to make good on IoT opportunities.

Security: As more people and companies get on the IoT network, there would be a significant amount of critical data passing from one device to another. Already, SMBs in India are facing a serious security threat to their assets. With more connectedness comes an even more serious security threat. In fact according to a recent survey, 70% of the devices can be hacked without any effort due to poor encryption and backdoor loopholes. India could be in a big soup if it exposes its systems to an ever growing network of cyber criminals.

Lack of understanding: though there is a decent amount of funding going the IoT way, it is still by and large an urban phenomenon. What's more – it is still considered a new-fangled technology setup that is not backed by enough use cases in the industry. It is a classic chicken-or-egg conundrum where it takes enough practical use to spur even more take up by the industry. Since adoption of IoT requires high infrastructural investments and the ROI is hazy at best, it requires a leap of faith.

Lack of uniform standards: Standards are lacking in data, security, integration and wireless protocols. It will be a key deterrent as IoT systems mainly target environments with a heterogeneous mix of technologies. It will take some time for a de-facto industry to emerge and it is not clear if that would happen by 2020. Meanwhile, conflicting standards can actually inhibit innovation as resources would have to be diverted from product innovation to managing unanticipated and varied compliance requirements.

Low business justification: At present, we see that there are concerns related to demand and supply economics of IoT. For Original Equipment Manufacturers (OEMs), the business justification is improvement in equipment design and a corresponding increase in market share, followed by increase in revenues for services after sale. For end users, the business justification is in avoiding unplanned downtime and revenue losses due to equipment related issues. For all other sectors, such justification is not clear.

Limited Internet connectivity: the internet is the back bone on which India's IoT will be rolled out and the IoT aspiring businesses are no exception. Already, there is a huge urban-rural divide in the availability of functional internet. Since the vision of a 15 billion USD market includes the rural applications of IoT, it is time the government encourages infrastructural investments in rural India. The aim to connect India's 2.5 lakh village panchayats digitally is a step in the right direction but is yet to come to function. As digital India's goals keep getting pushed down the years, Indian IoT dream would also remain unfulfilled.

Lack of skilled workforce: According to a recent Labour Bureau Report, skilled workforce in India is only 2% while the trainable or skilled work force over 15 years old is a sad 6.8%. IoT being a new technology, the gap is likely to be even more. Since a successful implementation of

IoT systems involves cutting edge expertise is big data, cloud and security, the government must lose no time in channelizing its effort into education that delivers. What's the use if churning out engineers year after year without seeing that they are also going to be gainfully employed? It makes sense for the government to alter the curriculum and drive its vision across educational institutions. Without this, IoT vision will just be another pipe-dream.

India has missed the manufacturing bus in the 20th century. Today, it is on the brink of an IT revolution. Companies of all sizes can utilize the IoT networks to deliver new and useful services, improve customer engagement and create new streams of revenue for themselves. To achieve that, they would have to leverage on the synergies between domain experts and IoT experts which means, expertise is key even to make a leap of faith. This, more than the other challenges seems to be the hardest nut to crack.

6. IoT TECHNOLOGY

The main technologies in the development of IoT are:

- Devices ID - Development of RFID (Radio frequency Identification)
- IP Address - By changing over to IPv6
- Network, WiFi enabled connectivity
- Powering the devices which will need Nano-generators
- Agreement of all the developers for standardization (open platform)

The amount of data will be huge and we will have to device new technologies to carry this [5]. Already 5G is being developed to carry gigabytes of data. Latency of data will improve with 5G. The technology for IoT will have to meet the following needs.

- Sensing and data collection capability (sensing nodes)
- Layers of local embedded processing capability (local embedded processing nodes)
- Wires and/or wireless communication capability (connectivity nodes)
- Software to automate tasks and enable new classes of services
- Remote network/cloud-based embedded processing capability (remote embedded procession nodes)
- Full security across the signal path

The 3C's of IoT

- **Communication** : Necessary for all devices connected whether stationary or moving, we will need GPS enabled devices.
- **Control and Animation**: we may need to control devices remotely like AC, Lights etc., even today many DTH operators like TATA Sky enable control of programs on the mobile phones which can be connected from anywhere in the world. Today many IoT companies are working with electronics goods manufacturers like Samsung, LG to enable controls from smart phones by putting API's etc.

- **Cost Savings:** By collecting data of the Plant and Machinery, the companies will get performance data of all the machines working in a process thus economizing on their operation costs by analysing the performance data.

We need to address the following four issues.

- **Define sensors** – what output we need from these sensors to help industry in the developments. Many RFID companies are developing these sensors.
- **Build IoT network** and security Foundation – we need to work with open standards as the devices are interconnected there are greater security concerns.
- **Collect as much data as possible** – this is very important because if we keep on collecting raw data without any plan, we may end up in big confusion and management of data will become very difficult. There will be then Tsunami of data.
- **Review and scale** of IoT providers – there are going to be four categories in this 1) sensor and RFID providers 2) M->M device management platforms 3) Solution delivery platforms (SDP) 4) Apps that will enable IoT devices to respond.

7. IoT VISION

The vision of future internet based on standard communication protocols consider the merging of computer networks, internet of Media(IoM), Internet of Services(IoS) and Internet of things(IoT) into a common global IT platform of seamless networks and network of things [6].

IoS is denoting a software-based component that will be delivered via different networks and internet. Research on SOA, Web/Enterprise 3.0/X.0, enterprise interoperability, service web, grid services and semantic web will address important bits of the IoS puzzle, while improving cooperation between service providers and consumers.

IoM will address the challenges in scalable video coding and 3D video processing, dynamically adapted to the network condition that will give rise to innovative application such as massive multilayer mobile games, digital cinema and in virtual worlds placing new types of traffic demands on mobile network architectures.

The future network of networks will be laid out as public/private infrastructure and dynamically extended and improved by edge points created by the “things” connecting to one another [7]. In fact, in the IoT communications will take place not only between people but also between people and their environment. Communication will be seen more among terminals and data centres (e.g. Home data centres, Cloud computing, etc.) than among nodes as in current networks. Growth of storage capacity at lower and lower costs will result in the local availability of most information required by people or objects.

The future internet will exhibit high levels of heterogeneity as totally different things, in terms of functionality, technology and application fields are expected to belong to the same communication environment. The internet of things will create dynamic network of billions or trillions of wireless identifiable “things” communicating with one another and integrating the

developments from concepts like pervasive computing, ubiquitous computing and Ambient intelligence.

Internet of Things hosts the vision of ubiquitous computing and ambient intelligence enhancing them by requiring a full communication and a complete computing capability among things and integrating the elements of continuous communication, identification and interaction. The internet of things fuses the digital world and the physical world by bringing different concepts and technical components together.

The internet of Things will bring tangible business benefits, such as the high resolution management of assets and products, improved life-cycle management and better collaboration between enterprises many of these benefits are achieved through the use of unique identification for individual things together with search and discovery services enabling each thing to interact individually, building up an individual life history of its activities and interactions over time [7]. Improved sensor and device capabilities will also allow business logic to be executed on the edges of a network – enabling some existing business processes to be decentralized for the benefit of performance, scalability and local decision-making. For example, algorithms could be used for intelligent decision making based on real-time readings from sensors that are used to monitor the health of patients or the condition of vehicles in order to detect the early signs of problems or deterioration of condition.

IoT will create the possibility of merging of different telecommunication technologies and create new services. One example is the use of GSM, NFC (Near Field Communication), low power Bluetooth, WLAN, multi hop networks, GPS and sensor networks together with SIM card technology.

8. CONCLUSION

The Internet of Things is a vision that includes and surrounds diverse technologies at the confluence of nanotechnologies, biotechnologies, information technologies and cognitive sciences. Over the next 10 to 15 years, the Internet of Things will develop rapidly and will shape a new "information society" and an "economy of knowledge", but the direction and pace of events will be difficult of predicting. It will become a technocracy from democracy and anarchy because technology will control everything. The world will become Panopticon (a small number of devices that control large numbers). The buildings of Panopticon were allowed to monitor without their knowledge of the detainees. Similar will be the case where so many devices will monitor, nobody will know who is monitoring what.

Google's Eric Schmidt once said that there will be so many sensors and devices around you that you will forget internet. This means that there will be so many connected devices around you that you will forget the Internet.

The world will become an open book, nothing to hide. Internet security will be totally unsafe, which will lead to the complexity of the wrong elements that cause chaos. However, in positive terms, a very personalized, highly interactive and very interesting world is expected to emerge. The Internet of everything will change the way we work: more information, better decisions, more agile supply chains, more reactive production and greater economic value. The basis of the city of the future will be the Internet of everything and those who adopt this technology are at the forefront. In fact, if we consider the spectrum of possibilities for thin Internet in the period 2020-2025, at this stage we cannot say much, since the technology is still in the process of being perfected, the industry is in a reconfiguration phase.

9. REFERENCES

1. Arshdeep Bahga and Vijay Madiseti, Internet of Things, A Hands-on Approach, Universities Press.
2. Alessandro Bassi et al. "Enabling Things to Talk: Designing IoT solutions with the IoT Architectural Reference Model" Springer Open.
3. IoT. Wikipedia 2015 https://en.wikipedia.org/wiki/Internet_of_Things.
4. From the Internet of Computers to the Internet of Things.
<http://vs.inf.ethz.ch/publ/papers/Internet-of-things.pdf>
5. Internet of Things – From Research and Innovation to Market Deployment
http://internet-of-things-research.eu/pdf/IoTFrom%20Research%20and%20Innovation%20to%20Market%20Deployment_IERC_Cluster_eBook_978-87-93102-95-8_P.pdf
6. Internet of Things Architecture <http://www.iot-a.eu>.
The Internet of Things: How the Next Evolution of the Internet is changing everything
http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

Robotics and advanced Manufacturing.

K.Vineela,

Assistant Professor, Faculty in computer science
AV College of Arts, Science & Commerce,
GaganMahal, Domalguda, Hyderabad, Telangana.

Email-vkvinnu563@gmail.com

Mobile Number: 9949712832

ABSTRACT

Advance manufacturing is the use of innovative technology to improve the product of the Technology adoption and ability to use the technology remain competitive and value define the advance manufacturing sector and we discussed about the manufacturing machines and how the software are developed in advance machine development in manufacturing process and the advanced manufacturing techniques in mechanical, Electrochemical and we discussed the advance hybridization process. With its strong tradition in innovation and close ties to global research China has become the biggest and fastest-growing country in the global industrial robot market for its changing manufacturing environment and improving quality-of-living standards, with foreseeable demand of robots not only in the manufacturing industry, but also other service and societal sectors. Researchers to join efforts and form a critical mass in robotic research to maintain the regional, and towards global lead in service robotic research. Working on the cutting-edge of robotics that is closely associated with the future economy

1. INTRODUCTION:

Robot comes from the Czech word robota, meaning drudgery or slave-like labor. "A reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks".

The 21st century is a century for robotics. Robots have long borne the potential to bridge the gap between the cybernetic world (the internet of things) and the physical world. As the most promising candidate to theme the next major industrial revolution succeeding the present third (digital) industrial revolution, robotics is set to play an ever increasingly important role in society for its influence in every aspect of life in Hong Kong, including medicine and healthcare, building service, manufacturing, food production, logistics and transportation.

2. OVERVIEW:

The first modern industrial robots, called Unimates, were developed by George Devol and Joe Engelberger in the late 50's and early 60's. The first robot patents were by Devol for parts-transfer machines. Engelberger formed Unimation and was the first to market robots. As a result, Engelberger has been called the "**father of robotics**".

2.1 USE:

Most robots are designed to be a helping hand. They help people with tasks that would be difficult, unsafe, or boring for a human to do.

2.2 COMPONENTS:

- Software based control panel
- Computer Interface for control and monitoring
- Mechanical robot hardware

3 AREAS OF MANUFACTURING:

90% of all robots used today are found in factories. These robots are referred to as industrial robots. Although many types can be found in manufacturing today the most common are jointed arm robots. Ten years ago, 9 out of 10 robots were being bought by auto companies - now, only 50% of robots made today are bought by car manufacturers. Robots are slowly finding their way into warehouses, laboratories, research and exploration sites, energy plants, hospitals, even outer space.

Aerospace, Automotive manufacturing and supply, Chemical, rubber and plastics manufacturing, Electrical and electronics, Entertainment-movie making, Food stuff and beverage manufacturing, Glass, ceramics and mineral production, Printing Wood and furniture manufacturing.

4 TASKS IN MANUFACTURING:

- Assembling products
- Handling dangerous materials.
- Spraying finishes.
- Inspecting parts, produce, and livestock.
- Cutting and polishing.
- Welding.



Figure:1 Using robotics in Hospitals



© CanStockPhoto.com - csp46663987

Figure 2: Robots are Cutting hair



Figure 3: Robots are assembling parts of a car

5 A MINDSET SHIFT ON AGRICULTURE IS UNDERWAY

Many people don't picture farmers using technology. The common person would likely visualize a farmer riding a tractor or field workers harvesting crops by hand. With that viewpoint, it's understandable why many STEM students don't see agriculture as a career they wish to pursue. They're more interested in engineering and computer programming, not hard labor in a field. But agriculture isn't just farming anymore.

Farmers use technology daily. Automated drones already monitor fields and collect data on crops, and agricultural robots are being developed to do the fieldwork. Robots have successfully planted, tended, and harvested crops. Contrary to popular belief, many college graduates currently don't possess the skills that agriculture will need in the near future.

To meet the needs of modern agriculture, students will need to pursue science, technology, engineering, and math (STEM) learning paths. In order to revolutionize farming technology, students will need training in the use of autonomous vehicles, robotic farming methods, and the design and application of agricultural robots.



Figure 4:Robots are using in agriculture

6 ADVANTAGES OF ROBOTICS;

- **Competitive Advantage:** Robots can do some things more efficiently and quicker than humans.
- **Mechanical:** Robots never get sick or need to rest, so they can work 24 hours a day, 7 days a week. Greater output per hour with consistent quality. Continuous precision in repetitive operation. Robots don't get bored, so work that is repetitive and unrewarding is no problem.

7 DISADVANTAGES OF ROBOTICS :

- Robots are not creative or innovative.
- Cannot think independently.
- Cannot make complicated decisions.
- Cannot learn from mistakes.
- Cannot adapt quickly to changes in their surroundings.
- Every successful business must depend on real people for these abilities.

8 CONCLUSION:

The robots of today are based on computer technology. The robotics industry is thriving. Higher production capacity can be achieved using robots. Higher quality products are manufactured using robot.

11.REFERENCES:

- [1]. <https://www.ugc.edu.hk/doc/eng/rgc/theme/hall/abs4.pdf>
- [2]. <https://www.slideshare.net/SoundarrajanMadesh/advanced-manufacturing-technology-67211425>
- [3]. <https://www.robotics.org/blog-article.cfm/Agricultural-Robots-The-Future-of-Job-Creation/183>
- [4]. <https://www.slideshare.net/anirudhreddy123/robots-in-manufacturing>

Secure IoT Infrastructure using Blockchain Decentralized Approach

K Naga Maha Lakshmi

Asst Professor, CSE Department, AAR Mahaveer Engineering College, Bandlaguda, Telangana.

Abstract:

IOT is the technology to connect the physical things such as (AC, fridge, car, phone) etc to the internet so that they can communicate with each other without the human interference. IOT devices usually consists of sensors and microcontrollers. Where security plays vital role while data transmission or communication between things over network. In this context a robust IoT solution approach should focus not only securing the infrastructure and devices, which forms the base of IoT system, but also should develop the accurate level of data privacy and building trust with regulators and customers. In this paper we suggest that blockchain is promising for IoT security, It assures that the data is legitimate over the communication process. Blockchain IoT solution could ensure secure messaging between devices on IoT networks.

Keywords: Internet of Things, Blockchain, Decentralization, IoT security.

1. Introduction

Nowadays, the Internet of Things has attracted the attention of researchers, becoming an important technology that promises intelligent human life, allowing communication between objects, machines and humans. IoT refers to a system that contains a real world object and sensors connected to the Internet or connected to the Internet using wired and wireless network architecture. IoT sensors can use a variety of connections, such as RFID, Wi-Fi, Bluetooth and ZigBee, allowing wider area connectivity with many technologies such as GOT, GPRS, 3G and LTE. IoT enabled themes to share information regarding the situation and environment with people, software systems and other devices. The Internet of Things is the process of retrieving and transmitting large amounts of data using real-time systems and links. As every business needs privacy and security, the primary concern is to protect all data and communications.

Now imagine that data flows across a variety of networks and devices, including analytical capabilities, machines, methods, and platforms. In

this way, the data must cross several administrative boundaries, each with its own rules. In this case, ensuring the secure operation of the IoT system and proper data management becomes difficult.

Therefore, it is important not only to protect your data but also to ensure secure data transmission in the right place, in the right form and at the right time. Blockchain technology is suitable on Internet (IoT) link for reliability, privacy and scalability issues. Blockchain technology is probably a magic bullet for connected objects. Blockchain technology can be used to track hundreds of millions of connected devices that allow transaction processing and coordination between devices; Allow significant savings for IoT industry manufacturers. This decentralized approach eliminates certain points of failure, creating a more volatile ecosystem for operation. With blockchain cryptographic algorithms, data becomes private[2].

In the short term, blockchain stabilizes existing IoT solutions with reliable data to inform decision-making and optimize processes. Meanwhile, new IoT applications that improve blockchain functionality are new and sophisticated traditional

offerings. In the long run, the block chain releases a billion-dollar market value and allows for true interoperability between IoT composite devices, businesses and industries.

2. Security Issues in IoT

IoT security issues [1] are noticed while data transmission or communication between two end points using IoT layer architecture such as perception layer, physical layer and application layer as shown in figure 1.

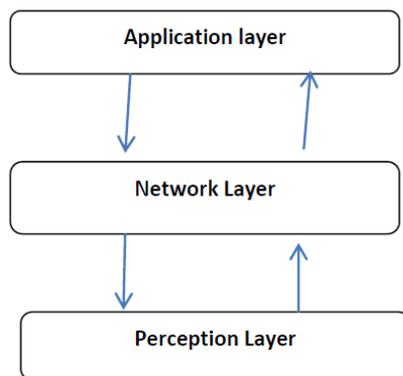


Figure: 1 3 -Layered IoT architecture

2.1 Layer wise Description:

The perception layer consists of various sensors devices such as ZigBee, RFID etc. which deals with the overall management and collection of information by the specific sensor devices. [9]

The network layers forwards information from the perception layers to the layers above it, and helps in privacy of the sensitive information sent from the sensor devices [8]

The application layer function manages the various applications related to IoT. It is responsible for delivering specific services to users; it also defines various applications of IoT [7]

Security issues with layers wise [3]:

1) Security issues in the perception layer: this is a lower level of IoT architecture. The Perception layer is the source of the information source using IoT [3].

Security issues in the perception layer include the security of device detection and data collection.

- IoT cannot provide a defence system and is vulnerable to attack caused by the diversity of the detection node, due to its limited power, uncertainty and the vulnerability of RFID, WSN and M2M terminals.
- RFID poses security issues such as information leaks, repeated attacks, information tracking, counterfeiting, clone attacks, and people-to-people attacks.
- capture gateway node, physical capture, unauthorized attacks, congestion, denial of service attacks, node replication attacks, and direct attacks.

2) Security issues in network layer:

Issues with network layers are

- Illegal access
- Confidentiality
- Data eavesdropping
- Integrity, DoS attacks, destruction, virus attack, man-in-the-middle attack and so on eventually which roots network security issues like data transfer needs of large number of nodes leading to network congestion, resulting in DoS attacks.

3) Security issues in application layer:

Security issues in the application layer include evacuation and tampering.

This layer is responsible for traffic management. It also provides software for various applications that convert data into intelligible or interrogated data [4].

Here we use blockchain technology to address these security issues. Blockchain technology has gained

popularity in recent years. Its role in securing the structure of bit coin crypto currency and other cryptographic assets is widely appreciated. This has proved to be a necessity for many industries and its growing awareness is becoming increasingly aware of this fact. The IoT ecosystem, as we

can see, now works in a centralized model. Various devices identify, connect, and validate cloud services that provide high data storage capabilities. Although these devices are close to each other, the Internet is still used as a communication tool. More prominently, if we can apply a centralized model for small IoT solutions, you can save money. You may not see problems with scalability and running costs. However, large IoT ecosystems can find these related problems and find a way to solve them. This requires a decentralized approach.

- Decentralized: a system where there is no node with a central authority over the other nodes. In other words, the failure of one node will not affect the decision-making process of the system.

3. Blockchain Architecture

Blockchain is a global online database that can be used by anyone with an Internet connection. Because it is on the Internet, it is "decentralized", which means that the blockchain registry is shared on all computers in the world, not centrally.

Blockchain infrastructure is divided into 6 layers. Each layer completes its core function and works together to achieve a decentralized trust mechanism [10].

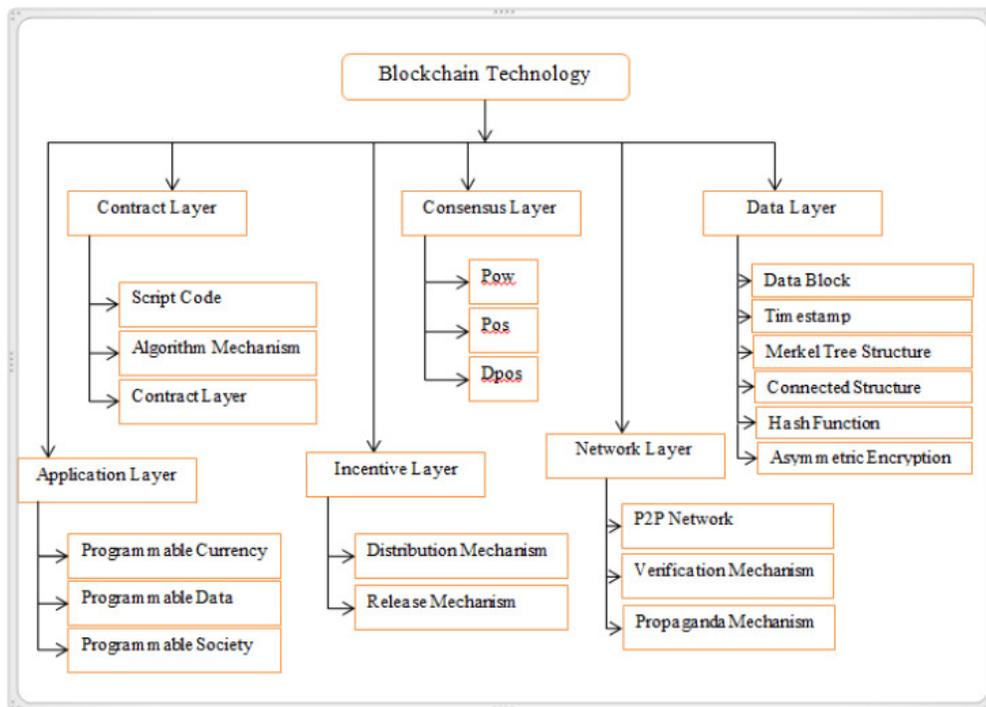


Figure 2. Blockchain infrastructure is divided into 6 layers

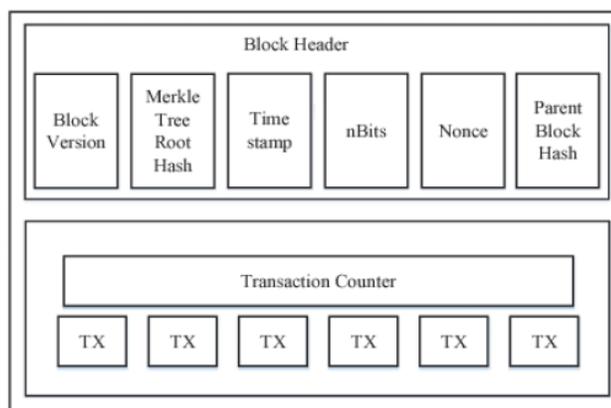


Fig. 3: Block structure

As shown in Figure 3, the block consists of a block header and a block housing. Block headers include[5]:

- (i) Block version: indicates that block validation rules are followed.
- ii) Merkle's tree root hash: this block consist of all transactions hash value
- (iii) Timestamp: Current time in seconds UTF
- (iv) nBits: valid block hash target record.
- (v) Nonce: A 4-byte field that usually starts with 0 and increased for each hash count.
- (vi) Parent Block Hash: A 256-bit hash value that points to the previous block.

4. Key Characteristics of Blockchain

In summary, blockchain has following key characteristics [6].

Decentralization: In traditional centralized transaction systems, each transaction must be tested by a central trusted agency, which inevitably leads to cost and performance problems on the central servers. Unlike the centralized mode, third party is no longer needed in blockchain uses consent algorithms to manage data stability on a distributed network.

Persistency: Transactions can be quickly verified and authentic miners do not accept invalid transactions. It is almost impossible to delete or cancel transactions once they are included in the blockchain. Blocks containing invalid transactions can be found immediately.

Anonymity : Each user interacts with the address generated by the blockchain, which does not reveal the true identity of the user. Note that blockchain does not guarantee complete privacy due to an internal constraint.

5. Securing IoT With Blockchain Approach

The combination of blockchain technology and the Internet of Things is one of the major trends[5].

- The decentralization of an IoT network allows it to address its security challenges such as Technological competences, including reliability, decentralization, scalability and autonomy.
- In the context of the Internet of Things, blockchain technology can be used to ensure successful multi-transaction processing, tracking and integration of millions of smart devices. In addition, since blockchain technology is cryptographic, its integration into IoT networks provides greater security and privacy.

6. Conclusion

In this paper, we conclude that how a blockchain decentralized approach can made secure IoT infrastructure. It will ensure proper security by ensuring the privacy and protection of data at all levels. We also explores the several IoT security issues with respect layers. In addition, blockchain can help resolve scalability issues and provide an effective functioning of the system as well.

References

- [1] sayedaliahmed, Elmustafa& Kamal Aldein Mohammed, Zeinab. (2017). Internet of Things Applications, Challenges and Related Future Technologies. world scientific news.
- [2] QuandengGOU, Lianshan YAN, Yihe LIU and Yao LI —Construction and Strategies in IoT Security System|| 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing.
- [3] TuhinBorgohain, Uday Kumar and SugataSanyal —Survey of Security and Privacy Issues of Internet of Things”.
- [4] Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari and MoussaAyyash —Internetof Things: A Survey on Enabling Technologies, Protocols, and Applications||ieeecomunication surveys &tutorials, vol. 17, no. 4, fourth quarter 2015.
- [5]A. Dorri, S. S. Kanhere, R. Jurdak and P. Gauravaram, "Blockchain for IoT security and privacy: The case study of a smart home," 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kona, HI, 2017, pp. 618-623.doi: 10.1109/PERCOMW.2017.7917634
- [6] Kouzinopoulos C.S. et al. (2018) Using Blockchains to Strengthen the Security of Internet of Things. In: Gelenbe E. et al. (eds) Security in Computer and Information Sciences. Euro-CYBERSEC 2018. Communications in Computer and Information Science, vol 821. Springer, Cham

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

[7] NRI, “Survey on blockchain technologies and related services,” Tech.Rep., 2015. [Online]. Available: http://www.meti.go.jp/english/press/2016/pdf/0531_01f.pdf

Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. 10.1109/BigDataCongress.2017.85.

[8] D. Lee KuoChuen, Ed., Handbook of Digital Currency, 1st ed.Elsevier, 2015. [Online]. Available: http://EconPapers.repec.org/RePEc:eee:monogr:9780_128021170.

[9] V. Buterin, “A next-generation smart contract and decentralized application platform,” white paper, 2014.

[10] Zheng, Zibin&Xie, Shaoan& Dai, Hong-Ning& Chen, Xiangping& Wang, Huaimin. (2017). AnOverview of Blockchain Technology: Architecture, Consensus, and Future Trends. 10.1109/BigDataCongress.2017.85

Machine Learning' -The Future Technology

Author1: G.Shanmukha SessaSai ,CSE department,Narayana Engineering College Gudur,Nellore District,A.P

Author2:A.S.Ramcharan, CSE department,Narayana Engineering College Gudur,Nellore District,A.P

Abstract—This paper helps you to understand, what is “Machine Learning”, and the aspects, opportunities regarding to the study of machine learning. It is a concept of inducing some experience into the machine or a device by giving previous experience use cases through the programming languages like python and can be applied to various fields where there is a need to take right decisions to perform a task. Artificial intelligence is a close related concept to machine learning. It uses the statistical methodologies to provide the computer system or a machine or else a device to learn and act according to a situation. It has been mentioned an under-research application to calculate the proportion of emotion in an image using machine learning image processing tools. Machine Learning is now mostly applied in financial sectors and can also be applied in many scientific and technical fields even in medical fields. It is going to be a buzzword in the next few years because of the development of new algorithms and also easy free availability of the online data.

Keywords—*Buzzword; Inducing; Multidisciplinary*

1. WHAT IS MACHINE LEARNING(ML) ?

The idea of “Machine Learning” comes from the mind of “Yann Le Cun”. Machine Learning is the study of inducing²decision-making skills into a system or a device through the previous use cases and it’s statistical data. You will be wondered that how a machine can make the decision in a given context like a human who uses their brain and analyze the previous experience that they have earned. But, Of course it is possible by inducing some experience of relative contexts. It answers the question, how we can construct a computer that is automatically improved. so, it is going to be a buzzword¹

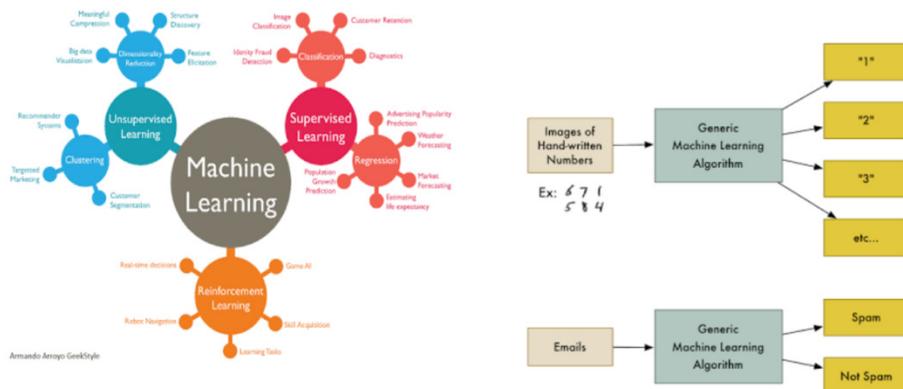
Machine Learning started it’s journey from the anxiety at the programming labs to apply into many of the fields like image recognition, voice recognition

and many more fields with the previous experience. It is easy to train a system or a device by giving the accurate suitable inputs and get the output very easily.



THE MAIN OBJECTIVE OF ML :-

The main objective of Machine Learning is to make the implementation of human intelligence process through the experiences in to the machine/or a computing system/ or any other electronic device. It got popularity because some of the people are interested in analyzing the human intelligence process and some are interested in implementing them. This helps the humans to take right decision or to analyze anything easily, accurately and very quickly



HOW DOES MACHINE LEARNING WORK :-

When a problem is given to a system induced with machine learning, it first analyzes the program by comparing each and every algorithm that is written accordingly to the early experiences faced in the past time, to find out the suitable best algorithm that should be

implemented to complete a given task. The implementation of best algorithm and maximum number of related past context algorithms.

ASSUMPTION

Let us consider a situation that; a machine is to be designed to detect fake note. Then we have to train the machine with a different test cases, i.e., we have to make the machine/system learn and classify the original note and the fake note. The machine is given with many fake notes and the original page features through algorithms. After that when you have given a note as an input then it compares with the previous algorithms and verifies the note as either a fake or original.



Machine Learning algorithms varies differently, in the way through which it represent candidate programs and the way in which they search through the space of programs (i.e., the optimization

algorithms with well-described convergence guarantees that estimates successive generations)

environment and after performing a task and many more tasks it receives actions, unlike An explicit expression of base truth. Semi-supervised and supervised learning strategies are used when fully or complete base truth is available, like labels for classification problem and real-values.

MACHINE LEARNING APPLICATION IN FINANCE FIELD:-

Machine learning had been a wonderful application in the finance field before the invent of mobile banking apps, or search engines, proficient chat bots.

In the modern days, with the latest technology and thinking capability, machine learning plays a vital role in the financial ecosystem, and for approving loans.

But now, It is used for the below mentioned purposes. They are,

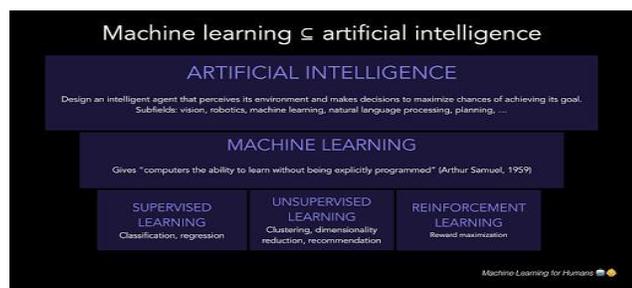
Algorithmic Trading Portfolio Management Fraud Detection
Loan Underwriting Customer Service Security
Sentiment Analysis

As, Machine learning brought a revolution in the way of understanding to human by the machines, With their strategiesto deal with Big Data, these application have became very essential in our day to day life applications.

CONCLUSION:-

In this paper the information regarding what actually Machine Learning means and described that the effective algorithm influences the best performance to make right choice from previous experiences and the inter disciplinary nature of Machine Learning.

The project is to calculate the percentage of emotion that is present in the given image by the analysis of machine learning system. We all know a famous quote i.e., "FACE IS THE INDEX OF YOUR MIND", By implementing this concept through the machine learning strategies, we analyze picture and identify the emotion of face in the given image.



So, let us consider another situation of identifying a suspicious person at any event or a program or a meeting. A person can be easily identified whose intention is to create violence and we can control it by not allowing him in with just a click of a picture and an analysis by the machine learning strategies.

We can even create a machine or an electronic gadget smart device that can easily answer anything with its prior experience. Not only using, the face recognition concept of machine learning, we can make use of voice recognition for the delivery of help signals to the police department to protect the women who are in a problem very easily and quickly.

So, with all these characteristic features Machine Learning will become a buzzword in our day to day life in the next few years.

REFERENCES

1. K. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, Cambridge, MA, 2012).
2. L. Valiant, Commun. ACM 27, 1134–1142 (1984).
3. V. Chandrasekaran, M. I. Jordan, Proc. Natl. Acad. Sci. U.S.A. 110, E1181–E1190 (2013).
4. S. Decatur, O. Goldreich, D. Ron, SIAM J. Comput. 29, 854–879 (2000).
5. S. Shalev-Shwartz, O. Shamir, E. Tromer, Using more data to speed up training time, Proceedings of the Fifteenth Conference on Artificial Intelligence and Statistics, Canary Islands, Spain, 21 to 23 April, 2012.
6. Mohammad Abu Alsheikh, Shaowei Lin,

Dusit Niyato¹ and Hwee-Pink Tan, "Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications", 2014, IEEE Communications Surveys & Tutorials, <http://www1.i2r.a-star.edu.sg/~hptan/publications/IEEECommSurvey2014.pdf>.

7.Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1:127, 2009. <https://www.iro.umontreal.ca/~lisa/pointeurs/TR1312.pdf>.

8.UC Berkeley, AMP Lab,

<https://amplab.cs.berkeley.edu/about/>

9.Berkeley Data Analytics Stack (BDAS),

<https://amplab.cs.berkeley.edu/software/>

10.Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2, 139-172.

APPLICATIONS OF BIGDATA IN MACHINE LEARNING

R.V.GANDHI,Asst Prof,(KMIT,HYD), Research Scholar (UOT,JAIPUR)

ABSTRACT:

Data revolution has changed the whole world and taking it into new generation. Machine learning and Bigdata is the main cause for data revolution leading to a complete data of new formats and unparalleled data bases. The expansion in enormous measure of information has prompted an open door for Machine Learning and Bigdata to meet up and to create Machine Learning strategies which are fit to hold present information types and to explore huge measure of data with negligible or no human intervention. Machine Learning is capable to give flawless results by implementing fast and effective algorithms to process data. Machine Learning is most emerging wide range of areas that is being researched. For a pure Machine Learning process, the more data provided to the system, the more it can learn from it, returning the results that are looking for, and that's why it works well with Bigdata. The Machine Learning can keep running at its at the highest level with Bigdata. If the data is less then the machine has less examples to gain from, and subsequently its results may be influenced. This paper discusses about the applications and challenges of Machine Learning techniques, advanced learning methods towards Bigdata.

KEYWORDS: Bigdata, Machine Learning

1. INTRODUCTION:

Data is ruling the world. The enormous data produced today is not being utilized properly. Many technologies in the present day depend upon data. Though data is available it is not being processed efficiently resulting in miscellaneous errors. Mobiles, Web technologies and Sensing devices, the amount of data is increasing at an unusual rate. For example the amount of data we deliver everyday is truly exciting. There are 2.5 quintillion bytes of data created everyday at current rate. Data is increasing at a rapid pace. By 2020 the new data generated for every individual per second will be the approximate amount of 1.7 mega bytes.

(I) Bigdata: Today, people and systems overload the web with an exponential generation of huge amount of data. The amount of data on the web is measured in exabytes (10¹⁸) and zettabytes (10²¹). By 2025, the forecast is that the Internet will exceed the brain capacity of everyone living in the whole world [1]. Big Data is vast in majority and complex data. Dissimilarity, storage and transport, privacy and security, and complexity problems with Big Data impede the progress at all stages of that can create value from data. There are various sources of Big Data. Basically Bigdata is described in 5V's. The volume, velocity, variety, veracity and value of the data the main aspects. The concept of volume represents of how much data is being produced and its is very necessary that this data can be used in a proper way for eminent results.

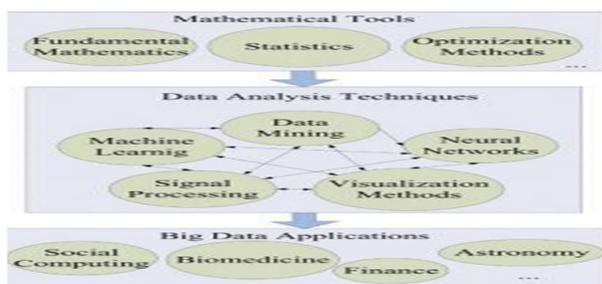


Figure 1: TOOLS FOR BIGDATA

(ii) Machine Learning: Machine Learning is about making the machine to learn by training them. To train we need data. Enomorous data may be needed for quality results. A knowledge base is necessary which can be obtained from the data. The objective is to formulate learning algorithms that do the learning naturally without human assistance or supervision.

Machine Learning is a sub area of artificial intelligence which empowers software applications to get into a state of self-learning without being explicitly programmed [2]. When presented to new data, these systems are empowered to learn, change, grow and implement by themselves. Machine Learning concentrates on the advancement of systems that can get to information and utilize it from themselves [3]. Machine Learning helps to identify the data and trends. The aim is to enable the computers to learn consequently and modify actions appropriately.

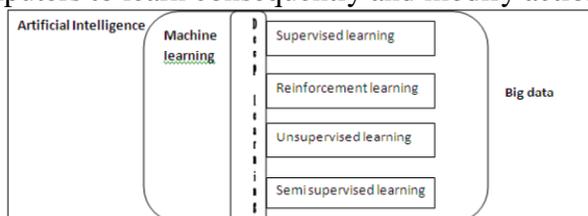


Figure 2 : MACHINE LEARNING IN BIGDATA

2. LITERATURE SURVEY:

Junfei Qiu, surveyed the latest Machine Learning techniques and introduced some of recent learning methods to solve Big Data problem; outlined the critical challenges, research trends, open issues of Machine Learning methods for Big Data processing. [4] Youming, proposed reference anatomy of Machine Learning for Big Data Analytics; analyzed research challenges and issues of Machine associated with Big Data [5]. Sreenivas R. Sukumar reviewed the Machine Learning challenges and analyzed the challenges at scale to Machine Learning in Big Data era [6]. The main problem found in different studies oriented to the processing of large volumes of data resides in the selection of suitable techniques for variable selection and classification [7]. The technique chosen depends on the type of information analyzed, this allows to obtain higher quality information, reduces the computational cost and improves processing times. Some of the most used criteria are: the dimensionality of the data, the relevant features [8] and the veracity of the information obtained. With these considerations we can select the most appropriate Machine Learning techniques that allow us to optimize the results obtained. Many Vital issues of machine learning for Big data are being used [9].

3. MOTIVATION TO THE PROBLEM:

Big Data problems include scenarios related to global economy, society administration, national security and traditional strategies struggle .While dealing with this large data, varying types, high speed and uncertainty many flaws occur. Learning from massively large data can brings significant opportunities for different sectors in business, health, climate, bio, medicine and many more. Most of the traditional machine learning techniques are good for processing structured data but they are lacking in processing unstructured data which requires massive computational efficiency, more scalability to handle the data with massive volume, varying types, great speed, uncertainty, inconsistency and incompleteness [10]. Therefore to design more optimal techniques which can solve huge sized unstructured data efficiently.

4. Machine Learning techniques:

Machine Learning is a field which is raised out of Artificial Intelligence (AI). Applying AI, better and intelligent machines can be built. complex and constantly evolving challenges can be solved. It is very necessary for the machine learn from itself. This sounds similar to a child learning from its self. So machine learning was developed as a new capability for computers. And now machine learning is present in so many segments of technology, that we don't even realise it while using it [11].

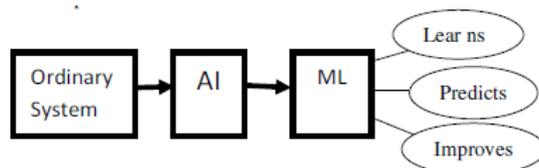


Figure 3 : APPLICATIONS OF MACHINE LEARNING

Suppose we provide a system with input data that contains a picture of a fruit. Then we do:

- First analyze data. Then it tries to find patterns such as color, size...etc.
- Based on these patterns, the system tries to predict different types of fruits belong to particular category,...etc and partition them.
- Finally, it keeps all tracks of the decisions; it took to make sure that it is learning. Then next time when we ask the machine to predict and segregate the different types of fruit. Then it does not go through the entire processes again. That's the machine learning works.

With the rise in big data, machine learning has become a key technique for solving problems in areas [12], such as:

1. Automotive, aerospace, and manufacturing, for predictive maintenance.
2. Computational biology, for tumor detection, drug discovery, and DNA sequencing.
3. Computational finance, for credit scoring and algorithmic trading.
4. Energy production, for price and load forecasting.
5. Image processing and computer vision, for face recognition, motion detection, and object detection
6. Natural language processing, for voice recognition applications.

Machine learning is a domain of research that formally focuses on the theory, performance, and properties of learning systems and algorithms [13]. Generally, the field of machine learning is divided into three sub domains[14]:

- a. Supervised learning.
- b. Unsupervised learning.
- c. Reinforcement learning.

Machine Learning algorithms can be divided into supervised, semi-supervised and unsupervised. In Supervised learning, both the inputs and their desired outputs are provided and an algorithm is used to learn the mapping function from inputs to outputs [15]. Classification and regression are the two main tasks of supervised learning. In classification the output is to predict the target class while in regression the output is to predict the continuous values. K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive-Bayes (NB) are the algorithms for classification [16]. Polynomial regression and linear regression are the algorithms for regression and in which algorithms like Neural Networks can be utilized for both classification and regression.

In unsupervised learning, only the input data is provided and the outputs are not known [17]. Clustering is the example of Unsupervised learning which includes the grouping of objects based on similarity criteria. K-means is the well-known example of this learning. Predictive analytics is based on Machine Learning to implement models used past data to predict the future. Algorithm such as neural networks, Naïve-bayes and Support Vector Machine can be used for this purpose. Semi-supervised learning falls among Supervised and Unsupervised learning. In semi supervised learning both labeled and unlabeled data are used for training-generally a maximum of unlabeled data and a minimum measure of labeled data [14]. Typically semi supervised learning is opted only when the extracted labeled data requires skilled and related resources to learn from it.

Learning Methods	Data processing tasks	Learning algorithm
Supervised	Classification/ Regression/Estimation	Support Vector
		Machine
		Naive Bayes
		Hidden Markov Model
Unsupervised	Clustering/ Prediction	Bayesian network
		Neural Networks
		K-means
		Gaussian Mixture Model
Reinforcement	Decision making	Q-learning
		R-learning
		TD learning

Figure 4 : COMPARISION OF MACHINE LEARNING METHODS

4.1 TYPES OF LEARNING METHODS

This subsection presents some recent learning methods that may play vital role in solving the big data problems.

1) Kernel-based learning: Kernel-based learning is proven to be very dominant methodology to efficiently enhance the computational capacity [18]. The notable advantage of this method is that both linear as well as non-linear vector kernel functional methods are present to deal with the non-linearity of data in N-dimensional feature space.

2) Depiction based learning: This kind of learning [19], is a solution to study valuable representations of the raw data. It is comparatively simpler to get knowledge information while processing through classifiers [60]. Some variants of representational learning [20] are evolved in past years.

3) Active learning: This learning chooses a subset of an unstructured and critical occurrence for purpose of labeling [21]. The active learner obtains larger accuracy using reduced number of occurrences.

4) Deep learning: These designs take more complicated, compartmented statistical patterns of inputs and manage to be robust for new fields as compare to traditional learning systems. “Deep belief networks (DBNs)” [22] [23] and” convolutional neural networks (CNNs)” are two deep learning methodologies.

5) Transfer learning: The prime intention of transfer learning is to derive knowledge features from input source and later implement the knowledge to the target task [24]. The main benefit is that it can efficiently apply knowledge, which has been learned previously in order to find solution for new problems in fast manner.

6) Parallel & Distributed learning: The data which is available in incomplete, inconsistent and unstructured format is first pre-processed, and then cluster forming is done [25]. Count of such distributed clusters is performed. Further one processing thread is assigned to each cluster in order to perform multi-threading in parallel and distributed manner.

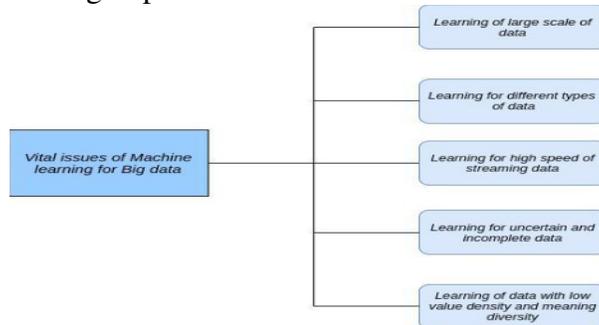


Figure 5 : LEARNING METHODS OF BIGDATA

5. MACHINE LEARNING APPLICATION TO BIG DATA :

Machine learning [26] is ideal for exploiting the opportunities hidden in big data. It delivers on the promise of extracting value from big and disparate data sources with far less reliance on human direction. An overview of the application to big data is given in the figure 4:

It is data driven and runs at machine scale. It is well suited to the complexity of dealing with disparate data sources and the huge variety of variables and amounts of data involved. And unlike traditional analysis, machine learning thrives on growing datasets. The more data fed into a machine learning system, the more it can learn and apply the results to higher quality insights.

5.1 Big data management tools

The entire data analytics industry nowadays has a buzzword, "big data," concerning how we're operating something with the enormous amount of information gathering up. "Big data" is replacing "business intelligence". To handle this massive amount of data available, we have listed out some significant tools that can be utilized to process big data.

5.1.1 Pentaho Business Analytics

It is a kind of software program that started as an engine, branching within big data by creating it simpler to absorb the information from the different sources. One can experiment with Pentaho's tool to many of the most popular NoSQL databases, they are - MongoDB, Cassandra, etc. One can drag furthermore drop the columns into aspects and reports as if the information issued from the SQL databases, once the databases are connected.

5.1.2 Karma sphere Studio and Analyst

It is kind of a specialized IDE that makes it simpler to create and run Hadoop jobs. This produces something better: As we set up the workflow, the tool engine displays the status of the test data at each and every step.

5.1.3 Talend Open Studio

This tool gives an Eclipse-based Integrated Development Environment for stringing data processing operations collectively with Hadoop. Its tools are intended to help with data integration, data quality along with the data management.

5.1.4 Skytree Server

Skytree allows a bundle that delivers many extra advanced ML procedures. All it needs is typewriting the right command in command line. It is more focused on the guts than the shiny GUI. Skytree Server is optimized to execute a no. of classical ML algorithms. It thought of as ten thousand times faster than different packages. It can explore through the data looking for clusters of similar objects, then rearrange this.

5.1.5 Splunk

It is a little distinctive from the other tools. It creates an index of the data as if the data were a part or a block of text. This approach is much alike to a text search method. Splunk will choose text strings and search around in the index. Its variant tool Shep guarantees bidirectional union of Hadoop and Splunk, enabling to interchange data within the systems and query Splunk data of Hadoop.

5.1.6 Jaspersoft BI Suite

It is one of the open source tool for mainly producing reports from database columns. The software tool is well-polished and already installed in many businesses turning SQL tables into PDFs that everyone can scrutinize at meetings. Jaspersoft is not specifically offering unique ways to look at the data, just more complicated ways to access and to locate data stored in the new locations.

6. RESEARCH DIRECTIONS :

The processing such huge sized unstructured, inconsistent, incomplete and vague data by computing machines is a challenging task. In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has grown as an important tool to perform big data analytics. The problem identification and further future research directions are presented as follows: The span of Big Data, is ruling the industries which may be considered as the next bound for modernization, competition and potency. A new boom of revolution is nearly

about to onset where the volume of data today is raging at an unusual rate [27] as a result of advancements and developments in Web technologies, social media, and mobile devices etc. Traditional strategies are hardly suffering when faced with this massive sized data. These traditional machine learning(ML) routines and procedures are not inherently practical or scalable enough to manage the data with the properties of massive volume [28], varying types, great speed, uncertainty and incompleteness.

Based on the precious knowledge, we need to create new techniques and methods to excavate big data. Machine Learning demands to deeply discover itself for processing big data, so that the knowledge extraction and reasoning for uncertain concepts from unstructured and huge sized data can be done in a computationally efficient manner.

The aim of our research is to develop new efficient methods for the analysis of big data sets. Our future research directions are as follows: -

We will contribute some optimal and computationally efficient big data analytics techniques to analyze different type of data sets. This may be achieved by selecting strategies of Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform data analytics.

As, today, processing of massive sized unstructured, inconsistent, incomplete and imprecise data by computing machines is a challenging task. In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has become an important tool to perform big data analytics. To perform operations in the data, present in higher dimensions may be more computationally complex procedure as well as the computational overhead is huge in further training and testing phases of classification. We will employ these modern machine learning techniques to process big data, which also gives the guarantee of dimensionality reduction and other parameters selection of data sets.

7. CONCLUSION:

This paper discusses about various types of learning methods. Further it gives knowledge on some of the significant and practical issues of machine learning for big data analytics. An extensive survey of related work and methods which have been developed in past, is presented. Later, we have listed out some tools which can be employed for big data management and analysis. Future research goals which are useful for further research on this domain are presented.

8. REFERENCES:

- [1] K. Davis, D. Patterson, Ethics of Big Data: Balancing Risk and Innovation, O'Reilly Media, 2012.
- [2] Lidong Wang, Cheryl Ann Alexander, "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 52–61, 2016.
- [3] M.Rouse, "Machine Learning Definition" 2011. <http://whatis.techtarget.com/definition/Machine-learning>.
- [4] Yuhua Xu and Shuo Feng ,Junfei Qiu, Qihui Wu, Guoru Ding, "A survey of machine learning for big data Processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67.

- [5] Junfei Qiu and Youming Sun, “A Research on Machine Learning Methods for Big Data Processing”, International Conference on Information Technology and Management Innovation (ICITMI 2015).
- [6] Sreenivas R. Sukumar, “A Research on Machine Learning Methods for Big Data Processing”, Conference Paper August 20.
- [7] A Survey of Machine Learning Methods for BigData,Zoila Ruiz1, Jaime Salvador1, and Jose Garcia-Rodriguez.
- [8] W. Fan and A. Bifet. Mining Big Data: Current Status, and Forecast to the Future. ACM SIGKDD Explorations Newsletter, 14(2):1-5, 2013.
- [9] D. Saidulu et. al. [10] discussed various types of data types, learning methods, vital issues in big data processing and application
- [10] Al-Jarrah, Omar Y. et al. “Efficient Machine Learning for Big Data: A Review.” *Big Data Research* 2 (2015): 87-93.
- [11]<https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4>.
- [12]<https://in.mathworks.com/discovery/machinelearning.h>
- [13] Junfei Qiu, Qihui Wu, Guoru Ding*, Yuhua Xu and Shuo Feng. “ A survey of machine learning for big data Processing”. Qiu et al. EURASIP Journal on Advances in Signal Processing (2016) 2016:67 DOI 10.1186/s13634-016-0355-x.
- [14] PETER HARRINGTON, “Machine Learning in Action”
- [15] Annina Simon, Mahima Singh Deo, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu,D.R. Ramesh Babu, “An Overview of Machine Learning and its Applications”, International Journal of Electrical Sciences & Engineering (IJESE);Vol1, Issue 1; 2015 pp. 22-24
- [16] TM Mitchell, “The discipline of machine learning”, (Carnegie Mellon University,School of Computer Science, Machine Learning Department, 2006)
- [17] TM Mitchell, Machine learning (McGraw-Hill, New York, 1997)
- [18] M. Chen, S. Mao, Y. Zhang, V.C.M. Leung. ”Big Data: Related Technologies Challenges and Future Prospects”, Springer, Cham, Heidelberg, New York, Dordrecht, London, (2014).
- [19] Nir Friedman, Ron Kohavi.” Bayesian Classification”, Stanford Artificial Intelligence Laboratory, (1999) robotics.stanford.edu/~ronnyk/bayesHB.pdf.
- [20] Y Bengio, A Courville, P Vincent. ” Representation learning: a review and new perspectives”. IEEE Trans Pattern Anal 35(8), 1798-1828 (2012)
- [21] EW Xiang, B Cao, DH Hu, Q Yang. ” Bridging domains using worldwide knowledge for transfer learning”. IEEE Trans Knowl Data Eng 22(6), 770783 (2010)
- [22] F Huang, E Yates. ” Exploring representation-learning approaches to domain adaptation”, in Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (Uppsala, 2010), pp. 23-30.
- [23] D Yu, L Deng.” Deep learning and its applications to signal and information processing”. IEEE Signal Proc Mag 28(1), 145-154 (2011).
- [24] D Peteiro-Barral, B Guijarro-Berdias. ” A survey of methods for distributed machine learning”. Progress in Artificial Intelligence 2(1), 1-11 (2012).
- [25] I Arel, DC Rose, TP Karnowski. ” Deep machine learning-a new frontier in artificial intelligence research”. IEEE Comput Intell Mag 5(4), 13-18 (2010).

- [26] Lidong Wang, Cheryl Ann Alexander. "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences, Vol. 1, No. 2, 5261, (2016).
- [27] Yichuan Wang, Nick Hajli. " Exploring the path to big data analytics success in healthcare", Journal of Business Research 70 (2017) 287-299.
- [28] C.L. Philip Chen, C.Y. Zhang. "Data-intensive applications, challenges, techniques and technologies: a survey on Big Data", Inf. Sci. 275 (2014) 314-347.

Machine Learning Database for Natural Language Processing

N Jayanthi, Assistant Professor, CSE,IARE, Hyderabad.

A. Ravi Prasad, GDC Piler.

Abstract

In today's digital world machine learning and predictive analytics are gaining a lot of attention. These are used in multiple domains such as bioinformatics, computational anatomy, natural language processing, speech recognition, etc. based on specific purpose machine learning can be used to train software or hardware. One contribution of machine learning is Machine Learning Database (MLDB). Machine Learning Database is a powerful and high performance database system specifically developed for machine learning, knowledge discovery and predictive analytics. The Traditional database use historical data and the select queries work on 'what happened?' "whereas Machine Learning Database use new data and the select queries work on "what will happen?" i.e. we get a prediction. The aim of this paper is to shed light on various components of MLDB and its application for natural language processing.

Introduction

In today's digital world machine learning and predictive analytics are gaining a lot of attention. These are used in multiple domains such as bioinformatics, computational anatomy, natural language processing, speech recognition, etc. based on specific purpose machine learning can be used to train software or hardware[2]. Various supervised or unsupervised algorithms are developed for obtaining accurate results from data. Even though Machine learning and Artificial intelligence are used very often but they differ in reality. Artificial Intelligence is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages where as Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed[2]. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. High optimization and accurate results are obtained by machine learning based implementation. Some of the sophisticated tools are Keras, Lime, Apache Singa, MLDB[3].

Machine Learning Database (MLDB)

Machine Learning Database (MLDB) is developed by Jeremy Barnes and François Maillet[1]. Machine Learning Database (MLDB) is a powerful and high performance database system specifically developed for machine learning, knowledge discovery and predictive analytics. MLDB is FOSS and is compatible with assorted platforms. It uses the RESTful API for the storage of data, exploring the data using Structured Query Language (SQL) and, finally, it trains machine learning models[4]. The following are the key features of MLDB are speed, scalability,

free and open source. The SQL support of MLDB is very user-friendly along with the support for Big Data processing. MLDB can process, train and make predictions using database tables that have millions of columns, with concurrent processing and no compromise on integrity. MLDB can be implemented on multiple platforms such as Jupyter, Docker, JSON, Cloud, Hadoop, and many others. It can be compatible with different application programming interfaces (APIs) and modules including JSON, REST and Python based wrappers. MLDB can be deployed easily on an HTTP endpoint that provides easy interface and fast deployment. MLDB supports enhanced form of SQL queries.

The procedures of MLDB are used for the training of machine learning models and these are implemented using functions. Given below is the list of functions and procedures that can be used in MLDB with high performance.

There are two editions of MLDB that are free, and are distributed as community and enterprise editions[4]. To run the MLDB enterprise edition, you need to enter the licence key to activate the software. A licence key can be created for first-time users on signing up at https://mldb.ai/#license_management and filling the required details in the registration form.

The classical MLDB distribution is the Docker image. While other distributions are available for virtual machines, the Docker image is executed as a container. This method is used for Linux flavours or private cloud deployments[4].

MLDB is an open-source database designed for machine learning. It can be installed wherever you want and send it commands over a RESTful API to store data, explore it using SQL, then train machine learning models and expose them as APIs. In MLDB, machine learning models are applied using Functions, which are parameterized by the output of training Procedures, which run over Datasets containing training data[2]. Functions are also available to SQL Queries and as REST Endpoints. An overview of MLDB is presented in figure 1.

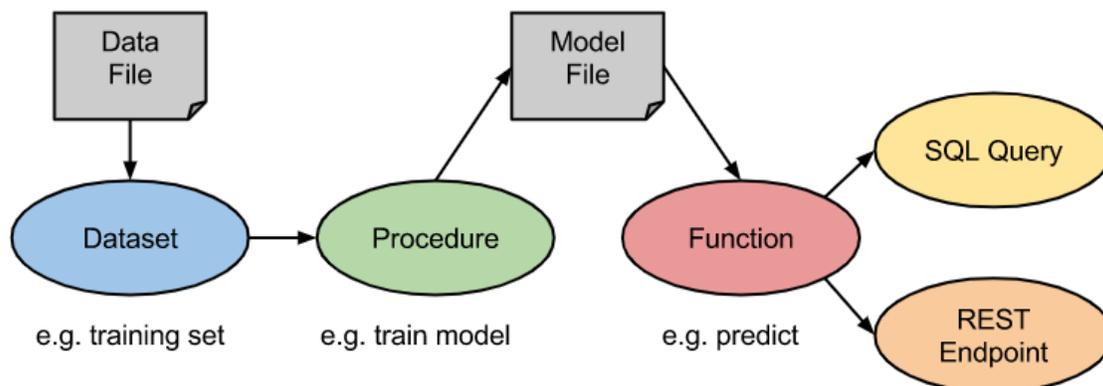


Fig 1 : overview of MLDB (taken from <https://docs.mldb.ai>)

Functions of MLDB

In MLDB functions can be named and reusable. These functions can accept arguments and return values, used to encapsulate SQL expressions, applied to machine learning models. All

MLDB Functions are also automatically accessible as REST Endpoints, which can be used to apply machine learning models in a streaming/real-time process. Some of functions are pooling, stemmer, tensorflow. graph etc[2].

Procedures

The procedures in MLDB are named and reusable with no return values and are used to implement long-running batch operations. These are applied to datasets mostly. The outputs of procedure can be datasets and files[2]. For transforming or for cleaning data, to train machine learning models and to apply machine learning in batch mode MLDB procedures are used. Some of the procedures are classifier.test, classifier.train, export.csv, import.git, import.json, import.sentiwordnet, import.text, import.word2vec, kmeans.train.

MLDB procedure has the following format <verb><resource><arguments>

A verb can be any one of the four types GET, PUT, POST and DELETE.

Resources can be datasets, procedures, functions, plug-ins

Arguments can be query – string.

Datasets

The datasets in MLDB are a combination of row, column, timestamp and value, these are named data points. Data points can take on multiple time stamped values which can be text or numeric. Datasets are given as inputs to procedures and can be stored or can be loaded from files[2]. Some of the examples of datasets are embedding, merged, sampled, sparse.mutable, tabular.

Files

Files can be used for saving in and loading of datasets. They can be created through procedures and parameters for files can be uploaded through functions. If data is stored in a system such as S3 or HDFS, it is possible to run multiple instances of MLDB so as to distribute computational load[2].

SQL Queries

MLDB SQL queries are based on SQL. These queries are similar to SQL queries and beyond them[2]. The queries can be used for specifying queries and for evaluation of expression. SELECT, NAMED, FROM, WHEN, WHERE etc.

Plugins

MLDB plugins are used for loading dataset types, procedure types and function types. As the plugins have access to a low level C API new dataset types, procedure types, function types and even new plug-in can also be defined. On creation of new dataset types users can integrate MLDB with new sources of data, creation of datasets can pull from other databases, creation of new procedures and functions types enables users to integrate their algorithms into MLDB. By using plugins ,MLDB's REST API can be extended this help to create a very good user

interfaces[2]. Some of the plugins are Hello, world! - Minimal MLDB plugin, DeepTeach - DeepTeach is the interactive deep image classifier builder, MLPaint - MLPaint is the Real-Time Handwritten Digit Recognizer etc.

MLDB can be used to solve a machine learning problem in the following steps:

Start with a file full of training data, which is loaded into a Training Dataset.

A Training Procedure is run, to produce a Model File

The Model File is used to parameterize a Scoring Function

This Scoring Function is immediately accessible via a REST Endpoint for real-time scoring

The Scoring Function is also immediately accessible via an SQL Query

A Batch Scoring Procedure uses SQL to apply the Scoring Function to an Unscored Dataset in batch, producing a Scored Dataset[4].

References

[1] <https://mldb.ai>

[2] <https://docs.mldb.ai>

[3] <https://www.kdnuggets.com/2016/10/mldb-machine-learning-database.html>

[4] <https://opensourceforu.com/2018/05/mldb-the-open-source-machine-learning-database-for-the-cloud-and-for-docker>

[5] <http://blog.mldb.ai/blog/posts/2017/02/elementai>

MACHINE LEARNING APPROACH FOR PREDICTING MEDICAL DIAGNOSIS

S. Prabhakar¹ and Dr. M.Sujatha²

¹Assistant Professor, Kakathiya University , Warangal, Telangana

²Associate Professor, Jyothishmathi institute of technology and science, Karimnagar, Telangana

ABSTRACT

The Medical data contains massive and complicated knowledge which will be needed so as to get fascinating pattern of diseases & makes effective choices with the assistance of various machine learning techniques. Advanced data processing techniques are wont to discover data in information and for medical analysis. This paper has analyzed prediction systems for polygenic disease, excretory organ and disease victimization a lot of variety of input attributes. the information mining classification techniques, particularly Support Vector Machine(SVM) and Random Forest (RF) are analyzed on polygenic disease, excretory organ and disease information. The performance of those techniques is compared, supported exactness, recall, accuracy, f_measure further as time. As a results of study the planned algorithmic program is meant victimization SVM and RF algorithmic program and therefore the experimental result shows the accuracy of 91.5%, 94.3 and 97.4 on polygenic disease, excretory organ and disease severally.

KEYWORDS

Data Mining, Clinical Decision Support System, Disease Prediction, Classification, SVM, RF.

1. INTRODUCTION

Computational health scientific discipline is rising analysis topic that involving varied sciences like medicine, medical, nursing, information technology, and statistics [1]. Data processing techniques applied to predict the effectiveness of big and sophisticated clinical information so as to diagnose unwellness and extract data to recommend effective medical help [2]. In life science, doctor's facilities introduced completely different data frameworks with lots of information to manage medical insurance and patient data but sadly, data do not appear to be well-mined to search out hidden information for effective call [2][3].

Clinical test outcomes square measure usually created on the idea of doctor's perception and {knowledge or skill} rather than on the knowledge enrich knowledge covert among the information and usually this procedure prompts accidental predispositions, doctor's expertise may not be capable to diagnose it accurately that affects the unwellness designation system [2][3]. In aid sector, the term data processing can mean to analysis the clinical knowledge to predict patient's health standing. So discovering fascinating pattern from tending info, totally different data processing techniques square measure applied with applied math analysis, machine learning and information technology.

Predictive systems predict some outcome on the idea of some pattern recognition, as shown in figure1. Unwellness detection is that the tactic by that patient's designation is performed on the idea of symptoms analyzed which can causes problem whereas predicting unwellness have an effect on [4]. As associate example, fever itself may be a proof of the various disorders that doesn't tell the tending skilled what precisely the unwellness is the results or opinion vary from one medico to a distinct, there is a demand to assist a medical medico which can have similar opinion definitely symptoms and disorders [5]. This might be done by analyzing information generated by medical data or medical records.

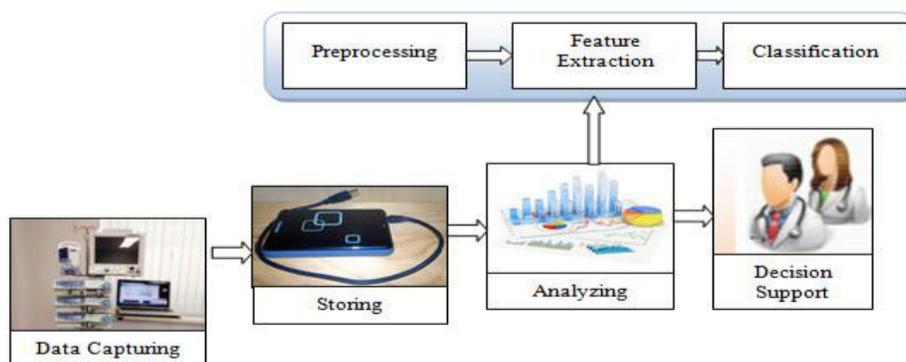


Figure 1: A Typical Medical Data Processing

As a result, the new info is commonly compared with previous records and optimistic identification is commonly done. prognostic {medical identification| diagnosis| diagnosing} may well be a web application that is in a position to predict a particular disorder on the idea of symptoms and provide diagnosis for some disorder that is in a position to be detected by rule. Care professionals use their previous knowledge and insights to achieve a specific call concerning any illness or disorder [6]. Inside the similar manner, this paper proposes totally different classification techniques for identification by victimization generic illness datasets. This paper brings into limelight all the advantages and disadvantages of victimization the varied data processing techniques for the prediction of diseases. It together accounts for the prediction rate for numerous techniques so, transferal out the comparison between every of them [7]-[10].

Data mining has been with success employed in data discovery for prognostic functions to create a lot of active and correct call [11]. the most focus of the paper is on classification yet as clump techniques. In clump method like-means, EM, Fuzzy c-means, etc, knowledge is partitioned off into sets of clusters or sub-classes [12]. Machine learning techniques like KNN, SVM, Naïve Bayes etc, may be accustomed classify totally different objects on the idea of a coaching set of knowledge whose outcome worth is understood.

Classification Techniques

Machine learning primarily based classification techniques are often accustomed to classify numerous objects supported a series of coaching information whose result price is thought. during this study four classification algorithms are used: KNN, SVM, Naive Bayes and C5.0. within the nearest neighbor KN, the article is classed by the bulk of its neighbors, with the article being allotted to the foremost normally used category among its nearest neighbors. In SVM (Support Vector Machines), information is initial reborn into a collection of points so classified into categories which will be separated linearly. The Naive model calculates the chance of a collection of information which will belong to a category exploitation the Bayes rule. The C5.0 algorithmic rule may be a call tree that recursively separates observations in branches to make a tree to enhance prediction accuracy. It's an improved version of the C4.5 and ID3 algorithms [10]. It additionally provides the powerful gain methodology to extend the accuracy of this classification algorithmic rule [11].

Clustering Techniques

The cluster method divides the information into cluster teams or subclasses. we have a tendency to used four cluster algorithms, particularly K-Means, EM, PAM, Fuzzy C-Means [12]. The K-Means classification algorithmic rule works by partitioning n observations in k-subclasses outlined by centroids, wherever k is chosen before the algorithmic rule begins. K-Means and EM are 2 repetitive algorithms. EM (expectation-maximization) may be applied mathematics model that depends on the unobserved latent variables to estimate the utmost chance parameters. Partitioning around medoids (MAP) is comparable to K-means that partitioning is predicated on the K-medoids methodology, which divides information into variety of disjoint clusters [12]. In fuzzy cluster, information components will belong to multiple clusters. this can be additionally referred to as soft cluster.

VII. RELATED WORK

In [1] neural networks, decision tree and naïve bayes machine learning approach is used to diagnose heart disease. For optimized feature selection genetic algorithm is used and obtained an accuracy of 100%, 99.62 and 90.74 respectively.

In [2], performed a prediction of heart disease detection using neural network with genetic algorithm based feature extraction. Back propagation based neural network weight optimization by Genetic algorithm is designed and obtained the 89% accuracy of prediction of heart disease.

In [3], author developed a heart disease prediction system using an approach of ANN with LVQ and achieved accuracy of about 80%, sensitivity of about 85% sensitivity as well as specificity of about 70%.

In [4] proposed a heart disease diagnosis is proposed using lazy data mining approach with data reduction strategies i.e. principal component analysis is used to get category association rules. The result analysis shows that J4.8 has 10.26 enhancement as well as 8.6% enhancement over naïve bayes.

In [5] presented a heart disease prediction system using data mining approach with two additional features i.e. obesity and smoking to boost the prediction rate. Neural networks, Decision trees and Naive Bayes was used in for predicting heart disease with an accuracy of 99.25%, 94.44% and 96.66% respectively.

A web based application has introduced in [6] using Naïve Bayesian algorithm which took symptoms from user and gave the diagnosis result to the user or patient. In [7] Association rule mining technique was used for diagnosis of diabetes. The authors concluded that the data mining techniques when used appropriately increases the computation and also the classification performance. These rules have the potential to improve the expert system and to make better clinical decision making. For predicting diabetes disease on weka tool, author in research work [8] had presented a comparison between Naïve bayes algorithm and decision tree algorithm and achieved system accuracy of about 79.56% and 76.96% respectively.

In [9] Decision Tree, Naive Bayes, and NBTree algorithms is used for liver disease detection with 10 features. The result analysis with respect to accuracy NBTree algorithm has the highest accuracy whereas with respect to computational time Naive Bayes algorithm performs better. In [11] author performed a comparative analysis on clustering and classification algorithms. The result analysis shows that the classification is better than clustering algorithms with an accuracy of about 81%.

In [12] author proposed classification with clustering technique i.e. KNN with FCM clustering and F-KNN with FCM clustering. It is clear that the Fuzzy KNN with Fuzzy c-means model produced the better result than the KNN with Fuzzy c-means model on both PIMA and Liver-disorder datasets. It is also clear that the use of Fuzzy c-means clustering algorithm for preprocessing of datasets improved the result in terms of classification accuracy and speed by reducing the number of tuples from the original datasets. From experiment, it is been found that KNN with Fuzzy c-means have accuracy of 97.02 and Fuzzy KNN with Fuzzy c-means have accuracy of 99.25 on PIMA dataset whereas on Liver disorder KNN with Fuzzy c-means have accuracy of 96.13 and Fuzzy KNN with Fuzzy c-means with accuracy of 98.95.

In [13] performed the chronic disease prediction by using data mining approach such as Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) for the diagnosis of diabetes and heart disease. The result analysis shows that SVM gives highest accuracy of 95.556% in case of heart disease and Naïve bayes gives accuracy of 73.588% in case of diabetes. Table I gives the comparative analysis of different existing techniques for heart, liver and diabetes diseases.

Table 1: Comparative Analysis of Different Techniques in terms of Accuracy

Name of Author	Technique	Accuracy
Bhatla et al.	Neural Network, Naïve Bayes and Decision Tree for heart disease detection.	Neural networks = 100 % Decision tree= 99.62 % Naïve Bayes 90.74 %
Amin et al.	Optimized Neural Network for heart disease detection.	Training data was = 89% Validation data = 96.2%.
Chen et al.	ANN based heart disease detection.	ANN = 80%.
Dangare et al.	Decision trees, Neural networks and Naive Bayes for heart disease detection.	Neural Networks = 99.25% Naive Bayes = 94.44 % Decision Tree =96.66 %
Iyer et al.	Decision tree and Naïve bayes algorithm for diabetes detection.	Decision Tree =76.96% Naïve Bayes= 79.56%
Uma Ojha and Savita Goel	Decision Tree, SVM, FCM	Decision tree (C5.0) =81% SVM =81% FCM = 37%
Chetty et al.	KNN and F-KNN for diabetes and liver disease detection.	KNN = 97.02% F-KNN = 99.25% for diabetes data. KNN = 96.13% F-KNN =98.95% for liver disease data.
	Naïve Bayes, Decision tree,	

Kumari Deepika and Dr. S. Seema	Support Vector Machine (SVM) and Artificial Neural Networks (ANN) for diabetes and heart disease detection.	SVM = 95.556% (heart disease) Naïve Bayes = 73.588% (diabetes).
------------------------------------	--	--

3. METHODOLOGY

3.1 Proposed Methodology

One of the interesting and important subjects among researchers within the field of medical and technology is designation malady by considering the options that have the foremost impact on recognitions. the topic discusses a brand new construct that is termed Medical data processing (MDM). Indeed, data processing ways use other ways like classification and bunch to classify diseases and their symptoms that ar useful for designation.

A sickness designation system is formed so as to predict totally different diseases like polygenic disease, nephropathy similarly as disease, etc. System's progress is mentioned below:

Step 1: Through the planned application user (doctor, patient, MD etc.) will input the attribute values of sickness and send it to the choice web for analysis.

Step 2: At call web, dataset of various diseases ar loaded and apply data processing algorithms to coach dataset. requested user inputs ar collected and processed on server to predict the designation result.

Step 3: For analyzing tending knowledge, major step for mining approaches like preprocess data, replace missing values, feature choice, machine and make decision are applied on train dataset. On the decision support system end different classification algorithms would be executed on train dataset and ready to classify the test dataset.

In the proposed method Support Vector Machine and Random Forest represents cluster levels for various subspaces. The ensemble model uses voting technique for classification . Finally, the target results are compared with class labels of the testing phase.

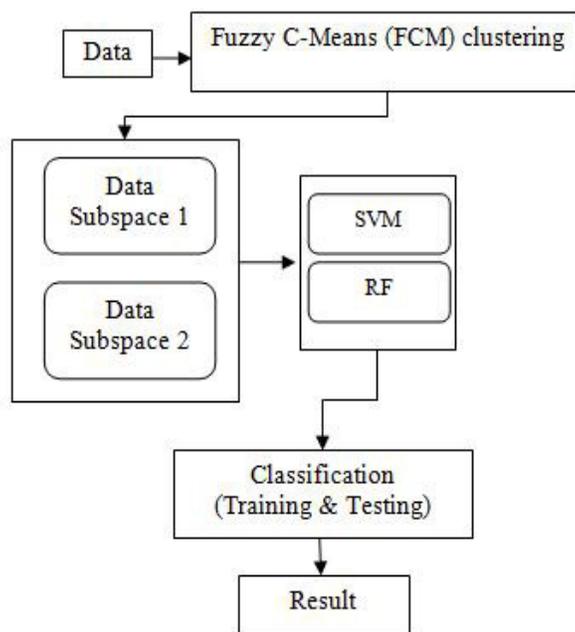


Figure 2: Proposed Model

Support Vector Machine

Support vector machine is a machine learning approach that can be used as classifier as well as for regression. SVM classifies the data into different classes by finding hyperplane (line) which separates training data into classes. SVM does not overfit the data and gives best classification performance in terms of precision and accuracy.

SVM does not make any strong assumptions on data. It shows more efficiency for correct classification of the future data. SVM is classified into 2 categories i.e. Linear and non-Linear. In linear approach, training data is separated by line i.e. hyperplane.

Random Forest

Random Forest algorithm is capable of performing each classification and regression tasks. the fundamental principle of RF is that a group of weak learner's come together to create a robust learner. Random forest rule uses bagging approach to form the bunch of decision trees with random set of the information. The model is trained few times on random sample of the dataset to attain best prediction performance from the RF rule. In this ensemble technique of learning, the output of all decision trees within the RF is combined to form a final prediction. The final prediction of the RF rule is derived after polling the results of every decision tree.

Suppose there are N cases within the training set. Then these N samples are taken randomly however with replacement. These samples are training set for growth of tree. If $m < M$ is specific. The simplest split of this m is employed to separate the node. The value of m is constant whereas growing the forest.

Dataset Description

Pima Indians Diabetes Database

This study used data sets from the Pima Indians Diabetes Database of National Institute of Diabetes [14]. This dataset consists of 768 samples with 8 numerical valued attribute where 500 are tested negative and 268 are tested positive instances.

Chronic Kidney Disease Dataset

This study used data sets from the university of California Irvine (UCI) repository. The data set contains 400 patients, where 250 patients were positively affected by kidney disease and as many as 150 patients do not suffer from kidney disease.

Liver Disorders Data Set

This study used data sets from the university of California Irvine (UCI) repository. This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

Performance Measures

In this study, we used three performance measures: Precision, Accuracy, Recall, F_measure and Total execution time.

Accuracy is termed as ratio of the number of correctly classified instances to the total number of instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is the ration of actually true predicted instances out of the total true instances.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is the ratio of actual true instances out of all the items which are true.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure is the harmonic mean of both precision and recall.

$$F_Measure = \frac{2*(Precision*Recall)}{Precision + Recall}$$

Where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively.

4. RESULT ANALYSIS

Below Table 2 and 3 as well as Figure 3 shows the comparative analysis of proposed algorithm with some existing algorithms.

Table 2: Result Analysis of Diabetes Disease Detection

Diabetes Disease Detection			
Recall	Precision	Accuracy	F_measure
1	0.9821	0.9151	0.991

Table 3: Comparative Result Analysis of Diabetes Disease Detection

Accuracy Measurement	
Existing Work [14]	90.4%
Proposed Work	91.5%

Figure 3: Comparative Chart of Diabetes Disease Detection

Below Table 4, 5 and 6 shows the parameter values for different diseases such as diabetes disease, kidney disease as well as liver disease. Similarly figure 4 and 5 shows the corresponding parametric chart of different disease detection using proposed algorithm.

Table 4: Result Analysis of Kidney Disease Detection

Kidney Disease Detection			
Recall	Precision	Accuracy	F_measure
1	0.9875	0.943	0.9937

Table 5: Result Analysis of Liver Disease Detection

Liver Disease Detection			
Recall	Precision	Accuracy	F_measure
0.9667	1	0.974	0.9831

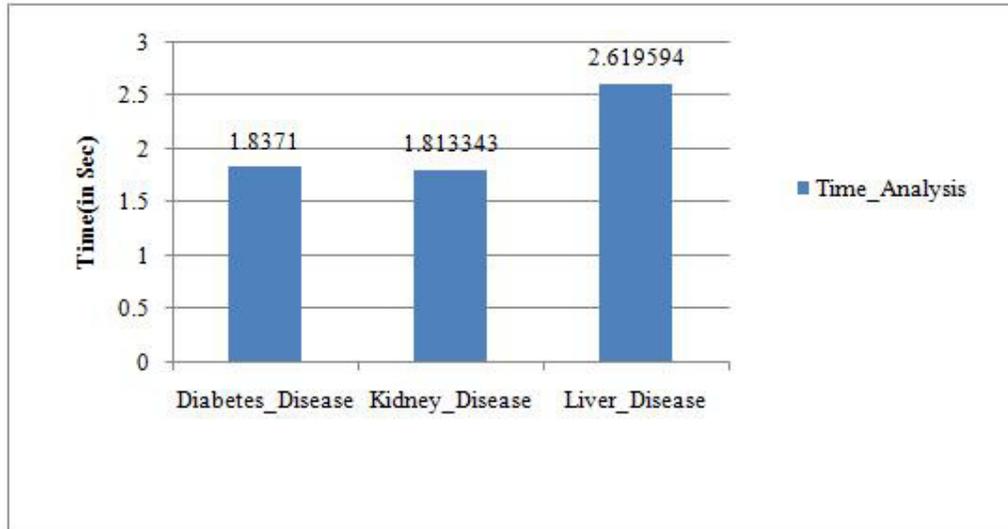


Figure 2: Time Comparison Chart of Different Disease Detection

5. CONCLUSION

This research paper is mainly focused to predict disease possibility using data mining or machine learning approach in order to enhance the accuracy or precision of the disease detection expert system. This paper also shows the related work study of different approaches such as neural network, naïve bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques. As a result of study the proposed algorithm is designed using SVM and RF algorithm and the experimental result shows the accuracy of 91.5%, 94.3 and 97.4 on diabetes, kidney and liver disease respectively. In future using data mining approach a new optimized intelligent system can be designed which can give accurate and efficient result.

REFERENCES

- [1] Nidhi Bhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", IJERT, Vol 1, Issue 8, 2012.
- [2] Syed Umar Amin, Kavita Agarwal, Rizwan Beg, "Genetic Neural Network based Data Mining in Prediction of Heart Disease using Risk Factors", IEEE, 2013.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

- [3] A H Chen, S Y Huang, P S Hong, C H Cheng, E J Lin, "HDPS: Heart Disease Prediction System", IEEE, 2011.
- [4] M. Akhil Jabbar, B. L Deekshatulu, Priti Chandra, "Heart Disease Prediction using Lazy Associative Classification", IEEE, 2013.
- [5] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", IJCA, Volume 47- No.10, June 2012.
- [6] P. Bhandari, S. Yadav, S. Mote, D.Rankhambe, "Predictive System for Medical Diagnosis with Expertise Analysis", IJESC, Vol. 6, pp. 4652-4656, 2016.
- [7] Nishara Banu, Gomathy, "Disease Forecasting System using Data Mining Methods", IEEE Transaction on Intelligent Computing Applications, 2014.
- [8] A. Iyer, S. Jeyalatha and R. Sumbaly, "Diagnosis of Diabetes using Classification Mining Techniques", IJDKP, Vol. 5, pp. 1-14, 2015.
- [9] Sadiyah Noor Novita Alfishahrin and Teddy Mantoro, "Data Mining Techniques for Optimatization of Liver Disease Classification", International Conference on Advanced Computer Science Applications and Technologies, IEEE, pp. 379-384, 2013.
- [10] A. Naik and L. Samant, "Correlation Review of Classification Algorithm using Data Mining Tool: WEKA, Rapidminer , Tanagra ,Orange and Knime", ELSEVIER, Vol. 85, pp. 662-668, 2016.
- [11] Uma Ojha and Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques", International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.
- [12] Naganna Chetty, Kunwar Singh Vaisla, Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach", International Conference on Advances in Computing and Communication Engineering, IEEE, pp. 568-572, 2015.
- [13] Kumari Deepika and Dr. S. Seema, "Predictive Analytics to Prevent and Control Chronic Diseases", International Conference on Applied and Theoretical Computing and Communication Technology, IEEE, pp. 381-386, 2016.
- [14] Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", IEEE, 2017.

Artificial Intelligence and Robotics via Games

First Author : K.Vishnu Vardhan Reddy, B-tech-CSE Department,Narayana Engineering College, Gudur,AP, subscript:vishnukummitha8@gmail.com

Second Author : M.Rohit, B-tech-CSE Department,Narayana Engineering College, Gudur, AP, subscript:rohit.mvs1997@gmail.com

To Reference: Krishna Prasad.

Abstract

The Department of Computer Science recently created two new degree programs, namely a Bachelor's Program in Computer Science (Games) and a Master's Program in Computer Science (Game Development). In this paper, we discuss two projects that use games as motivator. First, the Computer Games in the Classroom Project develops stand-alone projects on standard artificial intelligence topics that use video-game technology to motivate the students but do not require the students to use game engines. Second, the Pinball Project develops the necessary hardware and software to enable students to learn concepts from robotics by developing games on actual pinball machines.

INTRODUCTION

The Department of Computer Science recently created two new degree programs, namely a Bachelor's Program in Computer Science (Games) and a Master's Program in Computer Science (Game Development) (Zyda and Koenig 2008). In addition, we explore whether games can be used to teach traditional concepts from computer science in our regular computer science classes because games motivate students, which we believe increases enrollment, motivation and retention and thus helps us to educate more and better computer scientists. In general, games can be used to teach almost every area of computer science. For example, computer architecture is important for understanding how game consoles work, networking is important for building

networked games, human-computer interaction is important for designing user-friendly games, and algorithms are important for *The following students made the Computer Games in the Classroom Project possible: Kenny Daniel, Alex Nash, William Yeoh and Xiaoming Zheng. The following students made the Pinball Project possible: Hrudesh Doke, Darren Earl, Clark Kromenaker, Allen Pan, Selby Shlosberg, Jaspreet Singh, Daniel Wong, Ryan Zink and Fred Zyda. Our initiatives were partly supported by a grant from the Fund for Innovative Undergraduate Teaching the Rose Hills Foundation and the National Science Foundation under grants 0350584, 0413196 and 0113881. This abstract reuses ideas, photos and a small amount of text from our longer papers cited in the bibliography

Furthermore, games can be used to teach a variety of important job skills for computer scientists in both academia and industry, including technical skills (such as computational thinking, software engineering and programming skills), creativity, design skills, problemsolving skills and teamwork skills (such as collaboration skills with non-computer scientists). In this paper, we discuss two projects that explore how to teach traditional concepts from artificial intelligence and robotics using games as motivator. It is future work to evaluate the effects of our efforts.

Computer Games in the Classroom Project

The undergraduate and graduate versions of the artificial intelligence class at the University of Southern California are taken by both game students and regular computer science students. We explore how to coach their projects in terms of video-game technology. Video games can be used to teach artificial intelligence because artificial intelligence is important for creating more realistic or more fun games. Most video games use search algorithms for path planning. For example, Dragon Age: Origins uses sophisticated search algorithms with abstraction hierarchies to satisfy the runtime and memory requirements imposed by BioWare. Some video games also use planning or machine learning algorithms. Those students who are familiar with game development already understand game engines well while the regular computer science students might not want to learn about them, at least not at the same time as learning artificial intelligence. We therefore decided to create several stand-alone projects on standard artificial intelligence algorithms that use games as motivator but do not make use of game engines (Zyda and Koenig 2008). We now describe the three projects that we have developed so far, some of which have already been used successfully at other universities, namely the University of Nevada at Reno, the University of Central Florida and Massachusetts Institute of Technology. Each project text is about 15 pages

long and motivates the project, introduces an algorithm, gives examples of its use and then provides a variety of possible questions, including easy and difficult ones. Teachers need to select among them since each project text lists too many of them for a project of a reasonable size and some of them are difficult research questions. More information can be found on our web pages at

The Nintendo Wii is a big success because its motion-sensitive paddle lets users innovatively control objects on the screen with hand gestures. Still, some people say that all gaming instruments that we use today will look ridiculously old-fashioned and redundant in ten years ... In the near future, for example, users will likely be able to control objects on the screen with bare hands ... We limit our ambition to a static variant of the gesture recognition problem, where the computer has to classify hand gestures in single images. ... [C]omputers can easily get confused by ... different hands, angles, backgrounds, lighting conditions and other differences. In this project, we use (artificial) neural networks to recognize hand gestures in single images. Neural networks are among the most important machine learning techniques. They are too general to reliably classify a large number of hand gestures without any preprocessing of the images but are able to classify a small number of hand gestures out of the box. There exist more specialized gesture recognition techniques that scale much better but the advantage of neural networks is their generality. They are able to solve a large number of classification problems ..., and are thus often ideal solutions for simple classification problems and for prototyping solutions of more complex classification problems

Fast Trajectory Replanning with Variants of A:

The students need to code A* and extend it to the recently developed incremental heuristic search algorithm Adaptive A*, which requires them to develop a deep understanding of A* and heuristics and gives them an introduction to incremental heuristic search, a new area of search not yet covered in standard textbooks. The students then use Adaptive A* to move game characters in initially unknown gridworlds to a given target location (Koenig and Yeoh 2008).

Any-Angle Path Planning with Variants of A:

The students need to code A* and extend it to the recently developed any-angle search algorithm Theta*, which requires them to develop a deep understanding of A* and heuristics and gives them an introduction to any-angle search, a new area of search not yet covered in standard textbooks. The students then use Theta* to plan any-angle paths for game characters from a given start location to a given target location (Koenig, Daniel, and Nash 2008). Gesture Recognition with

Neural Networks The students need to use neural networks to recognize user gestures for video games, which requires them to develop an understanding of the back-propagation algorithm (Zheng and Koenig 2010), see Figure 1. This project extends a project from Tom Mitchell's Machine Learning book (Mitchell 1997).

Pinball Project

The standard computer science education tends to teach students only about software but not about interfacing it to mechanical systems. It thus does not prepare students well for robotics, which requires at least basic knowledge of electronics, signal generation, embedded systems, communication protocols, interface programming or real-time programming. Designing pinball games can be used for this purpose since pinball machines are simple (although rather unusual) robots. They contain actuators (such as solenoids), sensors (such as switches) and visual outputs (such as lights). We therefore developed a hardware and software interface between a PC and a solid-state pinball machine, see Figure 2. The students of the small pilot CS499 class “Designing and Implementing Games on Pinball Machines” at the University of Southern California then designed and implemented Pinhorse, a simple pinball game that features a true multi-player mode where each player directly influences the game of the other player. Pinhorse is a proof of concept game that demonstrates what can be accomplished with such an interface (Wong et al. 2010). To the best of our knowledge, this is the first time that anyone has managed to control an existing pinball machine completely, although others have tried before (Clark 1997; Bork 2005). We have recently made our interface available for research and teaching purposes and are now thinking about integrating artificial intelligence techniques into pinball games, for example, to adapt the difficulty of the game to the abilities of the players. More information can be found on our web pages at idm-lab.org/pinball together with an 11-minute YouTube video that demonstrates the features of Pinhorse.

Figure 1: Start of Project Text (Zheng and Koenig 2010)

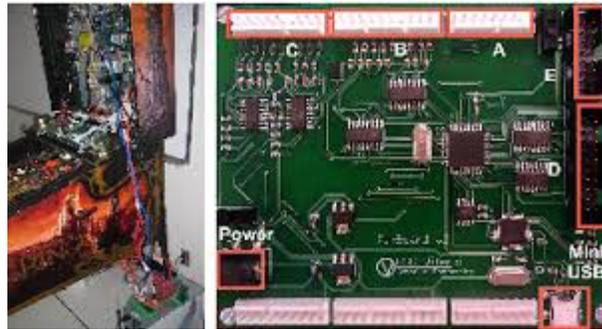


Figure 2:pinball Interface (1st and 2nd Generation)

Conclusion:

AI is at the centre of a new enterprise to build computational models of intelligence. The main assumption is that intelligence (human or otherwise) can be represented in terms of symbol structures and symbolic operations which can be programmed in a digital computer. There is much debate as to whether such an appropriately programmed computer would be a mind, or would merely simulate one, but AI researchers need not wait for the conclusion to that debate, nor for the hypothetical computer that could model all of human intelligence. Aspects of intelligent behaviour, such as solving problems, making inferences, learning, and understanding language, have already been coded as computer programs, and within very limited domains, such as identifying diseases of soybean plants, AI programs can outperform human experts. Now the great challenge of AI is to find ways of representing the commonsense knowledge and experience that enable people to carry out everyday activities such as holding a wide-ranging conversation, or finding their way along a busy street. Conventional digital computers may be capable of running such programs, or we may need to develop new machines that can support the complexity of human thought.

References

- 1.Bork.J, 2005. Controlling a pinball machine using Linux. Linux Journal 139.
- 2.Clark.D,1997. Progress toward an inexpensive real-time testbed.

3. The pinball player project. In Proceedings of the Real-Time Educational: Second Workshop, 72–79. Koenig, S., and Yeoh, W. 2008. A project on fast trajectory replanning for computer games for 'introduction to artificial intelligence' classes.
4. Technical report, Department of Computer Science, Daniel, K.; and Nash, A. 2008. A project on any angle path planning for computer games for 'introduction to artificial intelligence' classes. Technical report, Department of Computer Science Mitchell, T. 1997. Machine Learning. McGraw Hill. Wong, D.; Earl, D.; Zyda, F.; and Koenig, S. 2010.
5. Teaching robotics and computer science with pinball machines. In Proceedings of the AAAI Spring Symposium on Educational Robotics and Beyond: Design and Evaluation. Zheng, X., and Koenig, S. 2010. A project on gesture recognition with neural networks for 'introduction to artificial intelligence' classes.
6. Technical report, Department of Computer Science . Zyda, M., and Koenig, S. 2008. Teaching artificial intelligence playfully.

ML Integration with Everyday Data

Authors: Fiza Tarannum and Neha Ansari

Fiza.tarannum@infosys.com and neha.ansari@infosys.com

Abstract

Although machine learning (ML) offers many advantages when it comes to predicting outcomes and identifying with novel insights, collecting every day data is a challenging task. It is an important aspect to have a comprehensive study of data collection from a data management point of view. Data collection largely consists of data acquisition, data labeling, and improvement of existing data(preparation) or models. A roadmap to data collection in both real-time and offline are getting develop which can be further extended by storing them in google drive and further using it for sentiment analysis

Overview

Ideally, survey or research produces gigantic amounts of data. Later it becomes a part of the machine learning datasets. Those are further used to build models that aim to perform sentiment analysis of the data.

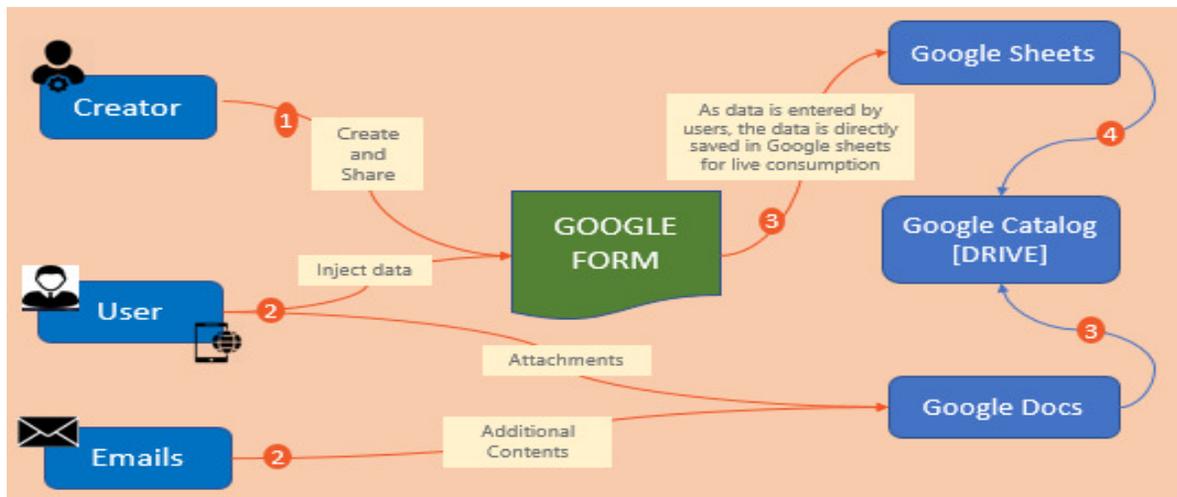
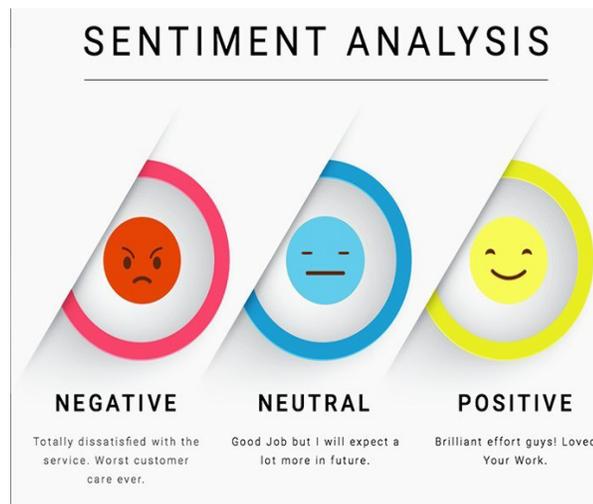
Of course, it is hard to store this amount of data on any physical storage, therefore a better approach is to store them in Google Cloud and save in Google Drive for processing.

A major focus of Machine Learning research is to automatically learn to recognize the polarity of sentences that is extracted from the text of reviews. Machine learning algorithms can achieve high accuracy for classifying sentiment of text based on Negativity, Neutrality and Positivity

Real World Scenario and Solution

In order to create a good Machine Learning System, it requires Data preparation capabilities, Algorithms - basic and advanced, Automation and iterative processes, Scalability and Ensemble modeling.

The data is stored in google drive and is extracted/read for further use in order to perform sentiment analysis analytic and report generation



DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)- 26 to 28 July 2019

The data is collected from several sources through google form and injected into drive as google spreadsheets. The Google Sheets are stored in Google Drive.

The data is pre-processed before using for sentiment analysis. By pre-processing it is meant, removal of Numbers, Punctuation, Lowercase, Stopwords and Stemming



Once the data is pre-processed it is then used to form sensitive analysis which produces two outcomes, Negative words (above picture) and Positive words (left picture).

The processing of data can be in both, real-time (using developer console) as well as offline (by csv). Thus, broadening the scope of data extraction, storage and scalability

Conclusion

An approach of using machine Learning as a platform to analyze one's sentiments, attitude or emotion towards entities or topic. Google Sheets that enables users to collaborate in real-time, is used to store data from which data can be exported in real-time as well as offline.

CROANN: chemical reaction optimization of artificial neural networks for effective forecasting of crude oil prices time series

Sarat Chandra Nayak

Department of Computer Science and Engineering,

CMR College of Engineering & Technology, Hyderabad - 501401, India Email:

saratnayak234@gmail.com; drsaratchandranayak@cmrcet.org

Abstract- Accurate forecasting of crude oil price time series is a difficult job as the series are extremely nonlinear and chaotic in nature. Also, they are affected by several economic as well as global scenarios. Various statistical and econometric models are suggested in dealing with this problem. However, they are able to capture the complex correlations between the data points on the time series up to certain extent. Computational intelligence approach such as artificial neural networks (ANN) are promising alternative to crack such problem. Performance of ANN models is solely depends upon its training. The adjustment of neuron weight and bias is the key factor of ANN training which requires intensive human interventions. To circumvent the limitations of gradient based ANN training, several nature-inspired optimization techniques are proposed. Chemical reaction optimization (CRO) is a recently developed metaheuristic inspired from the phenomenon of natural chemical reaction, which posses the characteristics of faster convergence, and finding the global optima. The intension of this article is to design an effective forecast by hybridizing the advantages of ANN and CRO. The hybrid model employed CRO for optimization of weight and bias vector of an ANN hence termed as CROANN. The proposed model is evaluated on prediction of crude oil price time series. Extensive simulation results and comparative performance analysis suggested the suitability of the proposed model.

Keywords: crude oil prices forecasting; artificial neural network; chemical reaction optimization; gradient descent based training; genetic algorithm; differential evolution

1. Introduction

In current global economic scenario the price of crude oil has a vital role in the economical growth of a country. The crude oil price has been facing arbitrary changes in its movement due to multiple socio- economical as well as political factors. The random fluctuations in the crude oil price time series make its forecasting difficult. Even a nominal change that occurs to the crude oil price can give a tremendous impact to the petroleum price, crude oil products and the global economy. The volatility in the crude oil price is mainly affected by factors like population, demand and supply, political climates, international relationship [1]. Thus an accurate and efficient prediction tool is crucial for crude oil price forecasting. Advances in computational intelligence techniques including artificial neural networks are used as better alternative to this domain [2]. Artificial neural network based models are applied successfully to forecast the crude oil price [3 - 6].

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

Artificial Neural Network (ANN) has the analogy with the thinking capacity of human brain and thus mimicking it [7 - 9]. The ANN can imitate the process of human behavior and solve nonlinear problems, which have made it popular and are widely used in calculating and predicting complicated systems. These are considered to be an effective modeling procedure for mapping input-output containing both regularities and exceptions as the case of financial time series. These advantages of ANN attract researchers to forecast financial time series with ANN based models. Dealing with uncertainty and nonlinearity associated with financial time series with ANN based forecasting method primarily involves recognition of patterns in the data and using such patterns to predict future event.

The adjustment of neuron weight and bias of ANN is the key factor of ANN training and requires frequent human interventions. The performance of ANN solely depends upon the adjustment of weight and bias vectors. To circumvent the limitations of gradient descent based ANN training, large number of nature and bio-inspired optimization techniques are proposed and applied [10]. Evolutionary computing techniques are based on the behavior of nature. Normally these algorithms are motivated by biological evolution and termed as evolutionary algorithms of metaheuristic. The ideas of imitating concepts from nature have great potential in developing algorithms to solve engineering problems. In recent past, applications of these techniques have achieved popularity in wide area of engineering, computer science, medicine, economics, finance, social networks and so on. Their performance depends upon several algorithm specific control parameters and there is no single technique performing well on all problems. Evolutionary training algorithms such as GA [11], PSO [12], DE [13], Ant colony optimization [14], CRO [15, 16] etc. are capable of searching optimal solutions better than gradient descent based search techniques.

However, their efficiency mainly depends on fine-tuning different learning parameters [17 - 23]. With the purpose of landing at the global optimum, an algorithm requires suitable selection of several learning parameters that ultimately makes the use intricate. Therefore, selection of apposite optimization technique to solve a particular problem involves abundant trial and error methods. This paves a path toward choosing an optimization technique requiring fewer learning parameters without compromising the approximation capability. This is the motivation behind choosing CRO for obtaining optimal ANN based forecasting model. Unlike other optimization techniques, CRO does not involve many parameters that must be specified at start rather only defining number of initial reactants is as much as necessary for implementation. As the initial reactants are scattered over feasible global search expanse, optimal solutions can be obtained with little iteration and hence significant reduction in computational time is achieved.

The objective of this study is to design an efficient forecasting model to predict the future price of crude oil prices time series. The proposed CROANN model uses a metaheuristic, i.e. CRO for finding the optimal weight and bias vector of an ANN. A possible ANN architecture is represented as an individual or molecule of CRO. The search process starts with a set of such molecule (potential ANN structure), applies exploration and exploitation through a set of chemical operator and finally lands at global optimal, i.e. best ANN structure on termination. Four real time series data such as daily, weekly, monthly and yearly crude oil prices are used for experimentation. The performance of CROANN is compared with three other models such as gradient descent based ANN (GD-ANN), genetic algorithm based ANN (GA-ANN), and differential evolution based ANN (DE-ANN).

The article is arranged as follows. An introduction about crude oil price prediction, methods used for this and ANN optimization is given by Section 1. Methods and materials are explained in Section 2. Proposed model is presented in Section 3. Experimental results are summarized in Section 4. Section 5 gives the concluding remarks followed by a list of relevant references.

2. Methods and materials

This section presents the basic ANN architecture used in this study and the CRO technique used to optimize the connectionist weight and bias vector for this ANN.

2.1. ANN

The mathematical description of ANN is beyond the scope of this chapter. We present only the description of ANN model used in this study. ANN architecture with one hidden layer of neurons is used as the base neural architecture as shown in Figure 1. Since there is no rule to choose the optimal number of layer and neurons, we choose them on experimental basis.

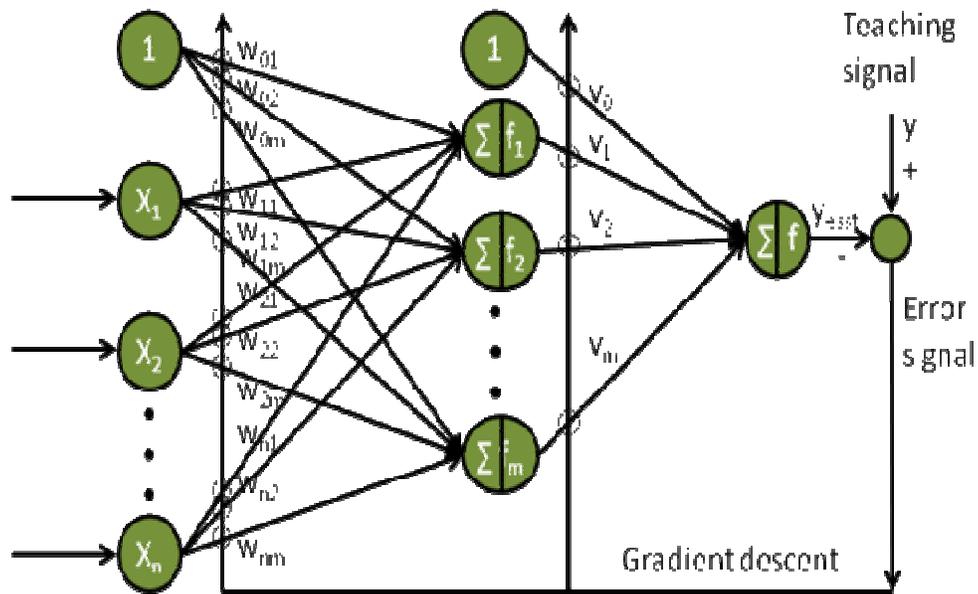


Figure 1 Architecture of 3-layer ANN

The error correction learning in this case is supervised learning, i.e. the target response for the system is presented at the output neuron. This model consists of a single output unit to estimate one-day-ahead data point in the financial time series. The neurons in the input layer use a linear transfer function, the neurons in the hidden layer and output layer use sigmoidal function as follows:



$$y_{out} = \frac{1}{1 + e^{-\lambda y_{in}}} \quad (1)$$

where y_{out} is the output of the neuron, λ is the sigmoidal gain and y_{in} is the input to the neuron. The first layer corresponds to the problem input variables with one node for each input variable. The second layer

is useful in capturing non-linear relationships among variables. At each neuron, j in the hidden layer, the weighted output Z is calculated using Eq. 2.

$$Z_j = b_j + \sum_{i=1}^n w_{ij} x_i \quad (2)$$

where x_i is the i^{th} component of input vector, w_{ij} is the synaptic weight value between i^{th} input neuron and j^{th} hidden neuron and b_j is the bias value and f is a nonlinear activation function. The output y_{out} at the single output neuron is calculated using Eq. 3.

$$y_{out} = b_o + \sum_{j=1}^m w_{oj} Z_j \quad (3)$$

where w_{oj} is the synaptic weight value from j^{th} hidden neuron to output neuron, Z_j is the weighted sum calculated as in Eq. 2, and b_o is the output bias. This output is compared to the target output and the error is calculated by using Eq. 4.

$$e = t - y_{out} \quad (4)$$

where e is the error signal, t is the target signal for i^{th} training pattern and y_{out} is the calculated output for i^{th} pattern. The error signal e and the input vector are employed to the weight update algorithm to compute the optimal weight vector. During the training, the network is repeatedly presented with the training vector and the weights as well as biases are adjusted by training algorithm till the desired input-output mapping occurs. The objective is to minimize the total error as in Eq. 4 with an optimal set of weight and bias vector of the ANN.

1.1. CRO

CRO is a metaheuristic proposed [15, 16] inspired from natural phenomena of chemical reaction. The concept mimics properties of natural chemical reaction and slackly couples mathematical optimization techniques with it. A chemical reaction is a natural phenomenon of transforming unstable chemical substances to a stable one through intermediate reactions. A reaction starts with unstable molecules with excessive energy. The molecules interact with each other through a sequence of elementary reactions and producing some products with lower energy. During a chemical reaction the energy associated with a molecule changes with the change in intra-molecular arrangement. Finally it becomes stable at one point called as equilibrium point. Termination condition is checked by performing chemical equilibrium (inertness) test. If the newly formed reactant has better fitness value, it is included to the reactant pool and the worse one is excluded. Otherwise a reversible reaction is applied. The overall process of ACRO is depicted in Figure 2.

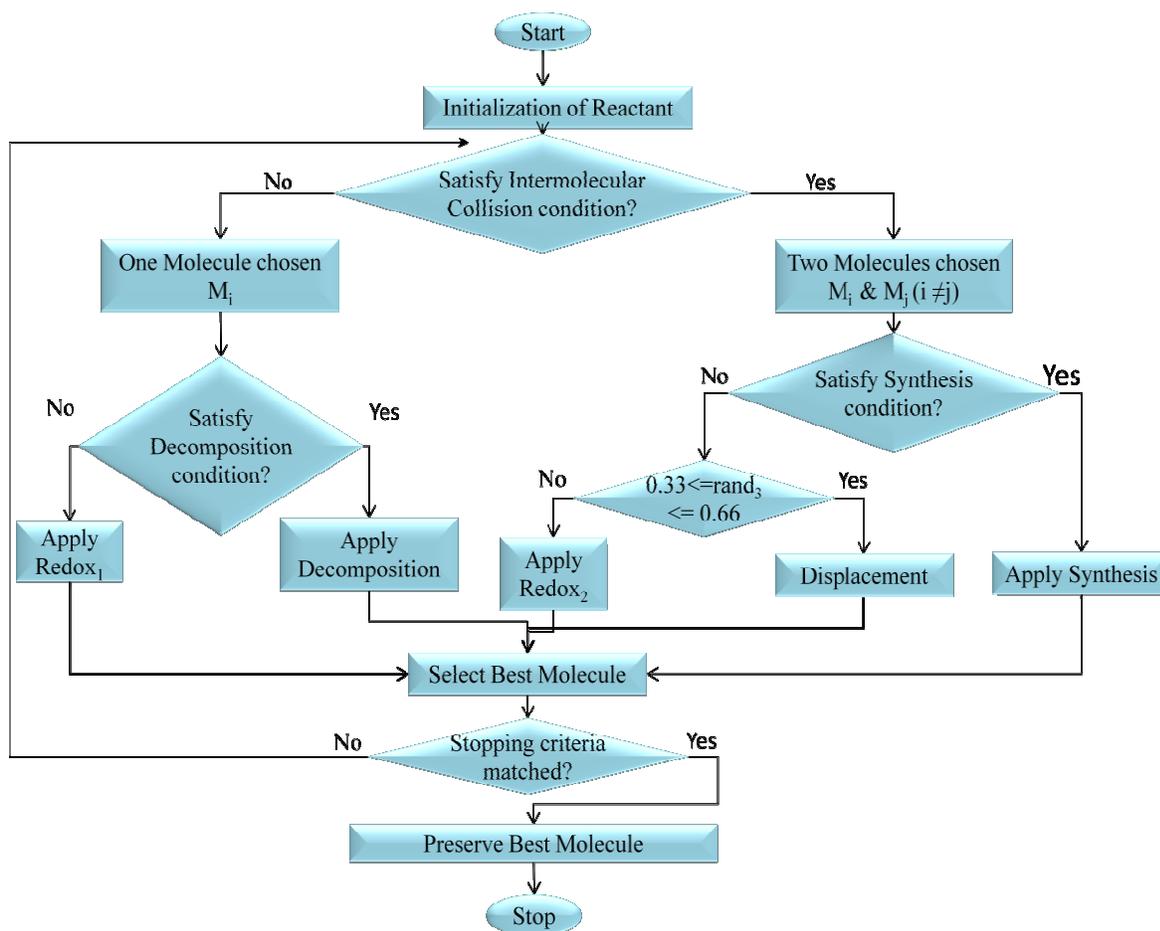


Figure 2 Process of chemical reaction optimization

2. Proposed CROANN based forecasting

In the proposed model CRO is used to train the ANN model. As stated earlier, a molecule (individual) of CRO can be viewed as a potential weight and bias vector for the ANN in the search space. At beginning, a set of such molecule called as reactant is initialized and for each such reactant five chemical operators as shown in Figure 3 are applied to perform the search process. The exploration as well as exploitation of the search space is achieved by these operators. The molecules are then evaluated in terms of error signal generation. The molecule with lowest error signal is considered as the best fit molecule. The selection process is then carried out with inclusion of better fit molecules. The above process continues till an optimal reactant found and the search process then terminates. The best molecule is the optimal weight and bias vector for the ANN model. The high level description of CROANN based forecasting is presented by Figure 3.

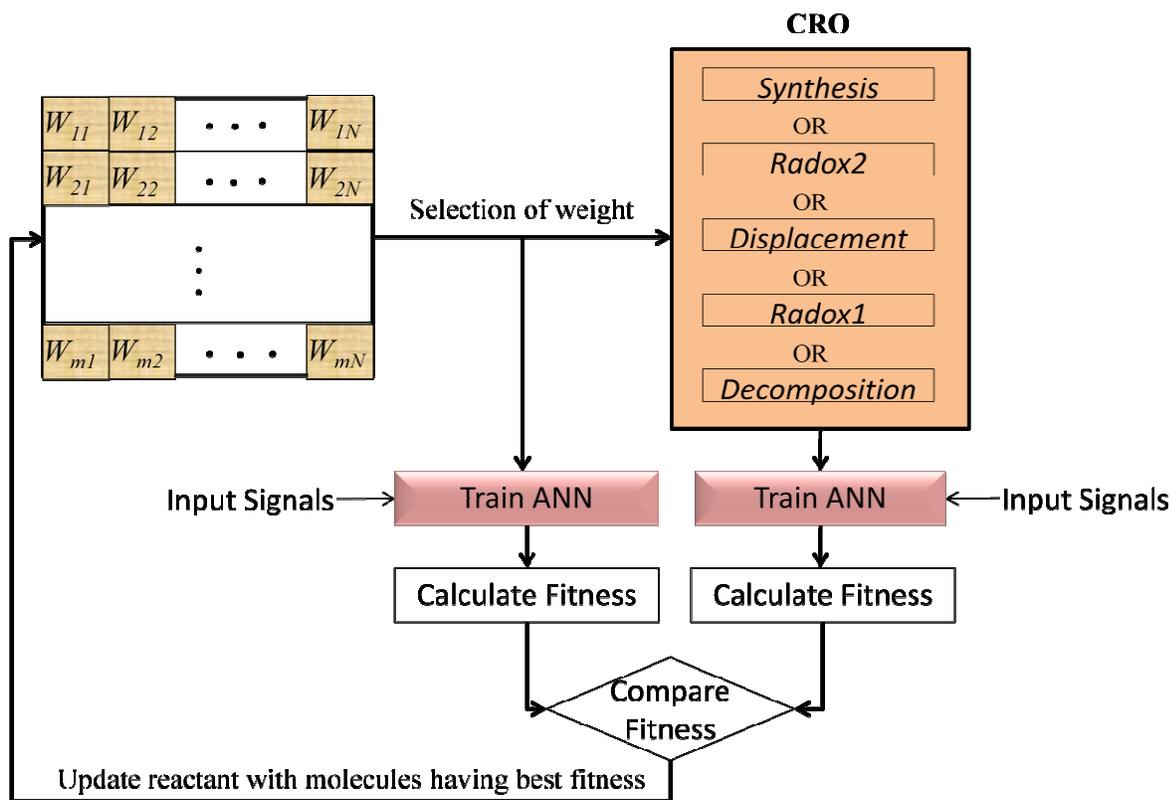


Figure 3 CROANN based forecasting

3. Experimental results and analysis

This section explains about the experimental data, experimental setup, input selection and normalization, performance metrics, experimental results, comparative study and result analysis.

3.1. Experimental data

The crude oil prices (Dollars per Barrel) are retrieved from US Department of energy: Energy Information Administration web site: <http://www.eia.doe.gov/> during the period April 1983 to July 2019. The crude oil price series are shown by Figure 4 - 7. The information about the dataset and descriptive statistics are summarized in Table 1 and Table 2 respectively. All the experiments are carried out in MATLAB-2015 environment, with Intel® core TM i3 CPU, 2.27 GHz processing and 2.42 GB memory size.

Table 1 The crude oil price time series

S . N o .	Crude oil price dataset	Period	No. of data points
1	Daily Cushing, OK Crude Oil	April 04	91

DST Sponsored National Conference on Recent Advancements on Computer Science
(CONRACS 2019)- 26 to 28 July 2019

	Dataset	1983 – July 02 2019	0 5
2	Weekly Cushing, OK Crude Oil Dataset	April 08 1983 – June 28 2019	1 8 9 1
3	Monthly Cushing, OK Crude Oil Dataset	April 1983 – May 2019	4 3 4
4	Annual Cushing, OK Crude Oil Dataset	1983 - 2018	3 6

DST Sponsored National Conference on Recent Advancements on Computer Science
(CONRACS 2019)- 26 to 28 July 2019

Table 2 Descriptive statistics from crude oil price time series

Crude oil price dataset (Dollars per Barrel)	Statistics					
	Min.	Max.	Mean	Std dev.	Skewness	Kurtosis
Daily Cushing, OK Crude Oil Dataset	10.4200	145.2900	42.9828	28.5640	0.9953	2.8829
Weekly Cushing, OK Crude Oil Dataset	11.0900	142.6000	42.8949	28.5241	1.0011	2.8921
Monthly Cushing, OK Crude Oil Dataset	11.3100	132.0000	42.8508	28.5115	0.9969	2.8629
Annual Cushing, OK Crude Oil Dataset	14.4000	99.7500	42.6214	28.0382	0.8772	2.3527

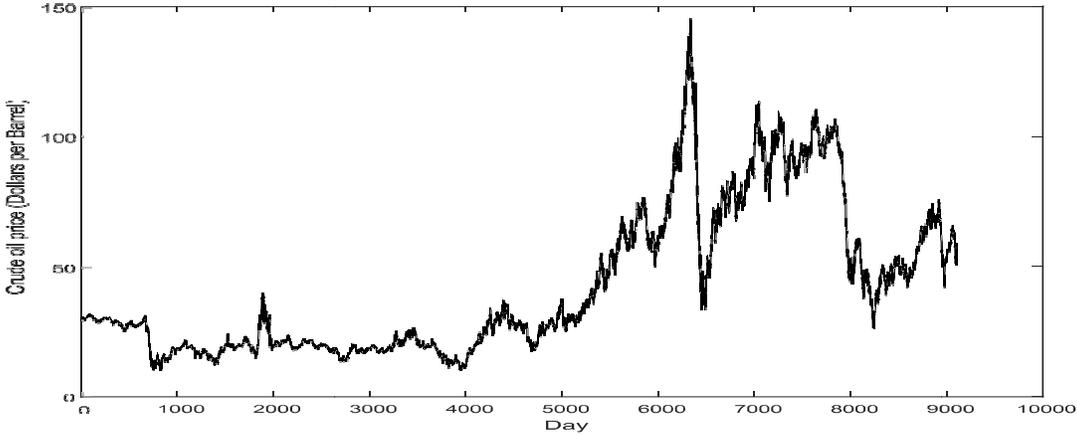


Figure 4 Daily crude oil price data

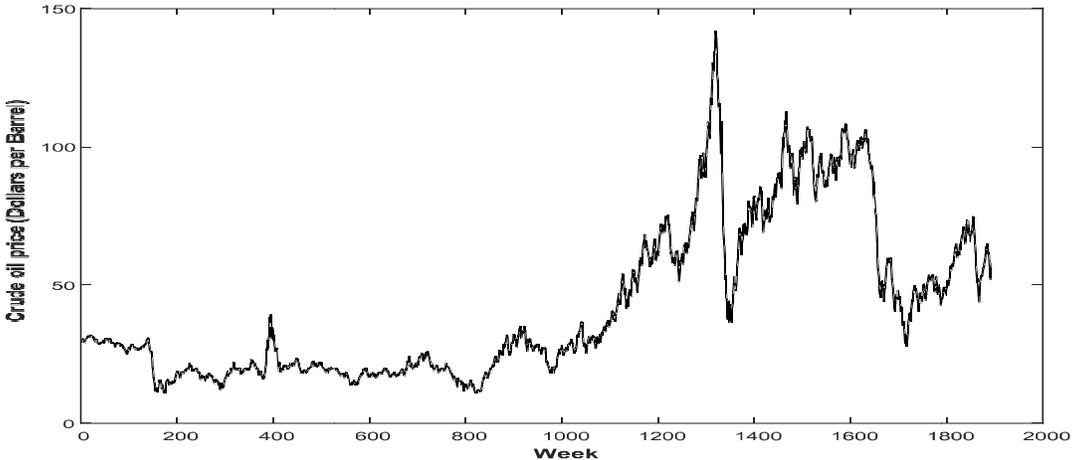


Figure 5 Weekly crude oil price data

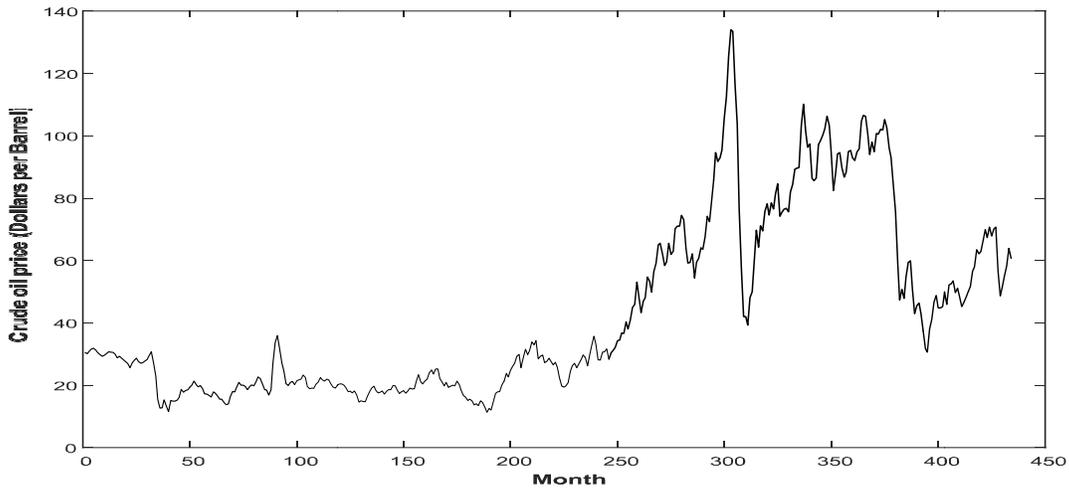


Figure 6 Monthly crude oil price data

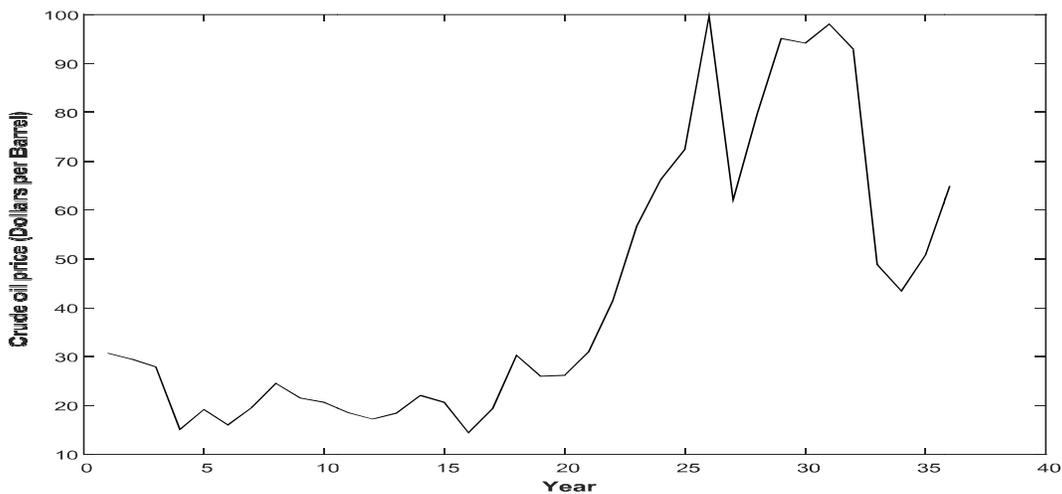


Figure 7 Yearly crude oil price data

3.2. Input selection and normalization

A sliding window of fixed size is used for selecting input for the forecasting model. On each sliding of the window, a new data point is incorporated and the oldest one is discarded. The window moves through whole financial time series and the selection of size of window is a matter of experimentation. For an example, the training and test patterns generated for one-step- ahead forecasting by sliding window technique is presented below. Here the bed length (window size) is written as *blen* and training length is represented as *l*. In general, the training data with window size = *blen* and training length *l* is:



And the respective test data is



3.3. Experimental results and discussion

All the three financial time series are normalized before feeding them to the ANN model [24]. The normalized data are then used to form a training bed for the network model. The model is simulated for 20 times for each training set and the average error is considered for comparative analysis. Since each time the sliding window moves one step ahead, only one new closing price data has been included into the training set. So there may not be significant change in nonlinearity behavior of the training data set. For that reason, instead of considering another random weight set (i.e. set of locations), we used the previously optimized weight set for the successive training. In this way, after the first training set, the number of iteration has been fixed to a small value, hence significant reduction in training time. During experimentation, different possible values for the model parameters were tested and best values are recorded.

To validate the performance of the proposed model, we developed four other models such as GD-ANN, GA-ANN and DE-ANN. All the four models are fed with the same training and test dataset. The error statistics in terms of minimum, maximum, average and standard deviation generated by four models from four datasets are summarized in Table 3. The estimated crude oil prices against actual prices are plotted in Figure 9 – 12. From Table 3 it can be observed that the CROANN always generated better error statistics from all datasets. The average error generated by the proposed model from daily, weekly, monthly and annual datasets are 0.0048, 0.0037, 0.0084, and 0.0159 respectively. These values are quite better than that of competitive models. Considering the performance of GD-ANN model as least, we calculated the performance gain of the three hybrid models over GD-ANN as in Eq. 5. The result is shown by Figure 8. It is observed that the performance gain of CROANN is much better than other two hybrid models.

$$\text{Performance Gain} = \frac{\text{Error}_{GD-ANN} - \text{Error}_{Hybrid}}{\text{Error}_{GD-ANN}} \times 100\% \quad (5)$$

Table 3 Forecasting errors from all models

Crude oil price dataset	Error Statistic	Forecasting Models			
		GD - AN N	GA - AN N	DE - AN N	CR OA NN
Daily Crude Oil Price Dataset	Minimum	0.00008	0.00007	0.00007	0.00005
	Maximum	0.0497	0.0482	0.0396	0.0375
	Average	0.0098	0.0083	0.0087	0.0048
	Standard Deviation	0.0081	0.0075	0.0075	0.0019

DST Sponsored National Conference on Recent Advancements on Computer Science
(CONRACS 2019)- 26 to 28 July 2019

Weekly Crude Oil Price Dataset	M i n i m u m	0 . 0 0 0 5	0.00 003	0.00 003	0.0 000 1
	M a x i m u m	0 . 0 4 6 5	0.04 02	0.04 36	0.0 426
	A v e r a g e	0 . 0 1 2 9	0.00 96	0.00 92	0.0 037
	S t a n d a r d D e v i a t i o n	0 . 0 1 5 5	0.00 75	0.00 77	0.0 038
Monthly Crude Oil Price Dataset	M i n i m u m	0 . 0 0 0 7	0.00 004	0.00 004	0.0 000 1

DST Sponsored National Conference on Recent Advancements on Computer Science
(CONRACS 2019)- 26 to 28 July 2019

	M a x i m u m	0 . 0 5 6 1	0.04 37	0.04 22	0.0 387
	A v e r a g e	0 . 0 1 5 0	0.01 33	0.00 96	0.0 084
	S t a n d a r d D e v i a t i o n	0 . 0 1 1 2	0.00 85	0.00 76	0.0 024
Annual Crude Oil Price Dataset	M i n i m u m	0 . 0 0 0 5	0.00 005	0.00 006	0.0 000 4
	M a x i m u m	0 . 0 4 5 9	0.03 86	0.03 92	0.0 184
	A v	0 .	0.01 66	0.01 75	0.0 159

e r a g e	0 2 8 9			
S t a n d a r d D e v i a t i o n	0 . 0 0 4 9	0.00 43	0.00 47	0.0 038

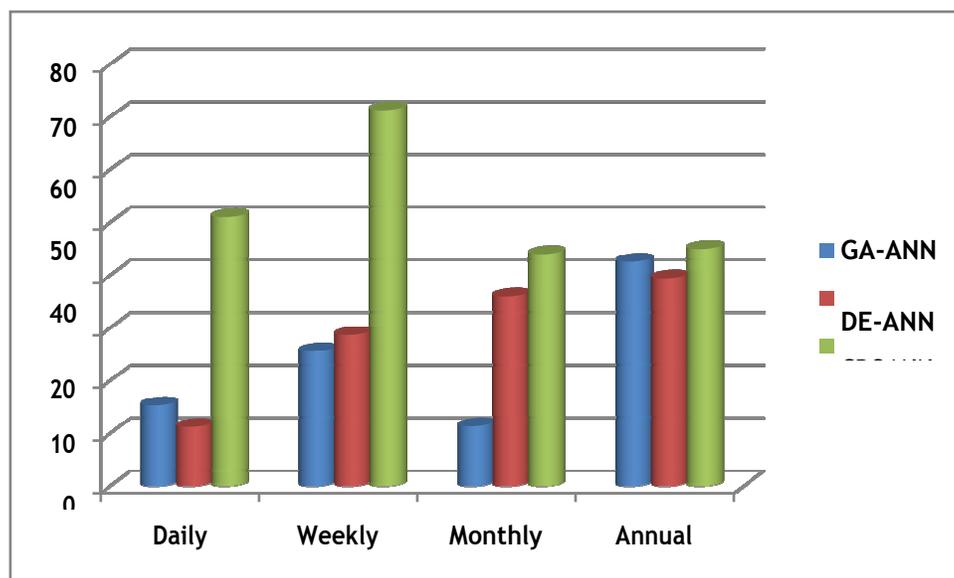


Figure 8 Performance gain of GA-ANN, DE-ANN, and CROANN over GD-ANN

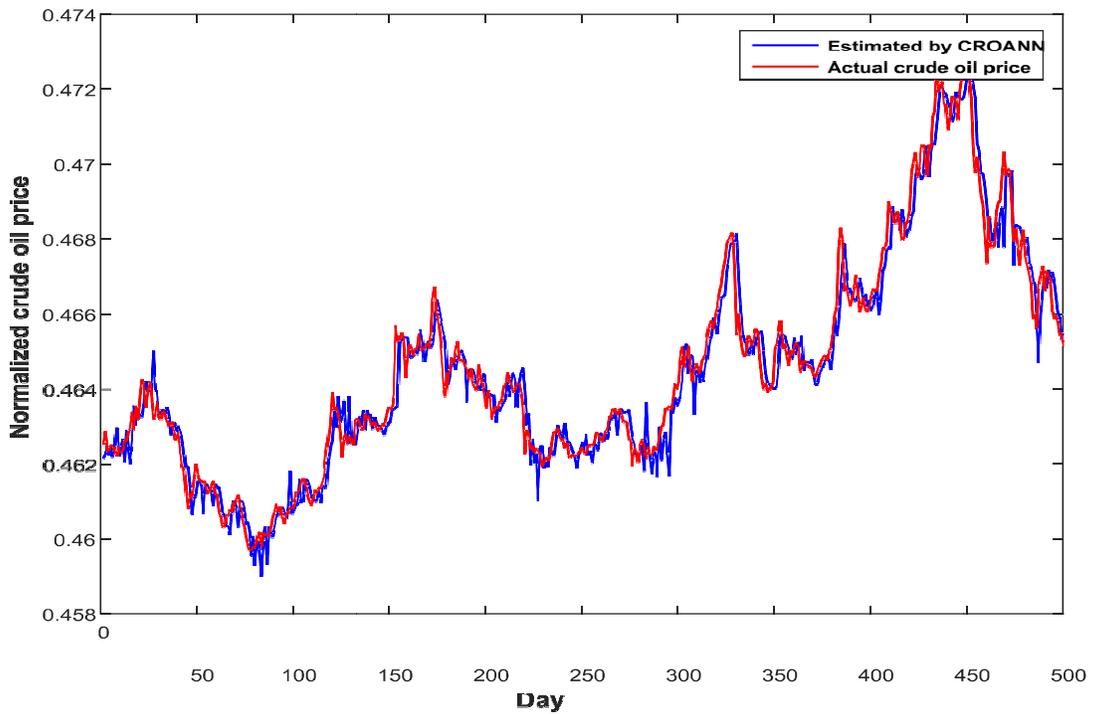


Figure 9 Estimated v/s actual daily crude oil prices

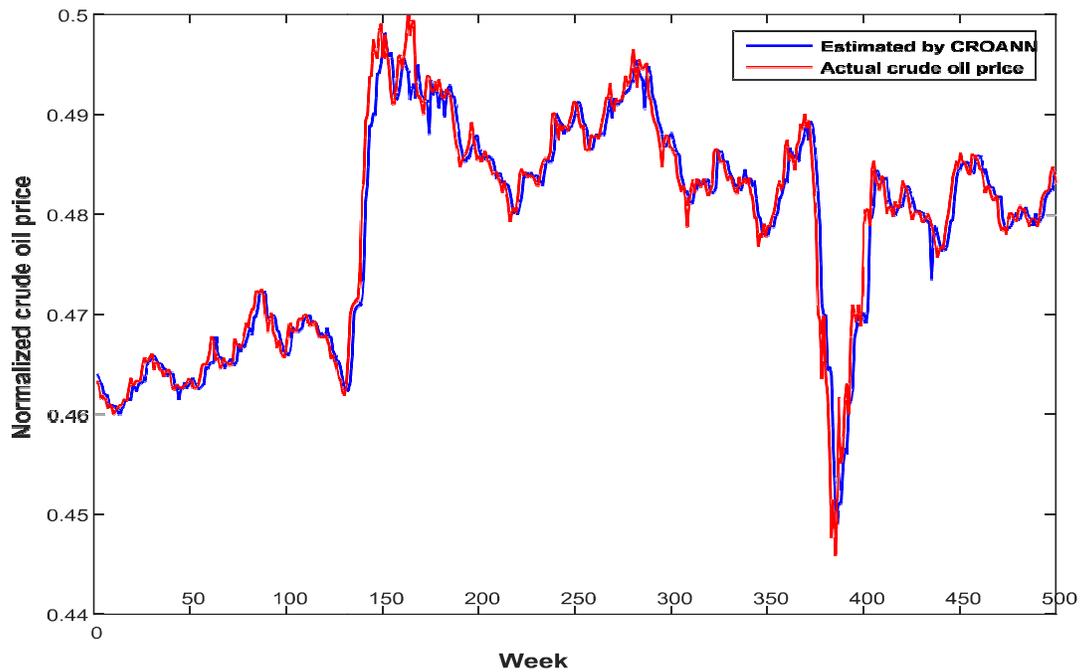


Figure 10 Estimated v/s actual weekly crude oil prices

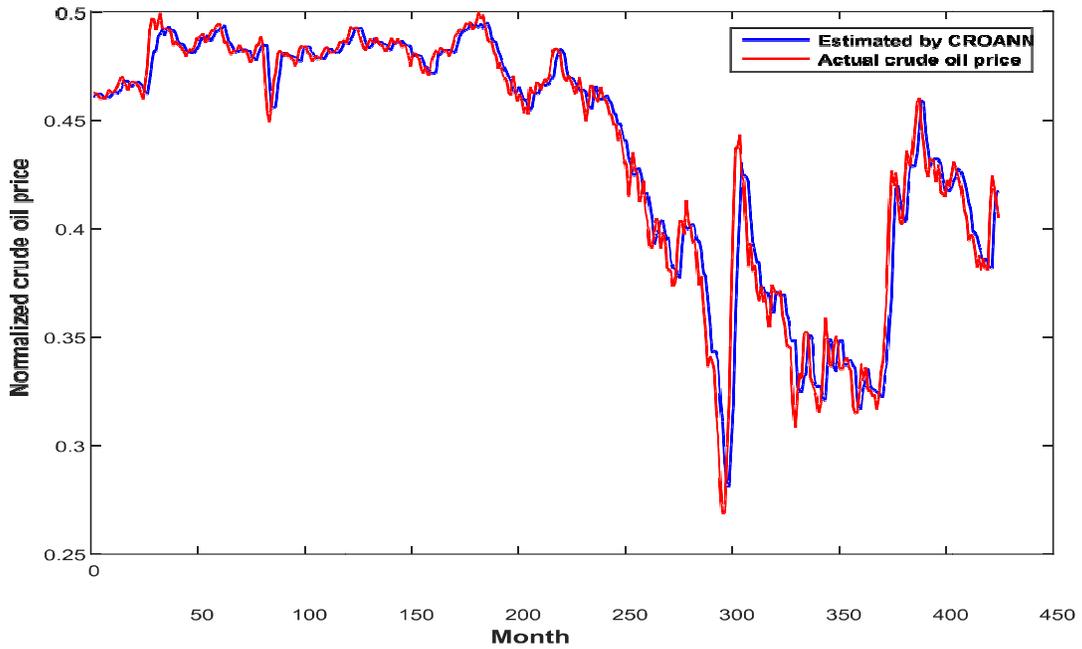


Figure 11 Estimated v/s actual monthly crude oil prices

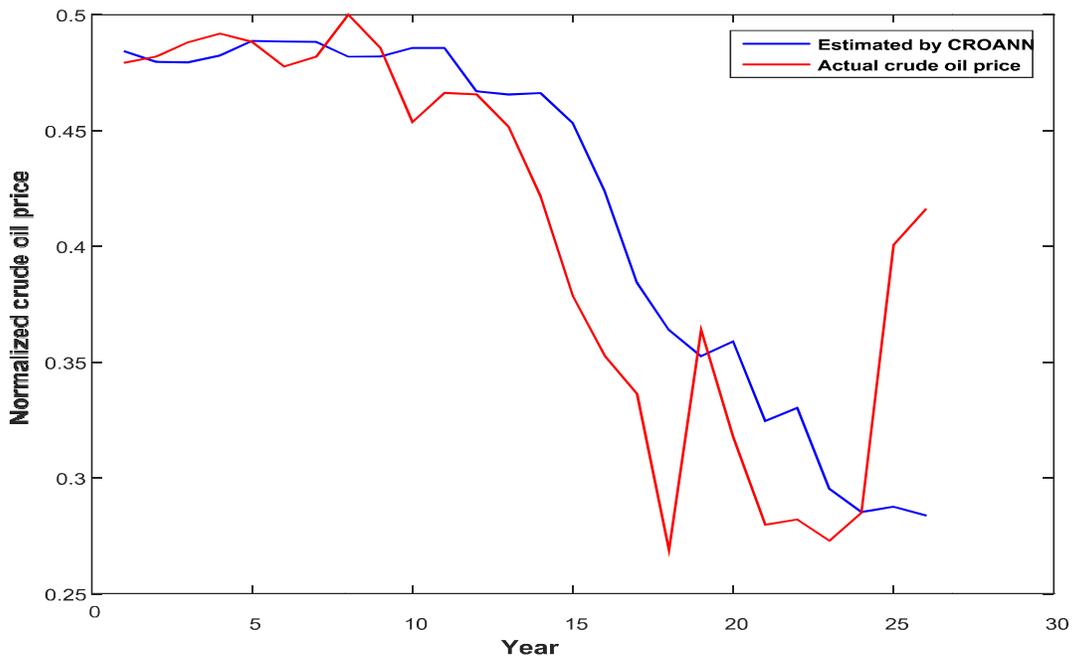


Figure 12 Estimated v/s actual annual crude oil prices

4. Conclusions

In order to capture the trend the crude oil prices time series follow, this article proposed a hybrid computational intelligent forecasting model termed as CROANN. The model synergies the advantages of CRO and ANN and able to capture the nonlinearities associated with crude oil prices time series. CRO was employed to optimize the weight and bias vector of a three layer ANN to minimize the forecasting error. CRO is able to overcome the well known limitations of gradient descent based training such as slow convergence, prone to local minima etc. in an efficient manner. Four real crude oil prices time series are considered for experimental work. The performance of the proposed model is compared with that of GD-ANN, GA-ANN and DE-ANN trained in a similar way and found much better. Also, among the three hybrid models the CROANN had shown better performance gain over other. This work may be extended for other time series and other nature inspired optimization techniques may be explored to search the ANN parameters.

Reference

- [1] Y. Nelson, S. Stoner, G. Gemis, & H. Nix, “Results of Delphi VIII Survey of Oil Price Forecasts,” Energy Report, California Energy Commission, 1994.
- [2] A modified neural network model for predicting the crude oil price, Intellectual Economics, 2016.
- [3] A review on artificial intelligence methodologies for the forecasting of crude oil price, Intelligent automation and soft computing, 2016.
- [4] Forecasting crude oil price using artificial neural networks: a literature survey, 2015.
- [5] Forecasting energy market indices with recurrent neural networks: case study of crude oil price fluctuations, Energy, 2016
- [6] Evolutionary neural network model for west texas intermediate crude oil price prediction, Applied Energy, 2015
- [7] S. Haykin, Neural Networks and Learning Machine, PHI, ISBN -978-81-203-4000-8, 2010.
- [8] V. Kecman, Learning and Soft Computing, Pearson Education, ISBN 81-317-0305-3, 2006.
- [9] S. Rajasekaran and G. A. Vijayalakshmi Pai, Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Application, PHI, ISBN-978-81-203-2186-1, 2007.
- [10] N. Shadbolt, “Nature-inspired computing,” IEEE Intell Systems, 19(1), pp. 2-3, 2004.
- [11] D. E. Goldberg, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, Reading, MA, USA, 1989.

- [12] J. Kennedy, R.C. Eberhart, Swarm intelligence. Morgan Kaufmann, San Francisco, 2001.
- [13] K. Price, R. Storn, J. Lampinen, “Differential evolution: a practical approach to global optimization,” Springer, Berlin, 2005.
- [14] M. Dorigo, T. Stutzle, Ant colony optimization. The MIT Press, Cambridge, MA, USA, 2004.
- [15] Lam, A. Y., & Li, V. O. (2010). Chemical-reaction-inspired metaheuristic for optimization. *IEEE Transactions on Evolutionary Computation*, 14(3), 381-399..
- [16] Alatas, B. (2012). A novel chemistry based metaheuristic optimization method for mining of classification rules. *Expert Systems with Applications*, 39(12), 11080-11088.
- [17] Nayak, S. C., Misra, B. B., & Behera, H. S. (2018). ACFLN: artificial chemical functional link network for prediction of stock market index. *Evolving Systems*, 1-26.
- [18] Nayak, S. C., Misra, B. B., & Behera, H. S. (2015). Artificial chemical reaction optimization of neural networks for efficient prediction of stock market indices. *Ain Shams Engineering Journal*
- [19] Nayak, S. C., Misra, B. B., & Behera, H. S. (2013, September). Hybridizing chemical reaction optimization and artificial neural network for stock future index forecasting. In *2013 1st International Conference on Emerging Trends and Applications in Computer Science* (pp. 130-134). IEEE.
- [20] Nayak, S. C., Misra, B. B., & Behera, H. S. (2017). Improving performance of higher order neural network using artificial chemical reaction optimization: a case study on stock market forecasting. In *Nature-Inspired Computing: Concepts, Methodologies, Tools, and Applications* (pp. 1753-1780). IGI Global.
- [21] Behera, A. K., Nayak, S. C., Dash, C. S. K., Dehuri, S., & Panda, M. (2019). Improving software reliability prediction accuracy using CRO-based FLANN. In *Innovations in Computer Science and Engineering* (pp. 213-220). Springer, Singapore.
- [22] Sahu, K. K., Sahu, S. R., Nayak, S. C., & Behera, H. S. (2016). Forecasting foreign exchange rates using CRO based different variants of FLANN and performance analysis. *International Journal of Computational Systems Engineering*, 2(4), 190-208.
- [23] Dash, C. S. K., Behera, A. K., Nayak, S. C., Dehuri, S., & Cho, S. B. (2019). An Integrated CRO and FLANN Based Classifier for a Non-Imputed and Inconsistent Dataset. *International Journal on Artificial Intelligence Tools*, 28(03), 1950013.
- [24] Nayak, S. C., Misra, B. B., & Behera, H. S. (2014). Impact of data normalization on stock index forecasting. *Int. J. Comp. Inf. Syst. Ind. Manag. Appl*, 6, 357-369.

AIRLINE BOT – THE TRANSFORMED EXPERIENCE

*Rincy Mariam Thomas^{*1}, Supriya Punna², Mr. C Kishor Kumar Reddy³ & Dr B V Ramanamurthy⁴*

^{*1,2,3&4}Stanley College of Engineering and Technology for Women, Hyderabad

ABSTRACT

Chatbots are revolutionizing the way customer's talk to businesses and more industries are starting to take notice. Chatbots have presented themselves as a forward-thinking and capable way of elevating the experience of flying commercial. The chatbot can quickly and accurately provide you with directions, answer FAQ's. The airline industry is one marked by fierce competition and a reputation for lackluster customer service. It seems that every day brings some new story about delayed flights or rowdy customers. With prices between competing airlines varying only slightly, service and experience is a significant differentiator. Often, it can be the difference between creating a loyal customer and creating one who spreads their poor experience to other potential passengers. The airline's chatbot sends booking confirmation, 24 hour check in reminder before the flight, and boarding passes are all sent through the chatbot. If you forget your gate, no problem. Airline chatbots can even alert you of delays and gate changes.

I. INTRODUCTION

Chat or speech is one meaningful form of communication between humans. Chat becomes more natural interaction than graphic base interface so will be broadly used in humanizing computer interaction to human. Chatbot worked by interpreting the message that given by the user, and then give response base on captured parsed meaning of the message.

The present day technology promote an up to date internet work and mobile communication which is applying an intense influence on the enlargement course of visitor industry while conducting tourists convenient, quick and brand new travel existence. Mobile travel booking service (MTBS), is an advanced way of trip booking, means that customer can make use of mobile devices to book hotels, resorts, air tickets and other tourism product or duty on the move via GPRS,3G/4G, Wi-Fi and other Wi-Fi network, which is famous with (Yang, Zhong, & Zhang, 2013). In comparison to the customary wired travel booking, the central characteristic of mobile travel booking is that the customer can book tourism products or can have service at anytime from anywhere (Wang & Liao, 2008). Anyways, mobile travel booking as a particular information system, if the user will not use effectively, it gets difficult for the travel service provider to gain profit. So, it is very important for this paper to identify the up to date intentions of the mobile travel booking service. Expectation-confirmation model (ECM) in Information System (IS) is a theoretical framework which is accepted all over to make the users understand the continuance behavior, and many other studies with this model have been expanded (Larsen, Sørrebø, & Sørrebø, 2009; Tang & Chiang, 2010; Lee 2010).

Service design is a holistic field that views service systems with the core principles of human-centeredness and co-creation (Sanders & Stappers, 2008). Keeping in mind about the needs of users and other stakeholders, service design goal is to provide value through service solutions which may contain both physical as well as digital elements. The digital field has improved rapidly especially in the context of digital services, the advancement of technologies and digital capacities has increased in the range of solutions which are available. A service has many ways to use the AI assistant – as a direct interface for the customer, a skill on the service backend supporting the service delivery or an assistant to employees by augmenting their capacities to deliver better service in the encounter with the customer. The AI makes sure that it plays an active role in the service which is offered to it. Services involve several actors (see Latour, 2005) and are based on interactions with touch points, which can be providers for

formulating marketing strategic interactions or products (Segelström & Holmid, 2012). Co-creation is an

important part of services (Holmlid, 2009).

Question answering (QA) systems are identified as an information retrieval system which aims to respond to the natural language queries and also returns the answers instead of the lists of documents. Although Question and Answer (QA) differs from standard information retrieval in the response format, both processes share a lack of interactivity. In the typical information-seeking session the user submits a query and the system returns a result; the session is then concluded and forgotten by the system. It has been argued (Hobbs, 2002) that providing a Question and Answer system with a dialogue interface which would encourage and accommodate the submission of multiple related questions and handle the user's requests for clarification.

In the last few years the variety in the chat bots has increased. Therefore there were many chat bots introduced especially in Facebook as they are the means of communication and they are opted the most as the designing of them will not include more cost.

In the late 90s, the diffusion of the internet initiated a debate concerning the transformation of the holiday-distribution channel. The so-called 'disintermediation effect', postulated that the potential of information and communication technologies and the corresponding reduction of transaction costs, could lead to the emergence of electronic markets for holidays (and holiday components), at the expense of brick-and-mortar tourism intermediaries (i.e. travel agencies and tour operators). As we approach 2020, the development and diffusion of robotics and artificial intelligence in services spark a second debate which gave rise to many questions.

As we know chat bots are the evolving trend of the current generation. In this paper we are mainly discussing about the Airline Chatbot and how its implementation is taking place in this world. Regarding this category the scope is very high because in previous days if we are in a plan to travel we used to book tickets before many days as per the relevant information provided by the Airlines members which is actually a very huge process. But now everything has become online where we can use the chat bots, they help us to solve all our queries about the trip on which we want to go. They give us the appropriate answers for the questions we ask in a natural process of communication. They are improved versions of the chat bots where all the options are placed in a single page where we need not need to browse much. Over all it is an interactive and a simple process where the AI assistants are appointed to fulfill our needs.

Today's most common way of planning holidays is – apart from going to the travel agency – the gathering of information from the internet. Many forms with many questions have to be filled in before you are able to get more information. Often users forget to really input all information, which results in error messages. Especially elder people are not used to work with computers. But exactly this part of society spends money in the field of tourism. That's why it makes sense to develop systems which are easy to control by this target group. Language is the most effective natural form of exchanging information.

On the other hand, internal chat bots may strongly influence and change the future organization communication and collaboration within the company. Apart from all these the answers returned by the chat bot will be understandable and will not be too lengthy and will also be apt to the question which is asked by the customer. It is also not a time taking process. The answers will be displayed on the screen within fraction of seconds. So at last AI is one of the fields with the potential of reshaping service interactions in the near future, of which AI assistants in the current market are visible examples. Chat bots is turning into a necessity and is opted by all the customers as it is uncomplicated.

RELEVANT WORK

Jizhou Huang et al. proposed a paper: "Extracting Chatbots Knowledge from Online Discussion Forums": An Online Discussion Forum is a web society that allows humans to discuss general topics like movies, sports and so on. In this paper, there is a unique approach for extracting a rich-quality pairs of online discussion forums, which

are actually extracted using cascaded frame work. In common this discussion will have seven discussion sections. In which each of it focuses on a specific discussion themes and involves several threads. Comparisons with related networks are also done to compare the manual knowledge building techniques with the manual knowledge building techniques with the most efficient in building a specific domain chatbot. A very brief explanation about cascaded networks and its role is mentioned in the research paper [1].

AM Rahman et al. proposed a paper: “Programming challenges of chatbot: current and future prospective”: The paper firstly tells us how to speak to a chatbot by taking few initial measures which is by giving the user names and then there is chance for the user to develop a conversation with the bot. After this, the needs of the people which they want to satisfy from the bots are satisfied. This scenario is seen commonly in business which provides better experience with less cost. The new techniques are added to the previously existing chatbots as it is a rising trend in this modern world. The people who develop the bots should be in a condition to understand certain qualities such as scalability, stability and flexibility issues. By this the present as well as the future prospective can be reached [2].

M.Dahiya et al. proposed a paper: “A Tool of Conversation”: This paper addresses the design and implement of a chatbot system. How smartly a chatbot can communicate with the users is being discussed but it is only based on the text only chatbot. When questions are being asked the chat bit will respond from a predefined pattern. Mainly a chatbot is built by pattern comparing, which the sequence of the sentence is identified and a saved response pattern is given to the user. It can be simply mentioned like this user -> ask question -> chatbot -> given response -> user. The Facts which are to be noted while designing a chatbot: 1) Selection of OS 2) Selection of Software 3) Cheating a Chatbot 4) Creating a Chat 5) Pattern Matching 6) Simple 7) Conservational and Entertaining. Implementation process involves: 1) Creating a Dialog Box 2) Creating a Database 3) Modules Description [3].

Ayesha Shaikh et al. proposed a paper: “A survey on Chatbot Conversational Systems”. A Chatbot is a human like conversational character. It is computer program which develops a conversation through auditory or textual methods. Its conversation and all the human like skills are due to the Artificial Intelligence. Previous Chatbots use simple keywords and pattern matching methods. For developing number of heuristic rules, language expert knowledge is necessary, these rules maintain the quality of the systems. The paper also mentions about the methods, using dialogue acts and POS- tagged tokens, long term memory and knowledge extractor, experimental results and discussions. So, this paper more deals with the type of conclusion that a chatbot does [4].

Ulrich Gnewuch et al. proposed a paper: “Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction”: This research paper generally deals how the robots as well as the humans response delays to the texted message. Specially, we encounter that when responses are dynamically delayed, the person who is using perceive chatbots are more of human -wish are more socially present and are more satisfied with the overall interaction than when responses are sent near instantly.

Chatbots response time represents a social cue that elicits social responses from users. Dynamic response delay is necessary while there is a case of specific characteristics. Our findings provide more an important initial step towards making chatbot-human interaction more natural [5].

Zhuling Zhong et al. proposed a paper: “Understanding Antecedents of Continuance Intention in Mobile Travel Booking Service”: This paper has become a pioneer of the integrated model for grasping users’ continuance planned behavior towards mobile travel booking service. The outcome of the mobile travel booking services is actually been affected by user satisfaction, perceived usefulness and the subject norm. The more significant among all these is the user satisfaction. In this paper finally the author listed out the strategies and he discussed about the outcomes [6].

Titta Jylkas et al. proposed a paper: “AI Assistants as Non-human Actors in Service Design”: The rapid improvement in technology and the usage of artificial intelligence in the past years are giving new hopes for the creation of service encounters. Now these service encounters are not only mandatorily created for the humans but also for the non-humans. An example for the above sentences is the AI assistants. Therefore in this paper we will be

discussing about the AI assistant by taking it as an example and how this non-human actor will play its role when it comes to service encounters [7].

Silvia Quarteroni et al. proposed a paper: “A Chatbot-based Interactive Question Answering System”: Question answering can be watched on the basis of information retrieval systems whose goal is to respond to the commonly or naturally used language queries by the process of giving back answers rather than a document format. So in this paper we report about the insight into the design, execution and assessment of a chatbot-build dialogue interface [8].

Darius Zumstein et al. proposed a paper: “Chatbots– An Inteactive Technology For Personalized Communication, Transactions And Services:” Chatbots as a produced information, communication and negotiation channel which makes possible for the businesses to extend their target audience by the means of messenger apps like Facebook, WhatsApp or WeChat. Brand new chatbots evolution in customer services and sales are exceptional. For example in the process of booking tickets, to find our destination when we plan for a field trip and all [9].

Papathanassis Alexis et al. proposed a paper: “R-Tourism: Introducing the Potential Impact of Robotics and Service Automation in Tourism:” Both Artificial Intelligence and Robotics are anticipating extending their ‘tipping points’ above the decade. Service automation and digitalization are previously seen in the tourism section. Innumerable application examples and instance of those technologies all over the complete holiday value-chain are defined and their dispersal drivers are dicussed [10].

Xiaojun Shen et al. proposed a paper: “Enhancing e-Commerce with Intelligent Agents in Collaborative e-Communities:” In this paper, we present the planning and enactment of an interdisciplinary research project including a quick-witted agent-based substructure for cooperative ecommerce applications. This architecture will not only allow appeal agent technologies in paperback manners, but it will also assimilate privacy law and codification into its technical scheme [11].

Alexander Maedche et al. proposed a paper: “Advanced User Assitance System”: Information technology (IT) potentiality is increasing at a magnificent pace, but the cognitive aptness of the users is not growing at an identical speed. Thus there is an interstice between the IT which is available and the abilities of the users. There is an experience proof that supports structures intentionally on the contrary [12].

Anand Nayyar et al. proposed a paper: “Virtual Reality (VR) & Augmented Reality (AR) technologies for tourism and hospitality industry”: whose intention is to call attention to top technologies for Tourism and Hospitality with consideration to AR and VR. Virtual Reality (VR) and Augmented Reality (AR) are considered as the most world- changing technologies of 21st Century [13].

Mariusz Krzysztof Żytniewski et al. proposed a paper: “Integration of knowledge management systems and business processes using multi-agent systems”: This paper presents the concept of an original solution ensuring integration of knowledge management systems and business process. The first part of the paper presents current research in the area of integration of software agents within business processes and the processes of knowledge processing. The second part presents the architecture of a software solution designed to support the modelling of business processes and improve these processes. The third part shows an example of using this architecture [14].

Arūnas Miliauskas et al. proposed a paper: “An Approach to Designing Belief-Desire-Intention Based Virtual Agents for Travel Assistance”: is to propose a Belief-Desire-Intention (BDI) architecture-based approach for a virtual agent design. The presented case of a chatbot assistant in a travel domain demonstrates the necessity of the BDI architecture modification. The approach is taken for multiple BDI agent instances with a shared external knowledge base [15].

Markus Berg et al. proposed a paper: “Website Interaction with Text-based Natural Language Dialog Systems”: describes the extension of an existing web-based booking interface in the tourism domain with a natural language

interface. This allows the user to interact with the system in form of a written natural-language-based dialog. The main focus lies in a user-centered, intuitive dialog design, which allows the system to guide the user effectively [16]. Milan van Eeuwen et al. proposed a paper: “Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers”: A research model is proposed based on the Technology Acceptance Model (TAM) and Innovation Diffusion Theory (IDT). Data is collected by means of an online survey among 195 participants. The proposed research model is tested by means of simple regression analysis and results are cross- validated using IBM Watson Analytics. All proposed hypotheses are supported [17].

Anita Nathania P et al. proposed a paper: “Android Based Chatbot and Mobile Application for Tour and Travel Company”: Mobile Applications are rapidly growing segment of global mobile market. This paper involves an application for the android base operating system for a travel agent which will conduct booking transactions for train tickets, airline tickets, hotel, theme park, and tour. This application is integrated with a chatbot, instant messaging applications [18].

Eko Handoyo et al. proposed a paper: “Ticketing Chatbot Service using Serverless NLP Technology”: The contribution of this research is to conduct some scenario that happening in ordering tickets. This research conducts that chatbot can help acts as customer service, based on the conducted scenario and show an F-measure score of 89.65% [19].

Bibek Behera et al. proposed a paper: “Chappie - A Semi-automatic Intelligent Chatbot” : Presently Chappie is being used as a routing agent wherein it can classify the requirement of user into one of the services provided by business based on the first few chats and then transfer it to an agent expert in that service. It uses natural language processing (nlp) to analyse chats and extracts intent of the user with a score similar to the likes of WIT1. Then it uses this information and AIML(Artificial Intelligence Mark-up Language) to make a conversation with the user. This is the marked difference between Chappie and existing chatbots like ALICE(Shawar and Atwell, 2003), which work solely on AIML [20].

II. PROPOSED CHATBOT

Airlines are using bots to deliver boarding passes and other itinerary information to their customers. The latest innovations automatically check your airline ticket to see if your fare has dropped and then negotiate a

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

refund. Online travel agencies are using them to help travelers find better deals on their itineraries.

The airline industry is one marked by fierce competition and a reputation for lackluster customer service. It seems that every day brings some new story about delayed flights or rowdy customers. With prices between competing airlines varying only slightly, service and experience is a significant differentiator. Often, it can be the difference between creating a loyal customer and creating one who spreads their poor experience to other potential passengers.

Chat bots have presented themselves as a forward-thinking and capable way of elevating the experience of flying commercial. While applications (detailed below) vary widely by airline, it is undeniable that these chat bots have arrived—and are not going anywhere anytime soon.

Answering common questions:

Even the most experienced traveler has questions to ask their airline. What gate does my flight depart from? What is the weight limit on checked luggage? When is the next connecting flight to Chicago?

In the past, these types of questions had to be handled by calling an airline. Doing so is inconvenient and eats up more of the customer's time than necessary, creating a negative experience that both parties would like to avoid. At the same time, airlines must pay humans to answer calls and field these relatively simple questions. Or, customers could navigate to the airline's website and spend a few moments searching around this unfamiliar space for the information they require.

Neither of these scenarios are optimal for the customer or the airline. They take up time, require undue effort, and increase costs. To resolve this, airlines such as Mexico's *Volaris* has created a Facebook Messenger chat bot capable of understanding and responding to hundreds of user questions. The chat bot answers questions at the rate of two humans, delivering answers more efficiently while also lowering customer service costs for Volaris. Since some questions regarding air travel are by nature complex, the chat bot is also capable of seamlessly transferring a user to a human agent to ensure they get the information they need promptly.

Booking and sales:

As messaging platforms such as Facebook Messenger become increasingly more supportive of chat bots, they are able to perform more complex tasks such as handling booking and payments. This gives users the opportunity to get the information they need without having to leave the Messenger platform.

An early mover in this sense is Icelandic airline *Icelandair*. Built in the Messenger platform, the Icelandair chat bot provides users the opportunity to search for and book a flight in a text-based conversational fashion. Rather than drop down menus, users enter the information themselves. This gives them more control over how the flight is booked. It also keeps the conversation in a thread so that they can later review their purchase and search information with ease.

Consolidating information:

Without chat bots, details critical to your flight end up spread across your digital ecosphere. Your ticket purchase information stays on the website where you bought the tickets. Your confirmation stays in your email inbox. Your boarding pass is stored in your phone's Passport or physically printed and carried. Your flight updates are sent via text message. Altogether, this makes keeping track of this info extremely difficult.

A full-stack chat bot such as that used by Dutch airline KLM allows you to store all critical flight information in one place: Facebook Messenger. Passengers can access their boarding pass, booking info, and flight details seamlessly. Plus, the chat bot is also capable of answering questions about your flight rapidly. It sports much of the functionality of an app, without the need for an actual download.

Additionally, the chat bot is able to actually make edits to your trip. If you are looking to change seats, a request

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

can easily be sent via the chat bot. Once it is confirmed, the updated ticket will be sent directly to you through Messenger. So a change that would previously have required the involvement of a customer service agent, as well as a website and email, has been swiftly streamlined into a single channel. Both the airline and the passenger save time and money, delivering the optimal flying experience.

There are two likely paths for airline chat bots moving forward. It's important to note that these are not mutually exclusive. These include:

More robust builds

Only a few airline chat bots fulfill a full load of operations. Some focus on booking, while others are more FAQ- focused. It's likely that once these companies start to see that their single function chat bot is working well, they will begin to build in other, more complex features. Air travel is an incredibly competitive field, with limited ability to differentiate on price. Therefore, airlines are constantly searching for opportunities to take the lead in creating the absolute best experience for their customers.

Proliferation to other platforms

At the moment, most airline chat bots exist in Facebook Messenger. There are a few reasons for this.

First, the Messenger platform is technologically advanced and offers a plethora of tools and capabilities to chat bot builders. It seeks to be as functional as anything you can access on a traditional website. Simply, it allows developers to offer a better experience to users.

Second, Messenger has an enormous amount of traction. At this moment, Messenger has 1.2 billion users spread across the world. The prominence of the platform makes it more likely that users will have access to that platform where their chat bot lives.

Once airlines have built more robust chat bots, it is likely that they will seek to expand to other platforms such as Twitter, Skype, or Kik among others. The messaging platform of choice is largely driven by what is popular in the user's specific location. Airlines can grow by adding their chat bot to platforms which are heavily used in areas they are looking to focus on. For example, WeChat is enormously popular in China.

How does my airline get on board?

Chat bots are significantly simpler and faster to build than applications. That means that a well-organized, nimble airline can have a chat bot up and running in no time. Chat bot building platforms such as [Snatch Bot](#) allow you to create a powerful chat bot that'll take your airline to a higher altitude of customer service.

The 6 steps to peak performance:

Happier customers:

Bots mitigate the frustration of customers with real-time responses to their queries and by reducing waiting time. They also enhance the self-service experience and prevent any possible miscommunications.

Information at one place:

Chat bots for airlines make it easy to track all critical flight details. Using them, the passengers can keep details such as boarding pass, flight updates ticket purchase information etc. at one place.

Conversational interface:

Equipped with natural language processing and machine learning capabilities, chat bots can understand contextual references and drive a conversation with the customers in a human-like interaction.

Process efficiency:

Chat bots can bring in more accuracy by automating numerous mundane tasks and cutting down the scope for human-errors and raising an alert when human support is needed.

Cost efficiency:

Intelligent workflows result in optimum utilization of resources and prevent the need for a high headcount. Thus chat bots help in bringing down the cost of operations phenomenally.

Competitive edge:

Exceptional customer service gives you a distinct reputation in the market, pulling more customers to you and improve customer loyalty. Therefore, you can get a competitive edge with bots.

Building an Airline Bot

The chat bot built is an Airline Bot. An Airline bot was built using few basic enquiries and details where a drop down was given listing few places where the passengers get to select or choose the place where they want to visit. As soon as the destination is chosen the bot displays the details regarding the place and will give the available flight timings in which the passengers got to choose the timings with which they are comfortable. The cost of the ticket is also shown on the screen. If they are ok with it then the bot will ask the passenger to mention their full name and their mobile number which is linked with their bank account. The required number will be debited from the person's account and the process of booking flight tickets is then successful. The bot will forward the required details later to the customer.

So, we can say that Airlines using chat bots ahead over their competitors by providing a better customer service. The main objective of building an Airline Bot is that it is easily accessible by all the online airline services and can make the work of the customers very easy. Usage of these chat bots will make us to cut short our time when we want to

book tickets. It is definitely a stress free activity for the people who are engaged in their daily work. We all can use Airline Bots and can also promote others to use it.

III. RESULT & DISCUSSION

When comparing common touristic web portals one can identify many similarities. The following attributes exist on nearly every website:

Earliest possible begin of journey & latest possible end of journey

Trip length

Age of children (implicitly number) & number of adults

Destination

Chat bots allow customers to get in contact with companies whenever they want so, without paying attention to time zones, opening times and waiting loops of call and service centers. The chat bot knows its users like a good friend and offers them appropriate offers, solutions and services at the right time.

Chat bots change the way of informing, communicating and transacting between the company and its customers or other external stakeholders.

The accuracy of the Question and Answer session is incredible.

Using chat bots, consumers and businesses can communicate *24 hours day, 7 days the week*, independent of working or opening hours.

Chat bots make our work easy and simple in knowing the exact details about our trip plan, the only thing we need to do is visit the correct site of the airline booking chat bots.

There are several types of assistants for each and every task involved in the process of booking tickets who respond properly to our questions.

The language which is used in the means of communication is natural and understandable.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

The internal chat bots help in upcoming organization, cooperation in the premises of the company.

Permanent events and touristic information about the cities and the areas are provided.

New opportunities for cross-selling activities in public

transport. Reservation management made easy.

Disadvantages of the bot:

Inability to Understand – Due to fixed programs, chat bots can be stuck if an unsaved query is presented in front of them. This can lead to customer dissatisfaction and result in loss. It is also the multiple messaging that can be taxing for users and deteriorate the overall experience on the website.

Increased Installation Cost – Chat bots are useful programs that help you save a lot of manpower by ensuring the all-time availability and serving to several clients at once. But unlike humans, every chat bot needs to be programmed differently for a new business which increases the initial installation cost. This also increases the time needed to prepare for the program and plan everything effectively. Considering the last-minute changes that can always happen, this is a risky investment as updating the program will invite added costs to it.

Another important topic for both providers and users is *data protection*. If companies offer a stand-alone chatbot app, they are responsible for protecting and handling customer data adequately. This can turn into a disadvantage if proper steps are not taken.

Poor Memory – Chat bots are not able to memorize the past conversation which forces the user to type the same thing again & again. This can be cumbersome for the customer and annoy them because of the effort required. Thus, it is important to be careful while designing chat bots and make sure that the program is able to comprehend user queries and respond accordingly.

Customers could become frustrated as many bots work for a limited data base, they can't improvise. In other words, if they get confused, the conversation could run in a circle. That can lead to customers who become frustrated.

Complex chat bots can cost more based on the purpose. Not all business can use chat bots, some businesses are far too complex for chat bots to be practical.

Applications :

Reservation / purchase of event tickets. Booking of hotel, trip and flight tickets. Customer service.

Ease to plan a trip according to the tickets available.

A popular use of chat bots is to deliver updates and offers based on consumer's preferences and history. Companionship is a remarkable application.

Content delivery of the chat bots is generally appreciated as they indulge completely in what are the users expecting from them and will try to satisfy them always.

IV. CONCLUSION

Airline bot will assist in friendly user experience by getting immediate support from the machine. It will

diminish the stress in customer service. Programmed alerts will help in dangerous cases. Reactive time for the question is negligible and precise. In near outlook, AI will present itself on a superior picture and will be included in our daily routine. There is a requirement to incessantly look for innovative thoughts for improvement and to develop in already set up research. The chatbot architecture amalgamates a language model and computational algorithm to follow information online contact connecting a human and a computer using common language. This computerized human to computer conversational plinths works optimistically to provide resourceful service in diverse fields to help humans.

By means of APIs like Government Services, Sports, Weather and News, the chatbot will be capable to respond to the queries exterior of its dataset and which are presently occurring in the real world. It is likely to assure information requirements on a large scale, tumbling users' time and endeavour and mounting competence and effectiveness of the process. In the upcoming years, we can picturise chatterbots as talking books for kids, chatterbots for foreign language instruction and training chatterbots overall.

Chatbots will alleviate the drawbacks people encounter in many aspects of work. The higher quality chatbots will use different features in different passes. As the chatbots has set the trend it has an effective role to play in business. The challenging task refers to complex chatbots and also depends on which they are designed but not the simple chatbot. The best quality of the chatbot is to be simple and should have the talent to understand easily.

Thus, chat bots are a thing of the future which is yet to uncover its potential but with its rising popularity and craze among companies, they are bound to stay here for long.

REFERENCES

1. Jizhou Huang, Ming Zhou, Dan Yang. (2007). *Extracting Chatbot Knowledge from Online Discussion Forums. International Joint Conference on Artificial Intelligence*, 423 -428.
2. AM Rahan, Abdullah Al Manum, Alma Islam. (2017, December). *Programming challenges of Chatbot: Current and Future Prospective. Humanitarian Technology Conference*, 75-78.
3. M.Dahiya. (2017, April). *A Tool of Conversation: Chatbot. International Journal of Computer Sciences and Engineering*, 5(5), 158-161.
4. Ayesha Shaikh, Geetanjali Phalke, Pranita Patil, Sangita Bhosale, Jyoti Raghawatan. (2016, November). *A Survey on Chatbot Conversational Systems. International Journal of Engineering and Computing*, 6(11), 3117-3119.
5. Ulrich Gnewuch, Stefan Morana, Marc Adam, Alexander Maedche. (2018, November 28). *Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human Chatbot Interaction. Association for Information Systems in Electronic Library*, 1-16.
6. Zhuling Zhong, Jing Luo, Mu Zhang. (2015). *Understanding Antecedents of Continuance Intention in Mobile Travel Booking Service. International Journal of Business and Management*, 10(9), 156-162.
7. Titta Jylkäs, Tytti Vuorikari, Mikko Äijälä, Vésaal Rajab. (2018, August). *AI Assistants as Non-human Actors in Service Design. Academic Design Management Conference*, 1436-1444.
8. Silvia Quarteroni, Suresh Manandhar. (2017). *A Chatbot-based Interactive Question Answering System*, 83-90.
9. Darius Zumstein, Sophie Hundertmark. (2018). *Chatbots – An Interactive Technology for Personalized Communication, Transactions and Services. IADIS International Journal on WWW/Internet*, 15(1), 96-109.
10. Papathanassis Alexis. (2017). *R-Tourism: Introducing the Potential Impact of Robotics and Service Automation in Tourism. "Ovidius" University Annals*, 17(1), 211-216.
11. Xiaojun Shen, Shervin Shirmohammadi, Chris Desmarais, Nicolas D. Georganas and Ian Kerr. (2006, February). *Enhancing e-Commerce with Intelligent Agents in Collaborative e-Communities. University of Ottawa, DOI: 10.1109/CEC-EEE.2006.43*.
12. Alexander Maedche, Stefan Morana, Silvia Schacht. (2016, August). *Advanced User Assistance Systems. Business and Information System Engineering*, 58(5), 367–370.

13. Anand Nayyar, Bandana Mahapatra, DacNhuong Le, G. Suseendran.(2018).Virtual Reality (VR) & Augmented Reality (AR) technologies for tourism and hospitality industry. *International Journal of engineering and technology*, 7(2.21), 156-160.
14. Mariusz Krzysztof Żytniewski.(2016 January). Integration of knowledge management systems and business processes usingmulti-agent systems. *International Journal of Computational Intelligence Studies*, 5(2), 180-196.
15. Arunas Miliauskas, Dal'e Dzemydien'e.(2018). An Approach to DesigningBelief-Desire-Intention Based Virtual Agents forTravel Assistance. *Institute of Data Science and Digital Technologies*, 94-103.
16. Markus Berg, Antje Düsterhöft.(2010). Website Interaction with Text-based Natural Language Dialog Systems. *University of Applied Sciences: Technology, Business and Design*,
17. Milan van Eeuwen. Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers. *University of Twente*, 1-15.
18. Anita Nathania P, Yulia, Vannesa Yuwono Putri.(2018 June). Android Based Chatbot and Mobile Application for Tour and Travel Company. *International Journal of Culture Technology*, 2(2), 21-29.
19. Eko Handoyo, M.Arfan, Yosua Alvin Adi Soetrisno, Maman Somantri, Aghus Sofwan, Enda Wista Sinuraya(2018).Ticketing Chatbot Service using Serverless NLP Technology. *International Conference on Information Technology, Computer and Electrical Engineering*, 325-330.
20. Bibek Behera. Chappie - A Semi-automatic Intelligent Chatbot, 1-5.

Digital Based Humanoid Robot(DBHR)

Ravi Vineet Sharma
rvineetsharma99@gmail.com

DBHR is the Robot which is used in typical whether conditions where Indian Army is facing Difficulties. DBHR is made of hard iron. DBHR is controlled by using Satellite signals and the signals will be given by the operator. DBHR is controlled by the Gamers in Laptop or PC who play games like Unknown Battle Grounds. Arduino micro processor and motor shield are used in DBHR. The complete code is written in C-language. It can rotate 360 degrees and can move the gun up and down.

OVERVIEW

DBHR is an initiative to overcome the problems faced by the Army.

Advantages:

- The main advantage of D.B.H.R is there will be no loss of soldiers life.
- It will be user friendly.
- It can kill the armed persons with in a short span of time .
- Since it is made of iron it's damage will be very less.

REALWORLD SCENARIO

The major problems faced by Indian army are:

- Indian army lives in the highest place of India(i.e. Siachen Glacier with temperature of -50°C).
- Indian army lives in a high temperature at the border of Indo-Pakistan at Rajasthan.
- They live at the places where terrorists attack frequently.

SOLUTION:

To avoid all the problems faced by Army we introduce the Robot named as DBHR.

The project aims at designing a Robot which is controlled through signals of satellite using satellite communication. The Robot can be rotated(360degrees) and the gun can be moved(upward,downward) through predefined keys assigned in the program. The proposed robot has an USB camera interfaced with ARM 11 processor for video surveillance.

The advent of new high-speed technology and the growing computer Capacity provided realistic opportunity for new robot controls and realization of new methods of control theory. This technical improvement together with the need for high performance robots created faster, more accurate and more intelligent robots using new robots control devices, new drivers and advanced control algorithms. This project describes a new economical solution of robot control systems. This project is mainly aiming to reduce the problems of a soldier at the typical weather conditions like Siachen Glacier. This Robot is controlled by a gamer in the operating room. Gamers who play battleground games can control this robot. They will get the real time experience of what they play.

CONCLUSION:

We conclude that DBHR is a robot which can reduce difficulties of the Indian Army.

The loss of life of soldiers will reduce if we use DBHR.

ANDROID BASED SIGN LANGUAGE TRANSLATOR

1 G. Pranay Goud 2 M. Ravi

¹ Student, Department of IT, J. B Institute Of Technology, Hyderabad, Telangana.

² Assistant Professor, Department of IT, J. B Institute Of Technology, Hyderabad, Telangana

ABSTRACT

In all around the world about 9.1 billion people are deaf and dumb. In their day to day life, they faced lot more problems on their communication. The deaf and mute people have difficulty in communicating with normal people who doesn't know sign language. So, to solve this problem, I introduced a sign language translator system for an android application. This system will allow user to upload images to server which will forward the images as input to the neural network in MATLAB. We have used supervised learning for training the neural network which is responsible for data classification. The response will be sent back to the server and server will sent back response to android device. This system will be implemented using client server architecture. This is an easy to use and inexpensive approach to recognise hand gestures accurately. This system will facilitate communication for millions of deaf and mute people and aid in communicating with people who don't understand sign language.

INTRODUCTION

Sign Language is the most natural and expressive way for the hearing impaired people. People, who are not deaf, never try to learn the sign language for interacting with the deaf people. This leads to isolation of the deaf people. But if the computer can be programmed in such a way that it can translate sign language to text format, the difference between the normal people and the deaf community can be minimized. The Sign Language development is different for each country or sub-continent. The following table presents the development of sign languages of influencing countries/subcontinent. The table below highlights the similarities and differences in their sign languages.

Table 1: Development of Sign Languages in different countries

S. No.	Country/Sub-Continent	Sign Language	Abbreviation
1	United Kingdom	British Sign Language	BSL
2	United States of America	American Sign Language	ASL
3	Commonwealth of Australia	Australian Sign Language	Auslan
4	Japan	Japanese Sign Language	JSL
5	People's Republic of China	Chinese Sign Language	CSL
6	Middle-East	Arabic Sign Language	ArSL
7	Republic of India	Indian Sign Language	ISL

Indian sign language (ISL) uses both single handed and double handed gestures to represent each alphabet and numbers. ISL alphabets are derived from British Sign Language (BSL) and French Sign Language (FSL). Most of the researchers in this area concentrate on the recognition of American Sign Language (ASL) since most of the signs in ASL are single handed and thus, complexity is less. Another attractive feature is that ASL already has a standard database that is available for use. A few research works carried out by the researchers in the recognition of ISL. Currently, more researchers have started doing research in ISL. Here this proposed system is able to recognize the various numbers of Indian Sign Language, this will reduce the noise and give accurate result. The important research problem in computer recognition is the sign language for enabling communication with hearing impaired people.

This system introduces efficient and fast techniques for identification of the hand gesture representing a number or alphabet of the Sign Language. Currently, more interest is created to do research in the field of sign language recognition system. Deaf and Dumb people rely on sign language interpreters for communications. A real time Sign Language Recognition system was designed and implemented to recognize gestures from the Indian Sign Language by hand gesture recognition system for text generation. The signs are captured by using camera in android device. This signs are processed and passed as input to the neural network for comparison. In order to perform the sign recognition, the features are compared with testing database. Finally, recognized gesture is converted into text.

SYSTEM ANALYSIS

EXISTING SYSTEM

The existing system have five accelerometers attached to the glove. For every gesture in sign language, there is a change in finger position. This change corresponds to change in coordinates of accelerometer. This data from the accelerometer is processed in Arduino Lily pad. The accelerometer used is Arduino compatible and it is connected to Lily pad through conductive threads. Arduino Lily pad is powered with power module which supplies 5v. This data is wirelessly transmitted and received by RF module. At the receiver side, Arduino UNO is used for processing the data received. For the corresponding change in coordinates, the code is written in Arduino IDE to display the character corresponding to sign made and also corresponding audio is also produced. TTS 256 (Text to speech) which takes in text and gives out corresponding audio. This is then given to speaker for

amplification of sound. The accelerometer is faster when compared to flex sensors because the change in coordinates can be easily recognised when compared to recognition of the change in voltage. As wireless approach is used, design is portable.

Drawbacks of Existing System

- We have to take the total equipment along with gloves whenever required.
- Sometimes hardware may not perform accurately.
- It is an expensive approach.

PROPOSED SYSTEM

The proposed system is based on client server architecture. Android App on the client side and MATLAB program on the server side. User captures the image containing hand gesture and uploads the image to the server by giving any name to the image.

For each and every upload from the android app, a row is created in the MySQL database. It is stored in the table images in projectmysql database. This row contains id number, upload URL and the name given for the image by the user.

The uploaded image containing hand gesture from the client side android app by the user is stored in the uploads folder on the server with the name as in the database id number.

The server program takes the image as input from uploads folder and resize the image using `imresize()` function in MATLAB[6] and then pass the image to the Neural Network in MATLAB. We used Alexnet, pretrained neural network for hand gesture recognition[4]. Neural Network is trained with the dataset containing thousands of images so that it can predict the hand gesture accurately. Processing is done and the corresponding converted text to the hand gesture is sent back to the server and from the server to the Android App in JSON format.

Advantages of Proposed System

- As it is software based, just an Android device with an app is necessary to recognise the hand gestures.
- It is easy to use as everyone uses a smart phone now-a-days.
- It is an inexpensive approach.

LITERATURE SURVEY

In olden days, in order to communicate with deaf people, the normal people used to write it on a paper and show to them. Similarly in case of mute people, in order to communicate mute people are going to write it on a paper and show to the normal people.

Later on sign language was came into the picture. A sign language is a language which chiefly uses hand movements and body language to communicate meaning and idea.

It was in 1970's the first glove based systems were introduced. The first prototype was brought in which each finger of the glove has flexible tubes having a photocell and light source. A bent in the tube caused reduction in the light amount that passed between the photocell and the source which was related to voltage of photocell.

In 1980's the MIT media laboratory came up with the MIT LED which had a LED system based on camera for tracking the body and limb movements.

Then in 1983, Gary Grimsby designed a digital entry data glove envisioned for generating alpha numeric characters from different hand posters identified by hard wired circuit board.

Later on, the data glove with sensors was introduced. This system has five flex sensors, an Arduino Lily pad, Bluetooth hc-05 and a battery.

Data glove is an electronic device that senses the movements of the hand and, also, the fingers individually and sends these movements to the computer in the form of analog or digital signals. These digital signals are then mapped to the task to be performed in the virtual environment. On this glove, various sensors are placed to detect the global position and relative configurations of the hand.

On this glove, flex sensors are used for each finger and they are connected to arduino. Based on movement of the flex sensors, the corresponding data is send to the arduino. Arduino is coded such that based on the data from flex sensors, it can show the corresponding meaning of the data that is gesture movement.

SYSTEM DESIGN

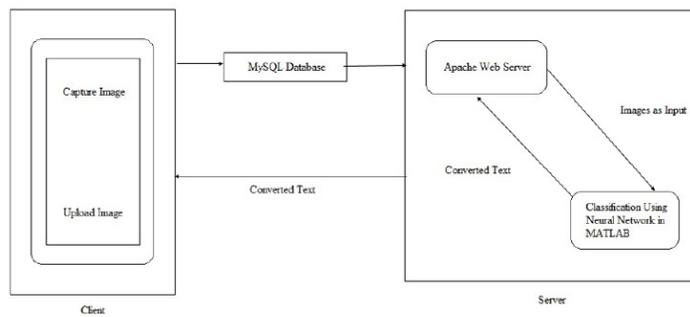


Fig : System Architecture

Client:

On the client machine there will be an Android application which will be used to capture images of the gesture. Once the image is captured, it will be uploaded on the server using the android application. Each and every upload data is stored in MySQL database. The client machine is also used to receive the converted text from the server after all processing is done on the server.

Server:

The uploaded image is stored in the uploads folder on the server. For this application I am using Apache Web Server. The image on the server is given as an input to the Neural Network in MATLAB. It classifies the image and the converted text is sent back to the client in JSON format.

IMPLEMENTATION

MODULES

“Android based Sign Language Recognition” mainly consists of four modules:

- ⊙ Android Application
- ⊙ Importing Dataset
- ⊙ Splitting the data
- ⊙ Building neural networks

© Prediction

© Sending converted text to Android app

MODULE DESCRIPTION

Android Application:

Initial step in our project is to develop an android application using Android Studio. For accessing camera in android device, camera2 class provided by Android Studio was used[1]. The android app will invoke the camera in-built in the android device. The android app is used to capture the images containing hand gestures. The captured images were then sent to the server. These images were then provided as input to MATLAB for pattern recognition and classification[3].

Importing Dataset:

Import image dataset which are having hand gesture contained images in different light conditions with labels.

Splitting the data:

After importing the dataset, we need to split the data into two parts like training and testing datasets. Based on split size, need to split the data.

Building neural networks:

We need to feed the convolution neural networks with trained data with different layers, after that giving the testing the data for analyzing the result. We are using Alexnet, pretrained convolutional neural network by changing the fully connected layer of alexnet using transfer learning and performed training and testing on this neural network[5].

Prediction:

After performing training and testing, give the hand gesture image to neural networks it will check the label for that image and then gives the result i.e., the converted text of the hand gesture.

Sending converted text to android app:

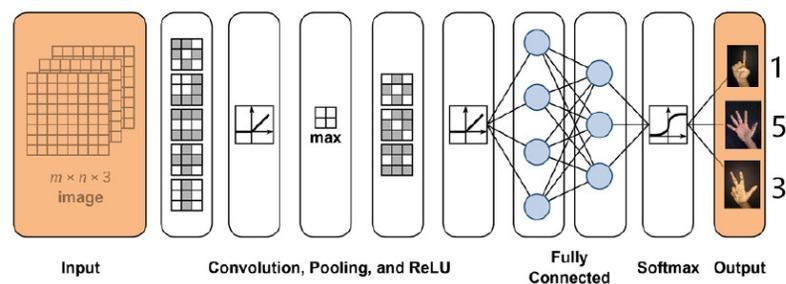
Finally, the converted text is sent back to the android app using JSON object. JSON object is used to send text from server to the android device. JSON stands for JavaScript Object Notation and is used to exchange data to/from a web server. There are four different classes provided by Android to manipulate JSON data. We have used JSONObject class to receive text from the server[2].

Algorithms:

Convolutional Neural Networks: Currently, CNNs are the most researched machine learning algorithms. It is one of the most popular algorithms for deep learning, a type of machine learning in which a model learns to perform classification tasks directly from images, video, text, or sound. CNNs are particularly useful for finding patterns in images to recognize objects, faces, and scenes. They learn directly from image data, using patterns to classify images and eliminating the need for manual feature extraction. As shown in, a CNN takes an input image of raw pixels, and transforms it via Convolutional Layers, Rectified Linear Unit (RELU) Layers and Pooling Layers. This feeds into a final Fully Connected Layer which assigns class scores or probabilities, thus classifying the input into the class with the highest probability.

CNN(Convolutional Neural Network) Algorithm:

Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernels), Pooling, fully connected layers (FC) and apply Soft max function to classify an object with probabilistic values between 0 and 1. We used alexnet, pretrained CNN for this project. The below figure is a complete flow of CNN to process an input hand gesture image and classifies the hand gestures based on values.



CONCLUSION AND FUTURE ENHANCEMENT

Sign Language Translator system was implemented. It helps the deaf people to communicate with the normal people. It focuses mainly on the recognition of Indian Sign Language from images that have been taken under controlled conditions. The design is easy to use and inexpensive when compared to existing design using Arduino Lily pad and accelerometers. The proposed sign language recognition system recognises the gestures in constrained environment like dark background. It solves the problem of deaf and dumb for communicating with normal people.

Further work should focus on hand segmentation method. This system can be further developed to recognise gestures in real time and in video format. Many other gestures of the sign language can be made part of the database.

BIBLIOGRAPHY

- [1] Android - Taking Photos [Online]. Available: <https://developer.android.com/training/camera/photobasics.html>
- [2] Android - JSON Parser [Online]. Available: https://www.tutorialspoint.com/android/android_json_parser.htm
- [3] MATLAB - Deep Learning [Online]. Available: <https://in.mathworks.com/videos/series/deep-learning-with-MATLAB.html>
- [4] MATLAB - Alexnet [Online]. Available: <https://in.mathworks.com/help/deeplearning/ref/alexnet.html>
- [5] MATLAB - Transfer Learning [Online]. Available: <https://in.mathworks.com/help/deeplearning/examples/transfer-learning-using-alexnet.html>
- [6] MATLAB - Imresize [Online]. Available: <https://in.mathworks.com/help/matlab/ref/imresize.html>

Wireless Sensor Networks: Applications and challenges

Dr.G.Rajitha Devi., Asst.prof in Computer Science
Government Degree College HayathNagar, Hyderabad.

Abstract

This paper describes the concept of sensor networks which has been made viable by the convergence of micro-electro-mechanical systems technology, wireless communications and digital electronics. First, the sensing tasks and the potential sensor networks applications are explored, and a review of factors influencing the design of sensor networks is provided. Then, the communication architecture for sensor networks is outlined, and the algorithms and protocols developed for each layer in the literature are explored. Open research issues for the realization of sensor networks are also discussed.

Keywords

Wireless sensor networks, Ad hoc networks, Application layer, Transport layer, Networking layer, Routing Data link layer, Medium access control, Error control, Physical layer, Power aware protocols

Introduction to Wireless Networks

A Wireless sensor network can be defined as a network of devices that can communicate the information gathered from a monitored field through wireless links. The data is forwarded through multiple nodes, and with a gateway, the data is connected to other networks like [wireless Ethernet](#). WSN is a wireless network that consists of base stations and numbers of nodes (wireless sensors). These networks are used to monitor physical or environmental conditions like sound, pressure, temperature and co-operatively pass data through the network to a main location .

A wireless network is any sort of computer network that uses wireless data connections to plug network nodes. Wireless networks are computer networks who are not connected by cables regardless of the sort. The use of a wireless network enables enterprises to prevent the costly means of introducing cables into buildings or as a connection between different equipment locations. The cornerstone of wireless systems is radio waves, an implementation that occurs at the physical higher level of network structure. Wireless technologies differ in a number of dimensions, most notably in just how much bandwidth they provide and how far apart communicating nodes can be. Other important differences include which perhaps the electromagnetic spectrums they choose and exactly how much power they consume

A Wireless Sensor Network (WSN) is by hundreds of small, low-cost nodes that are fitted with limitations in memory, energy, and processing capacity. In this particular form of networks, several problems is to learn each node. Recent advances in wireless communications and electronics have enabled the roll-out of low-cost, low-power and multi-functional sensors that are small in dimensions and communicate in a nutshell distances. Cheap, smart sensors, networked through wireless links

and deployed in vast quantities, provide unprecedented opportunities for monitoring and controlling homes, cities, along with the environment. Furthermore, networked sensors use a broad spectrum of applications within the defense area, generating new capabilities for reconnaissance and surveillance and various Tactical applications. Self-localization capability can be a highly desirable sign of wireless sensor networks. In environmental monitoring applications for example bush fire surveillance, water quality monitoring and precision agriculture, the measurement data are meaningless lacking the knowledge of the placement from the location where the data are obtained. Moreover, location estimation may enable many applications for example inventory management, transport, intrusion detection, road traffic monitoring, health monitoring, reconnaissance and surveillance. With all the advances inside the miniaturization and integration of sensing and communication Technologies, large-scale wireless sensor networks using a large number of low-cost and low-power sensors are already developed. Within a wireless sensor network, lots of money of tiny, battery-powered sensor nodes are scattered throughout a physical area. Each sensor in the sensor network collects data, as an example, sensing vibration, temperature, radiation along with other environmental factors .A wireless sensor network (WSN) includes hundreds to a large number of low-power multi-functional sensor nodes, operating within the unattended environment, and having sensing, computation and communication capabilities. The essential the different parts of a node undoubtedly are a sensor unit, an ADC (Analog to Digital Converter), a CPU (C.P.U.), an electrical unit as well as a communication unit. Sensor nodes are microelectro-mechanical systems (MEMS) that develop a measurable a reaction to a general change in some fitness like temperature and pressure. Sensor nodes sense or measure physical data in the area being monitored. The continual analog signal sensed through the sensors is digitized by an analog-to-digital converter and sent to controllers for more processing. Sensor nodes are of small size, consume extremely low energy, are operated in high volumetric densities, and will be autonomous and adaptive towards the environment.

Types of Wireless Networks

Basically, there are five types of wireless networks:

1. Wireless PAN
2. Wireless LAN
3. Wireless MAN
4. Wireless WAN
5. Global Area Network

Ad-hoc Networks

Ad-hoc networks are multi-hop wireless networks that can operate minus the services of the established backbone infrastructure. While such networks have obvious applications from the military and disaster relief environments, more modern works that contain motivated their use even in regular wireless packet data networks have raised their significance. The main objective on this paper should be to study the performance with the TCP transport layer protocol over ad-hoc networks thinking about an ad hoc network is normally unfamiliar to finish users with only seen small residential or business networks that use a standard router to send wireless signals to individual computers. However, the ad hoc network will be used a great deal in new sorts of wireless engineering, although until recently it turned out a rather esoteric idea. One example is a mobile

random network involves mobile devices communicating directly with each other. A different type of random network, the vehicular random network, involves placing communication devices in cars. Both these are examples of ad hoc networks designed to use a large variety of individual devices to freely communicate with no sort of top-down or hierarchical communication structure.

Ad-hoc Networks Characteristics

- 1. Multihopping:** A multihop network is a network the spot that the path from source to destination traverses other nodes. Random nets often exhibit multiple hops for obstacle negotiation, spectrum reuse, and conservation. Battlefield covert operations also favour a sequence of short hops to scale back detection by the enemy.
- 2. Self-organization:** The ad hoc network must autonomously determine its very own configuration parameters including: addressing, routing, clustering, position identification, power control, etc. Sometimes, special nodes can coordinate their motion and dynamically distribute from the geographic area to supply coverage of disconnected islands.
- 3. Energy conservation:** Most ad hoc nodes (e.g., laptops, PDAs, sensors, etc.) have limited power supply no power to generate their particular power (e.g., solar power systems). High efficiency protocol design (e.g., MAC, routing, resource discovery, etc) is important for longevity with the mission.
- 4. Scalability:** In certain applications (e.g., large environmental sensor fabrics, battlefield deployments, urban vehicle grids, etc) the random network can grow to thousand nodes. For wireless “infrastructure” networks scalability is actually handled by a hierarchical construction.

Applications of Wireless Sensor Networks

The applications for WSNs involve tracking, monitoring and controlling. WSNs are mainly utilized for habitat monitoring, object tracking, nuclear reactor control, fire detection, and traffic monitoring. Area monitoring is a very common application of WSNs, in which the WSN is deployed over a region where some incident might be monitored. E.g., a substantial variety of sensor nodes may very well be deployed over the battlefield to detect enemy intrusions rather than using landmines. When the sensors detect case being monitored (heat, pressure, sound, light, electro-magnetic flux, vibration, etc.), the big event needs to be reported to at least one in the base stations, which often can than take some appropriate action (e.g., send some text online or even a satellite). Wireless sensor networks are utilized extensively within the water/wastewater industries. Facilities not wired for power or data transmission can be monitored using industrial wireless I/O devices and sensor nodes powered by solar panels or battery packs. Wireless sensor networks are able to use numerous sensors to detect the existence of vehicles for vehicles detection. Wireless sensor networks may also be employed to control the temperature and humidity levels inside commercial greenhouses. If the temperature and humidity drops below specific levels, the greenhouse manager might be notified via e-mail or a cellular telephone text, or host systems can trigger misting systems, open vents, first turn on fans, or control a multitude of system responses. Because some wireless sensor networks are super easy to install, they've also been simple move if the needs with the application change.

There are lots of applications of WSN:

1. Process Management: Area monitoring is a very common using WSNs. In area monitoring, the WSN is deployed spanning a region where some phenomenon is usually to be monitored. A military example may be the use of sensors detect enemy intrusion; a civilian example would be the geofencing of gas or oil pipelines. Area monitoring is most important part.

2. Healthcare monitoring: The medical applications might be of two sorts: wearable and implanted. Wearable devices are applied to the body surface of the human or maybe at close proximity from the user. The implantable medical devices are the ones that are inserted inside your body. There are numerous other applications too e.g. body position measurement and of the person, overall monitoring of ill patients in hospitals and also at homes.

3. Environmental/Earth sensing: There are numerous applications in monitoring environmental parameters samples of which are given below. They share any additional challenges of harsh environments and reduced power supply.

4. Polluting of the environment monitoring: Wireless sensor networks have been deployed in lots of cities (Stockholm, London and Brisbane) to monitor the power of dangerous gases for citizens. These can leverage the random wireless links instead of wired installations that also make them more mobile for testing readings in several areas.

5. Forest fire detection: A network of Sensor Nodes is usually positioned in a forest to detect every time a fire has begun. The nodes is usually with sensors to measure temperature, humidity and gases which are produced by fire within the trees or vegetation. The first detection is necessary to get a successful action of the fire fighters; As a result of Wireless as Sensor Networks, the fire brigade are able to know when a fire begins you bet it can be spreading.

6. Landslide detection: A landslide detection system uses a wireless sensor network to detect the slight movements of soil and modifications to various parameters that will occur before or throughout a landslide. With the data gathered it may be possible to know the appearance of landslides before it genuinely happens.

7. Water quality monitoring: Water quality monitoring involves analyzing water properties in dams, rivers, lakes & oceans, and also underground water reserves. The application of many wireless distributed sensors enables the creation of a accurate map on the water status, and allows the permanent deployment of monitoring stations in locations of difficult access, while not manual data retrieval.

8. Natural disaster prevention: Wireless sensor networks can effectively act to avoid the results of disasters, like floods .Wireless nodes have successfully been deployed in rivers where changes in the water levels have to be monitored in realtime.

9. Industrial monitoring:

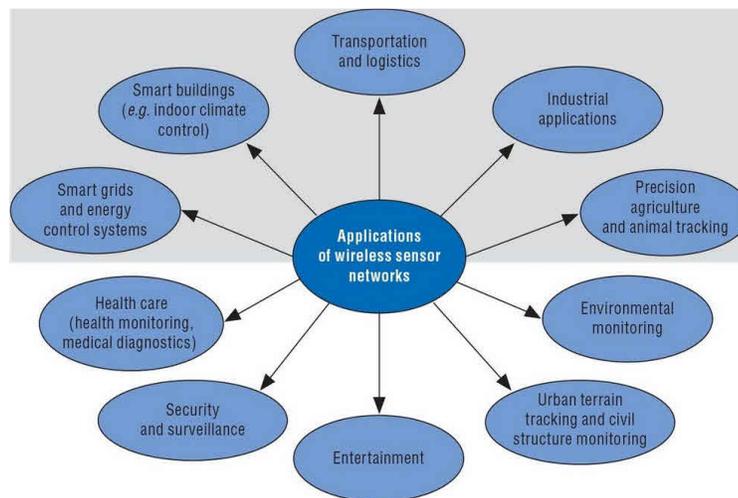
a. Machine health monitoring: Wireless sensor networks happen to be developed for machinery condition based maintenance (CBM) as they offer significant personal savings and enable new functionality .In wired systems, installing enough sensors can often be tied to the price of wiring. Previously inaccessible locations, rotating machinery, hazardous or restricted areas, and mobile assets can now be reached with wireless sensors.

b. Data logging: Wireless sensor networks are also employed for the gathering of web data for monitoring of environmental information; this is often as easy as the monitoring from the temperature in a very fridge to

the level of water in overflow tanks in nuclear power plants. The statistical information will then be employed to show how systems have been working. The main benefit of WSNs over conventional loggers is the "live" data feed which is possible.

c. Water/Waste water monitoring: Monitoring the high quality and level of water includes many activities including checking the quality of underground or surface water and ensuring a country's water infrastructure for your benefit of both human and animal .It may be helpful to protect the wastage of water.

d. Structural Health Monitoring: Wireless sensor networks enables you to monitor the fitness of civil infrastructure and related geophysical processes all around real time, and more than very long stretches through data logging, using appropriately interfaced sensors



Wireless Sensor Networks Applications

- These networks are used in environmental tracking, such as forest detection, animal tracking, flood detection, forecasting and weather prediction, and also in commercial applications like seismic activities prediction and monitoring.
- applications, such as tracking and environment monitoring surveillance applications use these networks. The sensor nodes from sensor networks are dropped to the field of interest and are remotely controlled by a user. Enemy tracking, security detections are also performed by using these networks.
- Health applications, such as Tracking and monitoring of patients and doctors use these networks.

- The most frequently used wireless sensor networks applications in the field of Transport systems such as monitoring of traffic, dynamic routing management and monitoring of parking lots, etc., use these networks.
- Rapid emergency response, industrial process monitoring, automated building climate control, ecosystem and habitat monitoring, civil structural health monitoring, etc., use these networks.

This is all about the wireless sensors networks and their applications. We believe that the information about all the different types of networks will help you to know them better for your practical requirements. Apart from this, for additional information about wireless SCADA, queries, and doubts regarding this topic or electrical and electronic projects, and for any suggestions, please comment or write to us in the comment section below.

Research Challenges in Wireless Sensor Networks

A brief history on the research in SN, but more interesting may be the overview within the technical challenges and issues is presented, from where we could cite several relevant items: WSN working in a harsh environment; the ability with the network (leastways the neighbors); the network control and routing; querying and tasking (should be as simple and intuitive as it can be); plus security issues (low latency, survivable, low probability of detecting communications, high reliability)..

1. Security: Security is often a broadly used term encompassing the characteristics of authentication, integrity, privacy, non repudiation, and anti-playback. The greater the dependency on the info supplied by the networks may be increased, the more potential risk of secure transmission of information in the networks has increased. To the secure transmission of numerous kinds of information over that happen to be renowned. In this section, we discuss the network security fundamentals you bet the techniques are meant for wireless sensor networks .

2. Cryptography: The encryption-decryption techniques devised for your traditional wired networks usually are not feasible to be employed directly for the wireless networks in particular for wireless sensor networks. WSNs include things like tiny sensors which really suffer from the possible lack of processing, memory and battery Applying the security mechanisms for instance encryption could also increase delay, jitter and packet loss in wireless sensor networks when applying encryption schemes to WSNs like, what sort of keys are generated or disseminated. How a keys are managed, revoked, assigned to your new sensor put into the network or renewed for ensuring robust to protect the network Adoption of pre-loaded keys or embedded keys could hardly be an efficient solution.

3. Steganography: While cryptography aims at hiding necessary of a message, steganography aims at hiding a good the message. Steganography is the art of covert communication by embedding a note in to the multimedia data (image, sound, video, etc.). The leading objective of steganography is to modify the carrier in a fashion that is just not perceptible and hence, it looks the same as ordinary.

4. Physical Layer Secure Access: Physical layer secure access in wireless sensor networks may very well be offered by using frequency hopping. A dynamic mixture of the parameters like hopping set (available frequencies for hopping), well time (interval per hop) and hopping pattern (the sequence in which the frequencies in the available hopping set is used) could be combined with a little expense of memory, processing and resources. Important points in physical layer secure access

will be the efficient design in order that the hopping sequence is modified in less time than is required to discover it and for employing this both sender and receiver should maintain a synchronized clock. A scheme as proposed in may be utilized which introduces secure physical layer access employing the singular vectors while using channel synthesized modulation. Attacks against wireless sensor networks may very well be broadly considered from two different levels of views. One is the attack from the security mechanisms and this band are brilliant from the basic mechanisms (like routing mechanisms). Ideas signalize the most important attacks in wireless sensor networks.

5. Localization: It is amongst the key techniques in wireless sensor network. The place estimation method is usually classified into Target / source localization and node self-localization. In target localization, we mainly introduce the energy-based method. Then we investigate the node self-localization methods. Considering that the widespread adoption on the wireless sensor network, the localization methods are wide and varied in several applications. There are some challenges using some special scenarios. With this paper, we present a wide survey these challenges: localization in non-line-of-sight, node selection criteria for localization in energy-constrained network, scheduling the sensor node to optimize the tradeoff between localization performance and energy consumption, cooperative node localization, and localization algorithm in heterogeneous network. Finally, we introduce the evaluation criteria for localization in wireless sensor network. The entire process of estimating the unknown node position inside the network is known as node self-localization. And WSN comprises a large number of inexpensive nodes which are densely deployed in a very region of interests to measure certain phenomenon. The leading objective would be to determine the location of the target. Localization is significant travelers have an uncertainty with the exact location of some fixed or mobile devices. One example has been in the supervision of humidity and temperature in forests and/or fields, where thousands of sensors are deployed by way of plane, giving the operator minimal possible ways to influence may location of node. An efficient localization algorithm might utilize all the free information from the wireless sensor nodes to infer the positioning of the individual devices. Another application will be the positioning of an mobile robot determined by received signal strength from your number of radio beacons placed at known locations around the factory floor. The primary function of an location estimation method to calculate the geographic coordinates of network nodes with unknown position in the deployment area. Localization in wireless sensor networks is the process of determining the geographical positions

of sensors. Only a number of the sensors (anchors) inside the networks have prior knowledge about their geographical positions. Localization algorithms utilize location information of anchors and estimates of distances between neighboring nodes to discover the positions in the rest of the sensors.

6. Power-Consumption: A wireless sensor node can be a popular solution when it is difficult or impossible to perform a mains supply towards sensor node. However, because the wireless sensor node is normally positioned in a hard to reach location, changing the battery regularly will not be free and inconvenient. An essential take into account the introduction of a wireless sensor node is making sure that there's always adequate energy accessible to power the system. The facility consumption rate for sensors in the wireless sensor network varies greatly good protocols the sensors use for communications. The Gossip-Based Sleep Protocol (GSP) implements routing and many MAC functions in a energy conserving manner. The effectiveness of GSP has already been

demonstrated via simulation. However, no prototype system has become previously developed. GSP was implemented for the Mica2 platform and measurements were conducted to discover the improvement in network lifetime. Results for energy consumption, transmitted and received power, minimum voltage supply necessary for operation, effect of transmission power on energy consumption, and different methods for measuring time of a sensor node are presented. The behavior of sensor nodes when they're all around their end of lifetime is described and analyzed.

7. Deployment: Sensor networks provide capability to monitor real-world phenomena in more detail and also at large scale by embedding wireless network of sensor nodes in the environment. Here, deployment is anxious with establishing an operational sensor network inside a real-world environment. On many occasions, deployment is often a labor-intensive and cumbersome task as environmental influences trigger bugs or degrade performance in a way that is not observed during pre-deployment testing within a lab. The real reason for this really is that the real life features a strong influence for the function of your sensor network by governing the output of sensors, by influencing the existence and excellence of wireless communication links, and also by putting physical strain on sensor nodes. These influences is only able to be modeled to your very restricted extent in simulators and lab testbeds. Home the typical problems encountered during deployment is rare. You can only speculate for the grounds for this. On one side, a paper which only describes what actually transpired during a deployment seldom constitutes novel research and could possibly be hard to get published. However, people might often hide or ignore problems that are not directly related to their field of research. It is additionally often tough to discriminate desired and non desired functional effects for the different layers or levels of detail.

References:

[1] **Localization in Wireless Sensor Networks**, King-Yip Cheng ,The University of Hong Kong
December 2006.

[2] **Wireless Networking Complete The Morgan Kaufmann Series in Networking Series Editor** ,
David Clark, M.I.T., Pei Zheng, Feng Zhao, David Tipper, Jinmei Tatuya, Keiichi Shima, Yi
Qian, Larry Peterson, Lionel Ni, D. Manjunath, Qing Li, Joy Kuri, Anurag Kumar, Prashant
Krishnamurthy.

[3] **AD HOC NETWORKS Technologies and Protocols, Edited by PRASANT MOHAPATRA**
University of California, Davis, SRIKANTH V. KRISHNAMURTHY University of California,
Riverside ©2005 Springer Science + Business Media, Inc.

[4] **A Microscopic Analysis of TCP Performance over Wireless Ad-hoc Networks**,
Vaidyanathan Anantharaman, Raghupathy Sivakumar

[5] **Vehicular ad hoc networks (VANETS): status, results, and challenges, Sherali Zeadally** ·
Ray Hunt · Yuh-Shyan Chen ,Angela Irwin · Aamir Hassan, © Springer Science+Business
Media, LLC 2010.

[6] http://en.wikipedia.org/wiki/Fixed_wireless.

[7] **Wireless Sensor Networks, The Morgan Kaufmann Series in Networking Series Editor**,
David Clark, M.I.T., Feng Zhao, Leonidas J. Guibas Morgan Kaufmann Publishers is an

imprint of Elsevier

[8] A **Comparative Study of Wireless Sensor Networks and Their Routing Protocols**, Debnath Bhattacharyya , Tai-hoon Kim , and Subhajit Pal, Sensors 2010, 10, 10506-10523; doi:10.3390/s101210506 www.mdpi.com/journal/sensors.

[9] **Wireless Sensor Networks**¹, F. L. LEWIS Associate Director for Research Head, **Advanced Controls, Sensors, and MEMS Group Automation and Robotics Research Institute** The University of Texas at Arlington 7300 Jack Newell Blvd. Sft. Worth, Texas 76118-7115

[10] **Challenges in Wireless Sensor Networks**, Er. Barjinder Singh Kaler, Er. Manpreet Kaur Kaler, RIMT-MAEC, Mandigobindgarh.

Social Applications in the Home Network

¹ Boga Jayaram,, ²Mysa Kalyana chakravarthy

¹ Assistant Professor, Department of Computer Science & Engineering, Balaji Institute of Technology & science,
Telangana,India,

² Assistant Professor, Department of Computer Science & Engineering , St.Marry's Engineering College,
Telangana, India

Email: ¹ jayaramboga@gmail.com , ² mkalyan8@gmail.com

Abstract — Home devices like set-top boxes and media gateways are a rich source of users' information. They can play a role to facilitate interactions with friends and relatives in social networks. However these interactions have to follow policies to safeguard users' privacy while maintaining the complexity low for the average user. In this article we prototype an architecture based for home devices to participate in social networks. Besides, we explore the privacy issues and how the privacy offered in social networks can be improved.¹.

Index Terms — Privacy in social networks, home networks, credential management, security policy.

I. INTRODUCTION

In the actual web landscape, there are two emerging forces: social networks and multimedia content delivery. Social networks are growing in importance as a convenient tool to keep our friends and relatives aware of our daily lives. We use them to share comments, photos, videos, and even to blog details of our routine especially important for those closer to us.

On the other hand Internet-based multimedia content delivery is starting to displace TV as the main source of entertainment. Both forces are complementary since users get more attracted by the Internet to find out about their social contacts, and the general audience looks for contents in the Internet.

This can be observed through the movement of commercial campaigns towards social networks, even house made campaigns. A recent example is the success of a Facebook campaign run by Jon and Tracy Mortner this Christmas in the UK. The campaign addressed the music market and got more than 800,000 fans and sold more than 500,000 copies of a CD (see [1]).

Nodes of a social network relate to each other by interests shared among them such as friendship, values, ideas, etc. Though social networks may appear as random, they are closer to the scale-free organization model [2]. Social networks are characterized by large hubs which make it

possible the “six-degree separation” [3] between any given individual. Users can thus easily target a group of people with common interests to share contents.

Nodes can be grouped into small sub-networks as for instance, family, friends, or fans of Star Wars, giving the user's view of the social network. Thus, a sub-networks, may be unaware of the existence of other nodes, and users can manage components and permissions of sub-networks to control how their contents may be (un)reachable by other nodes and sub-networks.

Industry of consumer electronics has been required to put an extra effort in converging devices with Web 2.0 services, to increase revenues. This enables the integration of devices with social networks, an emerging market already under exploitation by Telecommunication Companies and cell phone manufacturers.

Home devices, such as set-top boxes and media gateways are rich in user information such as preferences, habits or routines. They have knowledge to the movies and shows the user likes to watch, music he likes to listen to, and even the intended recording and playing schedule for them. That privileged access to user information turns them into appealing candidates platforms to deploy social network applications. However, they are also guardians of our private life and secrets. Hence different side effects might appear when integrating social networks with home devices, and our privacy may be compromised if users' information is not handle with care. This is specially worrying if children

and teenagers are involved, since they do not usually make a mature deliberation when deciding to publish personal information or media [4].

To enable social network interaction in home environments and to overcome privacy problems we present an architecture based on two major building blocks, which we prototype as two home gateway applications: the Social Enabler and its counterpart the Social Watchdog. Both applications mediate among home devices and social networks. The aim of the Social Enabler is to provide content adaptation and content sharing. The Social Watchdog provides security and privacy services governing the disclosure of personal information and tracking what others publish about home users.

The structure of the article is as follows. Section II gives an overview of the state of the art of privacy in social networks. Section III is devoted to describe how a social network application is deployed in our proposed architecture. Section IV deeply explains the functionality of the Social Enabler and the Social Watchdog respectively. In section V we review related works. Finally, section VI presents our conclusions and future work.

I. PRIVACY IN SOCIAL NETWORKS

Sub-networks of a social network are communities of interest where nodes represent users sharing contents and media. Social networks allow nodes to control how their contents are accessed and they differ in the support for the formation of such sub-networks with group definition and management policies and they may even facilitate fine grained access controls to enforce strict policies on the sharing of contents among the sub-networks.

However, as stated in [4], often users are not conscious about their privacy, or it is the case that the sub-networks has an open nature. There have been several efforts, coming from the databases and computer theory fields, to apply anonymity techniques to difficult an attacker to identify nodes.

In spite of that, [5] describe different active and passive attacks to successfully identify the nodes in the supposedly anonymous graph. They work under the hypothesis that the full structure of the anonymized graph can be obtained by the attacker. In the active attacks, some fake nodes are added to the network, and links among them are established to ensure that after the anonymity process, the fakes nodes can be identified in the anonymized graph. Once the fake nodes are identified the nodes connected to them can be identified as well. The passive attacks only require some new links between legal existent nodes. Once nodes are identified, this information can be used by purposes hidden to the user. Or worse, worms and bots can automatically collect information about the exposed users to build up profiles of the users for future illegal attacks, where the most common is the impersonation of the user in other contexts. Examples of such user's information are: maiden name, parents name and date of birth, pet's name, and the kind of information often used as security questions to retrieve our credentials or password information from other external sites.

Krishnamurthy et al. [6] characterize the potential privacy leakage in social networks. They identify what bits of information are currently being shared, how widely, and what users and social networks can do to prevent such sharing. Limiting access to just friends or those in a network is not fine-grained enough. They claim that it should be also possible to both deny and enable access to private information at the same level of granularity. They establish the necessity of a mechanism to identify the metrics bare minimum (minimum information needed by a third party application) and supremum (for applications exhibiting multiple behaviors).

A pragmatism proposal we find very appealing comes from Felt et al. [7], addressing the privacy risks associated with third party applications using social networking APIs. They design a privacy-by-proxy API that preserves privacy from external sites to the social network. They analyze 150 Facebook applications to determine the applicability of their API for limiting their access to real data. The API instead, provides third party applications with access to an anonymized social graph and placeholders for user data.

In this approach the application developers are enforced to respect the privacy of the users by using this reduced API, which uses encrypted user Ids, restricts reachability to friends, and unconnected nodes may be solved only by unidirectional relationships called contact lists.

In the context of this work, we deal with the privacy issues arising from publishing the wrong content, in the wrong network, made by mistake or due to immature publication decision. With respect to protecting privacy from third party applications we support the work in [7], and against malicious nodes we trust in the privacy controls of the social network. In our work, we consider social applications located in the home network sharing information within home user's sub-networks. Our aim is to preserve the privacy of this information as much as we can. We face different scenarios: i) closed sub-networks, ii) open sub-networks, and iii) mixed sub-networks, where membership and

information openness may differ. It is quite challenging to build a single working strategy for the three cases, since the third case requires the information to be publicly shared, whereas this is out of question in closed sub-networks. Besides, membership control is strictly enforced in closed networks, while loose and even non-existent in open sub-networks. We require specific policies to address all of them.

I. PROPOSED ARCHITECTURE

In this section we propose an architecture for integrating applications in a social network. The application may be deployed as part of the software located in a home device, in the home gateway itself, or even out of the home network, at the premises of a third party application provider.

Social networks often provide two interfaces: a web frontend for human users and a set of Representational State Transfer (REST) or Web Services for third party applications. Our proposed architecture includes the elements depicted in Figure 1.

The architecture provides home applications with two components which offer adapted interfaces to interact with the social network: content publishing and retrieval, and content adaptation, performed by the Social Enabler; and policy control, auditing, and credential management services performed by the Social Watchdog.

Besides we benefit from the already existing services offered by the home gateway in the home network: service discovery, access, and management, policy management, firewalling, etc.

The architecture has a middleware where basic and extended social services are composed and accessed by the social applications, which: identify the relevant content; handle the content; and present the result to the user

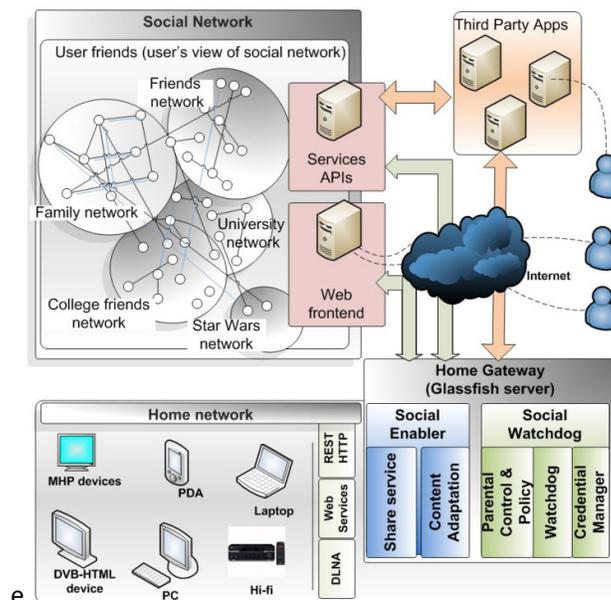


Fig. 1. Social network applications and the home gateway. The home gateway performs requests to Social Networks on behalf of the home devices.

Content identification may be performed either locally to the application or it can be discovered using the existing social services. Local identification may be achieved as described in the application Film Board as explained below in this section. Content handling may require publishing or retrieval from the social network. Presentation may require content adaptation to the device capabilities. These two are explained in the following subsections.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

The interaction of the home application with the middleware follows the REST model, so that requests carry the state with them. The application life-cycle is defined by two states, one where the requests are formed and sent to the middleware components, and other where the response is received and processed. Of course, this is a simplification and we usually have initialization states, for instance to identify the user and to set the user preferences, and finalization states. But we will always try to define the requests and responses respecting as idempotent. This allows that they may be sent several times (duplicated) and the result should be the same as if a single request and response. Timeouts and caching strategies are thus often used.

Requests always carry the name of the user, the operation requested, and the URI of the addressed resource. The ID of the device and the application ID may also be required.

Let us illustrate the architecture using an example: the *Film Board*, an application to test the Social Enabler and the Social Watchdog components. The Film Board collects feeds (comments) from social network users about movies and offers content identifiers to third party applications, like movie identifiers extracted from the IMDB. Content providers add tags containing Film Board's identifiers to content. Then, another application in the set-top box extracts identifiers from content and requests feeds to Film Board through the Social Enabler. The Social Enabler authenticates on behalf of the user and downloads feeds through Services API. In this way, a set-top box can show friend's comments about a movie or suggest a movie based on those comments

II. SOCIAL ENABLER, SOCIAL WATCHDOG AND IMPLEMENTATION

In this section we detail the main two components of the architecture: the Social Enabler and the Social Watchdog. Besides we give some implementation details on the prototype we have developed to show the feasibility of the approach.

A. Social Enabler

Both the Social Enabler (SE) and the Social Watchdog (SW) reside in the Home Gateway (HG). To target the majority of devices at home network, both SE and SW offer their services over different network interfaces and protocols as HTTP, Web Services and DLNA. The SE is in charge of retrieving contents from social networks on behalf of a user and presenting the content in an appropriate way considering device limitations.

1) Content Adaptation

Home devices are quite heterogeneous presenting different visualization areas, resolution, form factors and input devices. For that reason, a content adaptation service is needed.

Home devices willing to display information from social networks send requests to the SE. A request contains the username, the resource to be requested, a set of device constraints, and the operation requested (content adaptation). The SE authenticates to the social network on behalf of the user and retrieves the content. The authentication is of course performed in collaboration with (using the service offered by) the SW.

Resources requested are typically live streams and media (pictures, video). The live stream is list of feeds containing the activity of a group of nodes of a social network as family sub-networks or fans of Star Wars sub-networks. It can be also related to a topic (for instance a movie).

The social enabler uses a plug-in system to adapt live streams to an appropriate format. The SE can adapt live streams to HTML, DVB-HTML [8] and XML. HTML targets any web browser capable of rendering HTML including style sheets and ECMAScript and presenting media using Flash, JavaFX or a media player. DVB-HTML is used to present contents in devices implementing MHP 1.1. Finally, XML is used for previous MHP versions, JavaTV, OCAP [9] and others.

2) Share Services

Home devices might use the SE to share comments or media as pictures or video. Contents can be added to live streams accessible by several node sub-networks, for instance, a Hi-Fi system might post a comment as "Now hearing..." with the title of a song. A game console might publish the score or a screen capture to a gamers network. Requests are sent directly by home devices to the SE indicating the username, live stream, content type to be shared, and the operation requested (sharing).

The content itself is uploaded to the home gateway using HTTP or DLNA (Digital Media Controller profile). Upon

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

reception, the SE requests credentials and authorization from the SW which checks whether the device is authorized to publish content on behalf of the user or not.

B. Social Watchdog

The Social Watchdog (SW) is responsible for managing the home users' credentials, enforcing security and privacy policies and performing the parental control. This requires that every home user is registered in the SW, and also that devices and applications requiring interacting with the social networks are also registered. If some request from an unregistered user, device or application gets to the SW, a default policy for these cases will be applied, and usually a Denied request will be returned. The registration of users should be performed at the gateway, whereas devices and applications can be easily registered by authorized users.

1) Credential Manager

The SW handles and stores user's credentials for social networks acting as an authentication proxy for home devices preventing the user to give his/her credentials to every device at the home network. The SE relies on the SW for authenticating to social networks. Typically, home devices such as TVs, set-top boxes, media gateways receive input from user through a remote controller that lacks of an alphanumeric keyboard. For that reason, the SW associates credentials to a numeric PIN or other authentication suitable for a remote controller. This PIN will be requested by the SW when necessary.

The SW maintains a set of sessions with the social networks. Subsequent requests from the same device or application are served faster, because authenticating the user is not required unless the session expires. The number of sessions that may be stored in the SW is limited, as there is also a limit on the number of sessions per user.

Credentials for social networks typically involve pairs of user/password. Presently we are storing them directly in a database at the SW, but an identity management provider (IdP) infrastructure based in SAML, Liberty Alliance, or even OpenId may be used.

2) Parental Control and Policy

To watch over user's privacy, the SW maintains a policy that prevents devices from misbehaving and children from publishing inappropriate information. Every share request must be acknowledge by the SW which checks device permissions against the defined policies.

Different security and access control policy definition languages have been defined, being most of them declarative. Some are based in XML, among them we can cite the OASIS standard XACML [10] or PERMIS [11], defined by the NIST and based in roles. Others are declarative based in logic like the Security Policy Assertion Language (SecPAL) [12] from Microsoft. Others are object-oriented like Ponder [13, 14], and there others using databases to establish the relations among the users, resources and the authorization decisions.

The home gateway may use any policy definition language, and the SW is responsible for enforcing the policy retrieved from the home gateway. We use a neutral form as described in

[15] to ensure compatibility with the different policy languages. Following this approach, policies from different Access Control Engines (ACEs) may be combined. ACEs are required to extract policy items and requirements from their own controlled set of policies, so that all can be combined together. Besides this flexibility, we can formulate simple policy item statements like: "A set of devices (S_i) is allowed to publish textual comments (T) in a sub-networks (N_i)". Meaning that every request to publish a text feed in N_i coming from a device ID in the set S_i will be granted.

1) Watchdog

The SW generates periodical reports on social network activity. Home users can access to a summary of what has been published by itself and its devices. Moreover, the watchdog also generates reports on what friends publish about the user.

To accomplish this watchdog process, the SW builds a directed graph with a single path between the user and any node using node relations. First, the user configures the SW giving a weight measuring the relation strength of the user and each sub-networks. If it happens that a node is a member of different sub-networks, its strength value is the result of adding the weights of each sub-networks. Once the strengths of nodes have been computed, the weight of each sub-networks is recalculated as the average of its constituent nodes. The SW selects the path which gives the highest sum of relation strengths between the user and any node. The watchdog periodically fetches node activity looking for information related to the user. Node inspection is performed with a frequency proportional to the sum of relation strengths in the path divided by the number of nodes in the path.

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

There are policies establishing the users authorized to access what watchdog reports in the home gateway, but the default policy establishes that a user is always authorized to access the reports on her activity, and to set the preferences on sub-networks weights and the limits on the amount of reports to collect.

C. Implementation

Social Enabler and Social Watchdog have been developed using Java and deployed in a Glassfish server. To test both applications we developed a JavaTV Xlet that requests contents from Film Board and a popular social network and adapt the content using CE-HTML and XML. Figure 2 shows a screen capture of an Xlet using Social Enabler.

The message exchange for our testing application, called FilmBoard, is outlined in Figure 3. This application resides outside the Social Network. The DVB Receiver, acquire contents from the content provider and parses MPEG-2 Transport Stream tables. The receiver finds a tag about the movie that can be use to fetch feeds from a Social Network about the content. The DVB receiver invokes the social enabler to get feeds from the Social Network whereas request permission to the Social Watchdog.

The Social Watchdog requests feeds from the Social Networks which invokes the Filmboard external application and coalesces friends' comments into a single response.

When the Home Gateway receives the response from the Social Network, it uses Social Enabler's adaptation services to generate an appropriate presentation to be consumed by the DVB receiver.

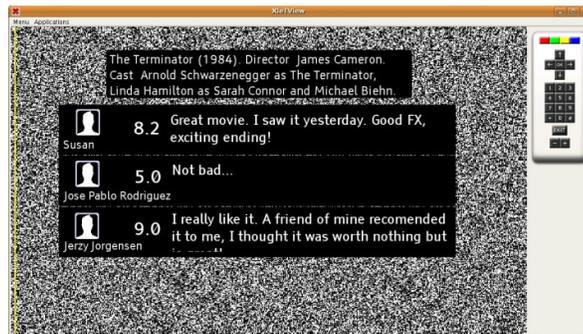
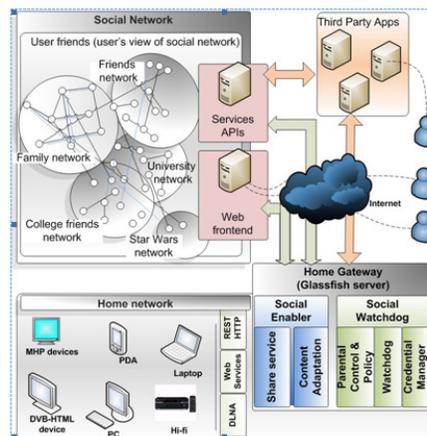


Fig. 2. Some comments (feeds) about a movie. Xlet developed for JavaTV requesting feeds to Social Enabler. Feeds are provided in XML



organizing techniques to enable users of similar interest to automatically come together and form clusters at a specific location in the conference hall. Other devices use the profile data introduced by the user to facilitate relationships to support meetings and social events like the nTAG provided by AllianceTech.

There are a number of devices in the consumer market to facilitate interactions of the user with social networks. Most PDAs and mobile phones are equipped with different types of software, from standard web browsers to sophisticated messaging applications or picture uploaders. Some of them include the possibility of revealing the

DST Sponsored National Conference on Recent Advancements on Computer Science (CONRACS 2019)– 26 to 28 July 2019

geographical location of the user. Besides several manufacturers of video cameras are shipping extra software for the user's PC to speed picture and video downloading, and including special codecs required by video portals.

RendezBlue.com is a social network which connects people in your proximity using a Bluetooth application in your mobile phone, charging you a small amount for every accepted invitation in your buddy list.

None of them addresses home devices, with their potential interest being so close to the user and having so much personal information. Besides, none of the aforementioned works in privacy topics takes into account different users including teenagers, and the possibility of defining and enforcing parental control and policies.

VI. CONCLUSIONS

Social networks are an emerging force in the Internet. Users find them a useful tool to keep in touch with friends and relatives, and with other people sharing the same interests. They already have a clear impact in commercial campaigns. Recognizing the importance of this communication vehicle, the industry is offering solution for users to interact with the social networks.

We propose using home devices, rich in user information, to interact with social networks on behalf of the user. Home devices like set-top boxes and media gateways are good candidates to run social applications, to facilitate interactions with friends and relatives in social networks. However this interactions have to follow policies to safeguard users' privacy while maintaining the complexity low for the average user.

Social networks have strong concerns with user privacy, and different techniques like anonymizing the graph of the network and differential privacy are being studied, together with pragmatically approaches aiming at restring the API offered to third party applications. In this work we deal with policies enforcing that only approved content may be shared in the network.

We have prototyped an architecture based for home devices to participate in social networks. Our home gateway applications use established standards like REST HTTP, WEB services, or DLNA. They help on the convergence of devices and social networks. They allow home devices to actively participate in social networks on behalf of the user. Moreover they respect and care privacy of home members and also monitor user's personal data available on the social networks alerting users when necessary.

We are evaluating two different areas of future work: policy management and risk modeling. We aim at improving the policy management for the home user, so that he/she can monitor and understand how the Social Watchdog applies policies from reading the reports in the Home Gateway, and also that he/she can understand the implications of defining new policies or modify the existent. As a second area of future work, we are exploring the possibility of defining a risk model that can help the Social Watchdog in the authorization process.

REFERENCES

- [1] Foster, P. "Rage Against the Machine beat X-Factor's McElderry to Christmas No 1", The Times, 2^{1st} December 2009.
- [2] Barabási, A. L., Bonabeau, E., "Scale-Free Networks". *Scientific American*, 288:50-59, May 2003.
- [3] Stanley M., "The Small World Problem", *Psychology Today*, 1967, Vol. 2, 60-67.
- [4] Gross, R., Acquisti, A., Heinz, H. J., "Information revelation and privacy in online social networks", Proc. of the ACM workshop on Privacy in the electronic society, Alexandria, VA, USA (2005).
- [5] Backstrom, L., Dwork, C., Kleinberg, J., "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography", in Proc. of the 16th int. conf. on World Wide Web, pp. 181-190, Banff, Alberta, Canada, 2007.
- [6] Krishnamurthy, B. and Wills, C. E. 2008. Characterizing privacy in online social networks. In ACM Proc. of the First Workshop on online Social Networks, pp. 37-42, Seattle, WA, USA, 2008.
- [7] Felt, A., Evans, D., "Privacy Protection for Social Networking Platforms", in Web 2.0 Security and Privacy Workshop of IEEE Symposium on Security and Privacy, Oakland, CA., USA, 2008.
- [8] ETSI, "Digital Video Broadcasting (DVB); Multimedia Home Platform (MHP) Specifications 1.1." TS 102 812, Nov 1, 2001.
- [9] OpenCable™ Application Platform Specification, "OCAP Home Networking Extension", OC-SP-OCAP-HNEXT1.0-I01-050519, May 2005.
- [10] Moises, T. "eXtensible Access Control Markup Language (XACML)" version 2 (2005).
- [11] Chadwick, D. W., Otenko, A. "The PERMIS X.509 role based privilege management infrastructure" in Proc. of the 7th ACM symposium on Access control models and technologies, Monterey, California (2002).
- [12] Becker, M.Y., Fournet, C., Gordon, A. D. "SecPAL: Design and semantics of a decentralized authorization language". Technical report MSR-TR-2006-120. Microsoft research, Cambridge (2006).
- [13] Daminou, N., Dulay, N. Lupu, E., Sloman, M.: "Ponder: a language for specifying security and management policies for distributed systems". Technical report, Imperial College, London (2000).
- [14] Twidle, K., Dulay, N., Lupu, E., Sloman, M.: "Ponder: a policy system for autonomous pervasive environments". In Proc. 5th International Conference on Autonomous and Autonomous Systems. ICAS, Valencia, Spain (2009).
- [15] Díaz-Sánchez, D. Marín-López, A., Almenárez-Mendoza, F. "Enhancing access control for mobile devices with an agnostic trust negotiation decision engine", *Personal Wireless Communications*, pp. 1571-5736, Springer Boston (2007).
- [16] Balasubramaniam, S.; Botvich, D.; Tao Gu; Donnelly, W., "Chemotaxis and Quorum Sensing Inspired Device Interaction Supporting Social Networking", *IEEE 65th Vehicular Technology Conference*, pp. 262 - 266, 2007

Criticality of Role of Optimizers in Machine Learning Techniques

Ajeet K. Jain¹, M. Swetha²,
Data Science Group (CSE),
Keshav Memorial Institute of Technology , Hyderabad 500029

Email: ¹ajeetjain.kmit@gmail.com , jainajeet1@rediffmail.com
²merugu.swethq@gmail.com

ABSTRACT

Deep learning (DL) algorithms in machine learning involve optimization in many contexts. For instance, performance measures of various models require a trade- off for solving an optimization problem. We often use analytical optimization to write proofs or design algorithms. Of all the many optimization problems involved in deep learning, the most difficult is neural network training. It is quite common to get involved into which optimizer is best suited for what and which kind of application in particular? The paper tries to delve into the peculiarities of these kinds of problems and also high lights a specialized set of optimization techniques - such as **AdaGrad**, **RMSProp** and **Adam** - which have been developed for solving DL and ML issues.

Keywords: Machine Learning, Gradient descent, Back propagation, Knowledge representation

INTRODUCTION

An Artificial Neural Network (ANN) is a group of connected I/O units where each connection has a weight associated with its computer programs. It helps you to build predictive models from large datasets. This model builds upon the human nervous system and helps us to conduct image understanding, human learning, computer speech, etc. The use of back-propagation is the essence of neural net training. It is the method of fine-tuning the weights of a neural net based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows us to reduce error rates and to make the model reliable by increasing its generalization. Fig. 1 depicts these concepts.

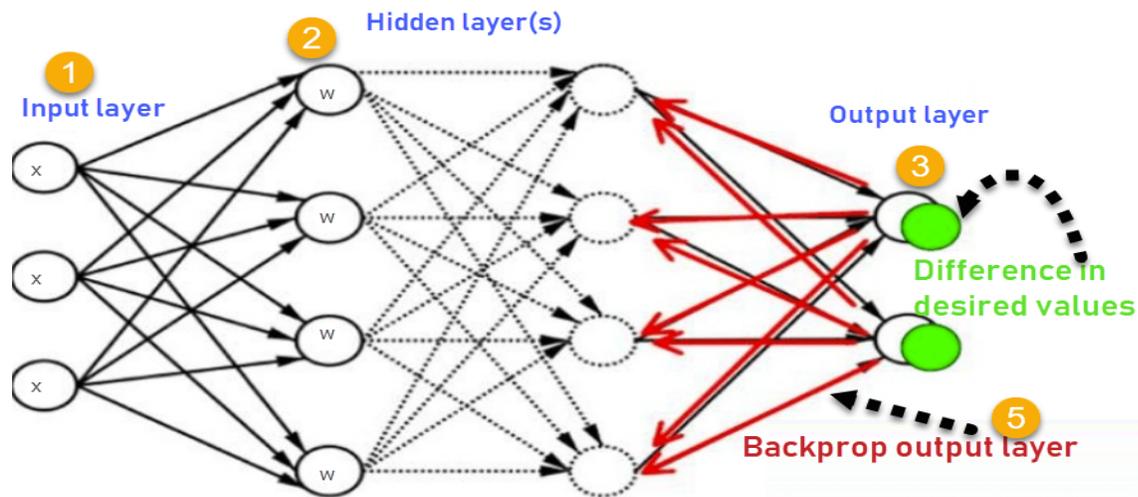


Fig. 1 Forward and backward NN

The goal of machine learning and deep learning is to reduce the difference between the predicted output and the actual output. This is also called as a Cost function or Loss function. For a deep learning problem, we usually define a loss function first. Once we have the loss function, we can use an optimization algorithm in attempt to minimize the loss. In optimization, a loss function is often referred to as the objective function of the optimization problem. By tradition and convention most optimization algorithms are concerned with minimization. If we ever need to maximize an objective there is a simple solution - just flip the sign on the objective function. The criticality of the role of optimizers are discussed in this paper taking 3 well known optimizers; viz. **AdaGrad**, **RMSPorp** and **Adam** and also their suitability with various challenges of using optimization in deep learning. Fig. 2 shows a block diagram of optimizer.

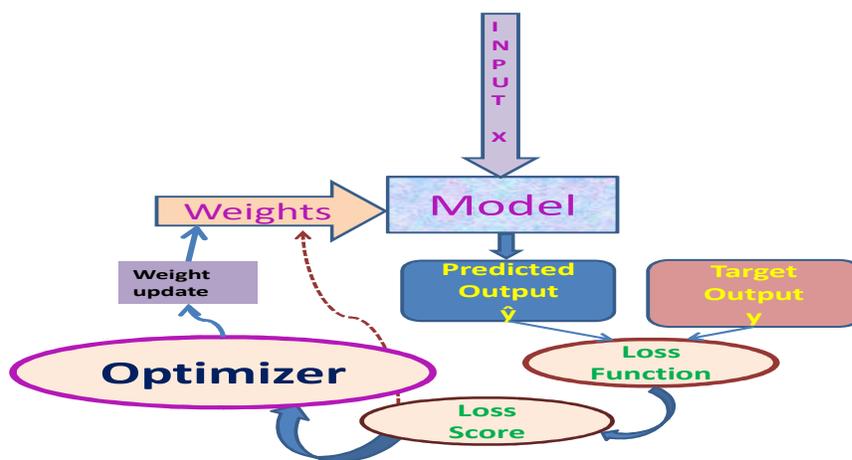


Fig.2 Schematic of Optimizers

OPTIMIZATION CHALLENGES

There are many challenges in ML and DL optimization – which typically include - local minima, saddle points and vanishing gradients.

Local Minima

For the objective function $f(x)$, if the value of $f(x)$ at x is smaller than the values of $f(x)$ at any other points in the vicinity of x , then $f(x)$ could be a local minimum. If the value of $f(x)$ at x is the minimum of the objective function over the entire domain, then $f(x)$ is the global minimum. For example, given the function:

$$f(x) = x \cdot \cos(\pi x) \text{ for } -1.0 \leq x \leq 2.0$$

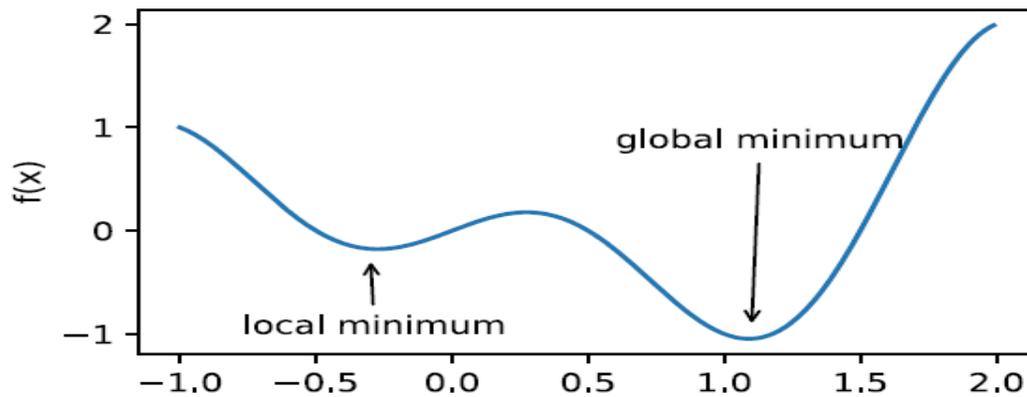


Fig. 3 Local and global minimum

The objective function in ML models usually has many local optima. When the numerical solution of an optimization problem is near the local optimum, the numerical solution obtained by the final iteration may only minimize the objective function locally, rather than globally, as the gradient of the objective function's solutions approaches or becomes zero. Further, some degree of noise may knock the parameter out of the local minimum. Indeed, this is one of the beneficial properties of stochastic gradient descent where the natural variation of gradients over mini-batches is able to remove the parameters from local minima.

Saddle Points

Saddle points are another reason for gradients to vanish. A saddle point is any location where all gradients of a function vanish but which is neither a global nor a local minimum. Consider the function $f(x) = x^3$. Its first and second derivative vanish for $x = 0$. Optimization might stall at the point, even though it is not a minimum - Fig.4.

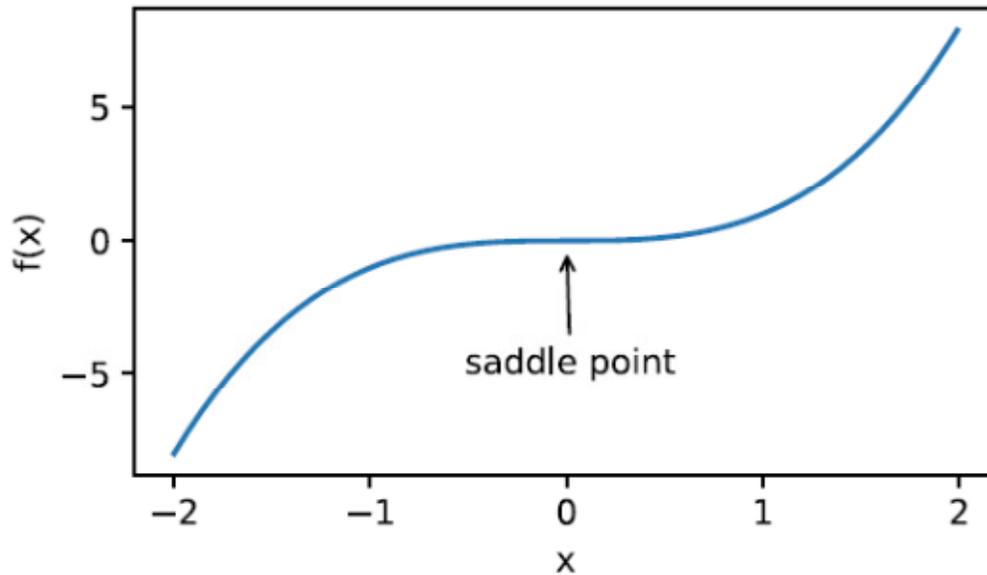


Fig. 4 Saddle points

Vanishing Gradients

Probably the most subtle problem to consider is vanishing gradient. For instance, if we want to minimize the function $f(x) = \tanh(x)$ and starting at $x = 4$. For this value, the gradient of f is close to nil. More specifically $f'(x) = 1 - \tanh^2(x)$ and thus $f'(4) = 0.0013$. Consequently optimization will get stuck for a long time before we make progress. This in turns out to be one of the reasons that training ML and DL models is quite tricky prior to the introduction of the **ReLU** activation function.

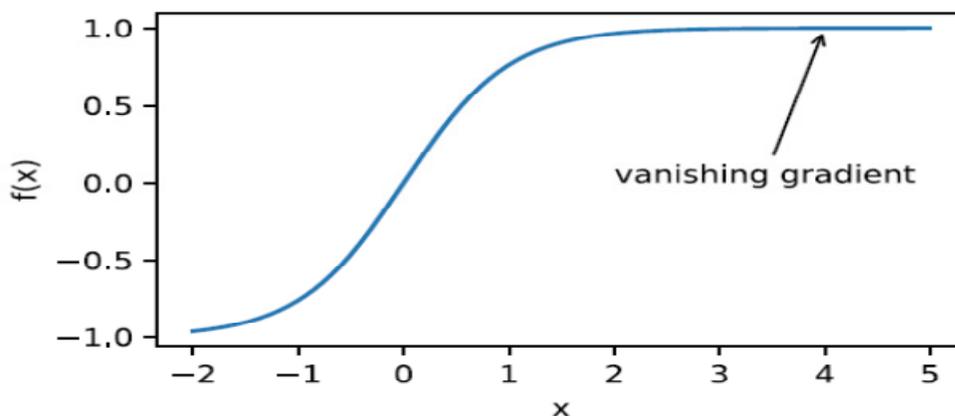


Fig. 5 Vanishing Gradient

Gradient Descent

GD is one of the most popular algorithms to perform optimization and mostly used in ML and DL techniques. Gradient descent is a way to minimize an objective function $J(\theta)$ parameterized by a model's parameters $\theta \in \mathbb{R}^d$ by updating the parameters in the opposite direction of the gradient of the objective function $\nabla_{\theta} J(\theta)$ w.r.t. to the parameters. The learning rate η determines the size of the steps we take to reach a (local) minimum. In other words, we follow the direction of the slope of the surface created by the objective function downhill until we reach a valley.

Role and Criticality of Optimizer

Optimizers update the weight parameters to minimize the loss function. Loss function acts as guides to the terrain telling optimizer if it is moving in the right direction to reach the bottom of the valley, the global minimum. Momentum helps accelerate Gradient Descent (GD) when we have surfaces that curve more steeply in one direction than in another direction. It also dampens the oscillation- as depicted in Fig. 6.

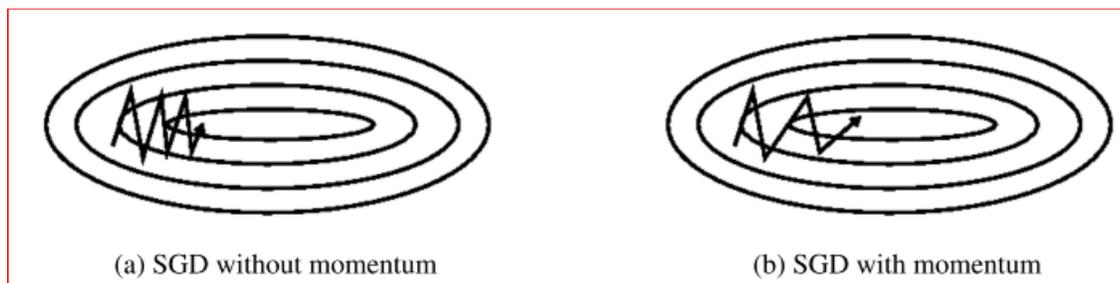


Fig. 6 Momentum

Convergence happens faster when we apply momentum optimizer to surfaces with curves.

Adagrad — Adaptive Gradient Algorithm

Adagrad is an adaptive learning rate method. In Adagrad we adopt the learning rate to the parameters. We perform larger updates for infrequent parameters and smaller updates for frequent parameters.

It is well suited when we have sparse data as in large scale neural networks. GloVe word embedding uses Adagrad where infrequent words required a greater update and frequent words require smaller updates.

We also use the same learning rate η . In Adagrad we use different learning rate for every parameter θ for every time step t .

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

G_t is sum of the squares of the past gradients w.r.t. to all parameters θ

RMSProp

RMSprop is an unpublished, adaptive learning rate method proposed by Geoff Hinton in Lecture 6e of his Coursera Class. RMSprop and Adadelta have both been developed independently around the same time stemming from the need to resolve Adagrad's radically diminishing learning rates.

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t$$

Adam

Adaptive Moment Estimation (Adam) is another method that computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients v_t like Adadelta and RMSprop, Adam also keeps an exponentially decaying average of past gradients m_t , similar to momentum. Whereas momentum can be seen as a ball running down a slope, Adam behaves like a heavy ball with friction, which thus prefers flat minima in the error surface. We compute the decaying averages of past and past squared gradients m_t and v_t respectively as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

m_t and v_t are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively, hence the name of the method. As m_t and v_t are initialized as vectors of 0's, the authors of Adam observe that they are biased towards zero, especially during the

initial time steps, and especially when the decay rates are small (i.e. β_1 and β_2 are close to 1). The authors propose default values of 0.9 for β_1 , 0.999 for β_2 , and 10^{-8} for ϵ . They show empirically that Adam works well in practice and compares favourably to other adaptive learning-method algorithms.

Discussion:

The paper has delved into the criticalities of optimizers in ML and DL and there is no single unique solution for fitting one- in- all. Different optimizers have their in-built features depending upon the data type and role to play. The Python mostly uses Tensor flow

Acknowledgement

The authors wish to thank the Director – KMIT and HOD and Coordinators to provide the support for the work and attending the conference.

References:

1. Adam: A method for Stochastic Optimization: arXiv:1412.6980v8 ; D P Kingma and Jimmy Lei Ba, July 2015
2. Multi objective optimization using theorem and scalability- E.M. Kasprazk and K.E Lewis ; Journal of optimization , 2010 pp. 34-41
3. Introduction to Machine Learning , Miroslav Kubat, 2nd Ed. Springer 2017
4. Deep Learning with Python; Francois Chollet, Manning Pub. 1st Ed. 2016
5. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Aurélien Géron, O'Reilly 2015