CARMEL VIDYA BHAVAN TRUST'S

# CHRIST COLLEGE PUNE

(Affiliated to Savitribai Phule Pune University)

# NATIONAL CONFERENCE

on

## *"Relevance of Research Methodology in Computer Science"*

*12th & 13th February 2016*

Sponsored by

**Savitribai Phule Pune University**

# PROCEEDINGS

ISBN NO : 978-93-84659-10-3

# Proceeding of National Conference on

# Relevance of Research Methodology in Computer Science

**Editor**

**Dr. R. Mohan**

**2016**
**International E - Publication**

www.isca.me , www.isca.co.in

# International E - Publication

427, Palhar Nagar, RAPTC, VIP-Road, Indore-452005 (MP) INDIA
Phone: +91-731-2616100, Mobile: +91-80570-83382
E-mail: **contact@isca.co.in** , Website: **www.isca.me** , **www.isca.co.in**



© ***Copyright Reserved***
***2016***

*All rights reserved.*

*No part of this publication may be reproduced, stored, in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, reordering or otherwise, without the prior permission of the publisher.*



**ISBN: 978-93-84659-10-3**

# PREFACE

It is with immense pleasure and profound gratitude to God Almighty that Christ College Pune initiated its first National Conference on **'Relevance of Research Methodology in Computer Science'** organized by the Department of Computer Science sponsored by Savitribai Phule Pune University on 12th and 13th February 2016.

Identifying the needs and benefits of developing research skills and thereby understanding and appreciating the value of research, we aimed at providing a beneficial platform to academicians and research scholars to be competent in their respective professional fields. The Conference brought together eminent intellectuals and enthusiastic researchers under one roof to exchange experiences, Knowledge, queries and insights. It was an effective stage for all the concerned stakeholders to introspect about the numerous challenges, dilemmas and relevance facing the current research methodology, enhancing knowledge and thereby significantly contributing to the new areas of technology.

We gratefully acknowledge the incessant support and valuable guidance of. Our Mangement for the refined execution of the conference We sincerely appreciate and thankfully remember the meticulous efforts of the reviewers who supported us to bring quality in research papers, all the experts who gave their valuable inputs, the encouragement of Savitribai Phule Pune University, all the authors for their pains taking efforts and the unconditional health provided by the staff and students of our college for successfully organizing the event. Special thanks to the printers and publishers for their efficient contribution in bringing out this edition.

# PRINCIPAL'S MESSAGE

Dr (Fr) Sony J Chundattu CMI
Principal

Greetings to all.....

'Education is not the learning of facts, but the training of the mind to think' Albert Einstien's , this well known saying has influenced us as educators of the modern world . We understand, it is a high responsibility vested on us to provide a significant meaning to this term, as per the present day need. Education has undergone a sea change from transformation of knowledge, to sharpening of intellect through, careful investigations and exploration of facts. To achieve this aim, the academicians have to primarily upgrade and keep themselves abreast of the technological developments as well as the changing and challenging recent trends in the field. We should comprehend that, dissemination and sharing of new knowledge is an important component of academic learning. Interactions, conversations, discussions and dialogues will definitely give way to confluence of various thoughts and opinions leading to critical reviews and interpretations. From this will emerge crucial conclusions that will effectively contribute to progress and development. It is this thought that initiated the two day national conference here at Christ College, Pune.

It is a great pleasure and privilege for us to be instruments in gathering eminent academicians, scholars as well as enthusiastic researchers from far and near, under one roof and provide ample opportunities to share their thoughts. It is a well accepted fact that development in the field of information and technology is the key to the progress of every nation. Computer science among the branches of sciences is a very exciting, progressing and challenging discipline and has a great impact in the growth of modern society. It brought life at our finger tips and has many fascinating marvels in store for us. In this context, we feel proud, to serve as a catalyst in encouraging young minds, to pursue research in this field and henceforth our topic for the conference- ''Relevance of Research Methodology in Computer Science".

I sincerely hope and wish that the conference provide exciting learning opportunities, endless creativity, proficient thinking, discovery of new knowledge and thus shape resourceful intellectuals....Thankyou....God bless you!!!!

# THE INSTITUTION

Christ College, Pune came into existence in the academic year 2007-2008. The college is affiliated to the Savitribai phule Pune. It has become the first choice of the brightest. The highly disciplined and clean campus provides the right atmosphere for systematic learning. Giving maximum importance to quality education, Christ college aims at the overall development of the students so as to equip them effectively face the challenges of the modern world.

Christ College promotes a worldview in which the entire humanity is one family. It is very secular in its outlook. Students of all caste, creed and language make it their 'home away from home' and enjoy the campus life here, in harmony.

# OUR HERITAGE

Christ College Pune is well rooted in the noble tradition founded on the educational vision of Blessed Kuriakose Elias Chavara, the founder of the CMI congregation, which manages the Institution.

The Carmelites of Mary Immaculate (CMI), the first indigenous religious congregation of India with more than 2000 members, is rendering service to humanity in pastoral, education and social fields, aiming at the integral development of nation.

# OUR STRENGTH

## VISION

Christ College Pune aims at imparting high quality and value based education to shape future leaders in multi-disciplines.

## MISSION

Christ College Pune is nurturing ground for individual's Holistic Development to make effective contribution to the society in a dynamic environment.

Christ College firmly accepts the universal fatherhood of god and thereby the universal brotherhood of humanity. It is a home away from home for all students irrespective of their caste, creed, language, gender or nationality. We constantly strive to provide all-round education to our students to make then constructive and dependable members of the society.

# INDEX


| Sr.No | Title | Page No |
|---|---|---|
| 01 | Big Data Applications: Special Reference to Indian Government Sector<br>Dileep Baragde & Dr.Mukund Sonawane | 01 |
| 02 | Unit Consumption under Fuzzy logic control of washing Machine<br>Prof.Petricia Leema Roseline | 07 |
| 03 | 3D Search Engine<br>Prof. David Thomas Vidhate. | 13 |
| 04 | A Review On Heuristics: Search Techniques In AI<br>Minal Rajendra Pawar | 17 |
| 05 | Automated Boolean Matrix Data Representation Approach for Efficient Mining<br>Renju Samuel and Linu Paulose | 19 |
| 06 | Solution to Design DFA<br>Prof.Nilima Shingate & Mr.Pradip Shingate | 23 |
| 07 | Complex event processing and data mining for smart cities<br>Rupali Memane | 27 |
| 08 | A survey on Web mining: Data mining approach<br>Nilesh Magar & Aniket Nagane | 31 |
| 09 | A Review On Heuristics: Search Techniques In AI<br>Minal Rajendra Pawar | 34 |

| | | |
|---|---|---|
| 10 | Automated Boolean Matrix Data Representation Approach for Efficient Mining<br>Renju Samuel & Linu Paulose | 36 |
| 11 | Collaborative Multiparty Data Mining<br>Deepak Joshy | 40 |
| 12 | Cryptography and computer Security<br>Prof.Sofeen Sayyed. | 44 |
| 13 | Cipher text conversion with encrypted polynomial equation<br>Dr. Kavita Khobragade & Shounak Joshi | 48 |
| 14 | Scope of Visual Basic for Applications in Data Analysis and Automation<br>Prof. Leena Sanu and Mr.Sanu Poulose | 51 |
| 15 | A Review on Forward and Backward Reasoning in AI<br>Kiran Umesh Kadam | 55 |
| 16 | Future and possibilities in Wireless Communication<br>Dr Prashant P. Mulay | 58 |
| 17 | Fuzzy Model Evaluation for Intensify Multimedia Data Security in Cloud Enviornment using Homomorphic Cryptosystem<br>Prof.Anjuna D Raj | 62 |
| 18 | A Review on Game Playing in AI<br>Sagar Sanjay Khomane | 69 |
| 19 | Technology : future application of lifi<br>Minal Chiranjit Dutta & Mrs.KavitaMahajan | 71 |

| | | |
|---|---|---|
| 20 | Big data<br>Prof. Pradnya Deshpande | 74 |
| | Network Security: A modified Playfair cipher based on Sparse Matrix | 78 |
| 21 | Kiran Tikar ,Dnyaneshwar Tikar & Sandeep Patil | |
| 22 | Predicting Patients Curability using Data mining Signatures S. Sreeji ME.;MBA | 83 |
| 23 | Review on Artificial Intelligence<br>Amit.R.Atole | 87 |
| 24 | Implementation Of Protus2.0 By Ontology For Advanced E Learning<br>Japala Saikrishna Rao | 88 |
| 25 | Secure Information Using Cryptography Based Authentication methods<br>Prof.Jankar Jayshree Prakash Lecturer | 91 |
| 26 | Segmenting and Hiding Data Randomly using Image Steganography<br>Ankita Rameshchandra Mithaiwala & Prof. Asha Hemaush Nagoriya | 96 |
| 27 | The study of Shimin wei Digital signature scheme based on Integer Factorization and discrete logarithm<br>A.B.Nimbalkar, Gadekar M.J. | 102 |
| 28 | Transposition Technique with Modified Caesar Cipher<br>Dr. Kavita Khobragade, Shubhankar G Keskar | 106 |
| 29 | Secure Data Deduplication in Cloud Computing<br>Mini Joswin | 110 |

| | | |
|---|---|---|
| 30 | A review in virtual private network<br>Asst.Prof. Shailesh Khandare | 115 |
| 31 | Quality improvisationin Game design techniques throughGame Development life Cycle (GDLC) Model.-<br>Author: Prof Sarita A Dhawale | 120 |
| 32 | A Watermarking Technique Without Data Modification for<br>Process Mining<br>Tintu Mary Xavier | 126 |
| 33 | Video Inpainting Forgery Detection: Blind Detection Method<br>Sree lakshmi Das | 129 |
| 34 | Classification Model for Intrusion Detection<br>Prof. Dheeraj Tandon<br>Prof. (Dr.) Harsh Dev | 135 |

# Big Data Applications: Special Reference to Indian Government Sector

**Dileep Baragde**
Ph.D. Scholar & Faculty,
G. S. Moze Senior College,
Pune
Email: baragde.dileep@gmail.com
Mobile: 9922446907

**Dr. Mukund Sonawane**
Asso. Prof. & HOD of Business Practices,
Ness Wadia College, Pune & Doctoral Guide,
SP Pune University, Pune
Email: sonawanemukunda@gmail.com
Mobile: 9890052318

**ABSTRACT:**

**Big data is used in multiple ways in many industries. Big data understands the business challenges and dynamics of the industry and it can help us to make use the most of all our information. When industries can analyze all of their available data, rather than a subset, they gain a powerful advantage over their competitors. Two things are required while dealing with big data such as Inexpensive, reliable storage. It analyzes new tools for structured and unstructured data. In many ways big data can help the Indian Government to implement more and go for the variations and implementing the different policies and government schemes. AADHAAR card project is example of the big data that will help people to get benefit of various schemes under it like education, scholarships, Biometric Ration card, and Pahal (DBTL) scheme. It can take Indian Railway to a new level by Improving System Performance at Scale. State Bank of India (SBI) is working on a project to take advantage of big data to solve key issues related to borrowers and predicting defaults. Airlines should use big data more effectively to boost ancillary revenue. It can also create an impact on Election and Voting system, Health care system Government projects etc.**

**Key words:** big data, big data applications, big data in government sector;

## Introduction

Big Data refers to datasets whose sizes are beyond the ability of typical database software tools to capture, store, manage and analyze. Big Data technologies as an innovative generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS".

Hadoop is a different sort of tool. Hadoop is aimed at problems that require examination of all the available data. For example, text analysis and image processing generally require that every single record be read, and often interpreted in the context of similar records. Hadoop uses a technique called Map Reduce to carry out this exhaustive analysis quickly. Hadoop takes advantage of this data distribution by pushing the work involved in an analysis out to many different servers. Each of the servers runs the analysis on its own block from the file. Results are collated and digested into a single result after each piece has been analyzed. Running the analysis on the nodes that actually store the data delivers much better performance than reading data over the network from a single centralized server. Hadoop monitors jobs during execution, and will restart work lost due to node failure if necessary. In fact, if a particular node is running very slowly, Hadoop will restart its work on another server with a copy of the data.

**Apache Hadoop** is a set of algorithms (an open-source software framework) for distributed storage and distributed processing of very large data sets (Big Data) on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (Map Reduce). Hadoop splits files into large blocks (default 64MB or 128MB) and distributes the blocks amongst the nodes in the cluster. To process the data, Hadoop Map/Reduce transfers code (specifically Jar files) to nodes that have the required data, which the nodes then process in parallel. This approach takes advantage of data locality to allow the data to be

processed faster and more efficiently via distributed processing than by using a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

Introduction

According to IDC1, the amount digital information produced in2011 will be ten times that produced in 2006: 1,800 Exabyte's. The majority of this data will be "unstructured" – complex data poorly-suited to management by structured storage systems like relational databases. Unstructured data comes from many sources and takes many forms – web logs, text files, sensor readings, user-generated content like product reviews or text messages, audio, video and still imagery and more. Large volumes of complex data can hide important insights. Are there buying patterns in point-of-sale data that can forecast demand for products at particular stores? Do RFID tag reads show anomalies in the movement of goods during distribution? Do user logs from a web site, or calling records in

A mobile network, contain information about relationships among individual customers? Can a collection of nucleotide sequences be assembled into a single gene? Companies that can extract facts like these from the huge volume of data can better control processes and costs, and better predict demand and can build better products. Dealing with big data requires two things: Inexpensive, reliable storage; new tools for analyzing unstructured and structured data. Apache Hadoop is a powerful open source software platform that addresses both of these problems. Hadoop is an Apache Software Foundation project. Cloud era offers commercial.

**Background**

Big data can be described by the following characteristics:

Volume: The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size.

Variety: It means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. It helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity: It refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability: It is a factor which can be a problem for those who analyses the data. It refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Veracity: The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity: Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data.

BIG DATA AND INDIA

In number of ways big data can assist the Indian Government to develop further and go for the changes and implementing the number of government policies and schemes. AADHAR is one of the biggest sources of big data scheme, a unique identification project. It is commenced by the Unique Identification Authority of India (UIDAI), is an agency of Government of India, which was established in February 2009, and will own and operate the unique identification Number Database. The authority is planning to provide the unique identification numbers to all Indians, but won't be providing the smart cards, that would depend on to the ministries of the country. The authority will maintain a database of residents containing the biometric and other data. Mainly the database would be containing the 12-digit unique number which is going to be issued by the Unique Identification Authority of India (UIDAI) for all the residents, all this numbers will be stored in the centralized database and will be linked to the basic demographics and biometric information – photograph, ten figure prints and iris of each individual. Random numbers generated will be free from any classification based on caste, religion and geography. This database would be very large and big but this will allow those people to have identity who do not have any sort of identification. Now if we talk about the number of entries under this scheme, total number of AADHAARs issued as of 30-Nov.-2015 is 940,372,234. This is 73.5% of the population of India. And that gives us the insight that how big the data would be within the end of this year, and processing and manipulating this data statistics would help the Government to make various decisions based on those statistics. AADHAAR is an ambitious government Big data project which is going to be one of the largest biometric database in the world by 2014, with the global capturing about 600 million Indian Identities and dealing with this kind of database is not an easy task, but India is going to be one of the first country who would be dealing with this much large amount of public biometric database.

**Literature Review:**

In 2004, Wal-Mart claimed to have the largest data warehouse with 500 terabytes storage (equivalent to 50 printed collections of the US Library of Congress). In 2009, eBay storage amounted to eight petabytes (think of 104 years of HD-TV video). Two years later, the Yahoo warehouse totaled 170 petabytes1 (8.5 times of all hard disk drives created in 1995).

Since the rise of digitization, enterprises from various verticals have amassed burgeoning amounts of digital data, capturing trillions of bytes of information about their customers, suppliers and operations. Data volume is also growing exponentially due to the explosion of machine-generated data (data records, web-log files, and sensor data) and from growing human engagement within the social networks.

The growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 Exabyte's of data were created and stored in 2005. The amount grew to 1,227 Exabyte in 2010 and is projected to grow at 45.2% to 7,910 Exabyte's in 2015.The growth of data constitutes the "Big Data" phenomenon – a technological phenomenon brought about by the rapid rate of data growth and parallel advancements in technology that have given rise to an ecosystem of software and hardware products that are enabling users to analyze this data to produce new and more granular levels of insight.

Gathering and processing of huge amount of data is not new, what is new is the speed at which we can now process that data. This is an era of collective intelligence. Every day we create approximately 2.5 quintillion bytes of data, it is estimated that 90% of the data which is been generated till now is created in last 2 years alone, and with the help of that thing we can predict what amount of data the world would be dealing within upcoming years, and that lead us to work with the term that is called "Big data". Big data is one of the Buzzing word all around now, there are lot of companies working on the processing and manipulation of these data, big data usually means the collection of data that is too big and massive generally in petabyte, and that data can come from many different sources like click streams, sensors, and many other places, it must be processed quickly and which is too hard to be managed by all the existing tools to process those massive amount of data records. Same way not only the giant companies deals with the massive amount of data but even the Government of any Country are canonical users of Big Data, as they also keep track of massive number of different records for their country which may include information about the people of the country, their growth and many more things. In this paper I am going to focus on some of the application of the Big Data in the field of Government, Public sector and the society in general. One of the biggest examples of using big data in the field of Government and politics has been given by the president of United States of America, Barack Obama. President Obama used the information of the voters of the previous election, like their email addresses their phone numbers, and team also tried to get feeds of various activities of people through the social sites like twitter and Facebook, and based on those feeds people used to receive emails regarding the presidential campaigns, what are the policies which are going to be developed which will help people in near future, what are the other facilities people will get after the election, People were directly targeted based on their previous data and their activities and they used to receive emails which will help people to understand more about the campaigns, and that lead to victory of President Barack Obama in 2012, and reelected as president of USA. Now we all know that elections are not solely by analytics. The policies of the two candidates contributed to the results, and there may be involvement of some other factors too. The main thing is that how can we use big data thing within the government and use those information and data records do more and more with less and less expanses which will help country and people.

**Discussion and Findings**:

BIG DATA AND ITS APPLICATIONS IN INDIA

Founded in 1853, Indian Railways is the world's second-largest railway network with nearly 7,000 stations and 72,000 miles of track. Every day, the railway carries over 23 million passengers on more than 12,000 trains. Passengers have always been able to purchase reserved tickets at railway stations—often waiting in long lines. In 2002, The Indian Rail Catering and Tourism Corporation (IRCTC), designed an Internet-based ticketing system. Initially, tickets were delivered by courier to customers and in they moved to a completely web-based e-ticketing system where tickets could be printed using a web-based application. In the beginning, the e-ticketing system generated around 100 e-tickets a day, but over the years demand has grown to where the system generates approximately 500,000 tickets daily.

In this they were facing many challenges. While the e-ticketing system definitely made it easier to buy reserved tickets, the experience was far from seamless. Indian Railways' unique Tatkal ticketing system offers the ability to book last minute trips one day before the date of travel starting at 10 a.m. In the past, the e-ticketing system would often crash at 10 a.m. because of Tatkal ticket demand, but the system experienced issues any time there were more than 40,000 concurrent users logged on to purchase tickets. Even when it didn't crash, the system was so slow that it would take customers up to 15 minutes to buy a ticket—and often the system would time out and not allow customers to finish their purchases. Because of the unreliability and slow performance of the e-ticketing system at peak hours, it was easier for customers to travel to train stations and wait in long lines in person to buy Tatkaltickets.It supports More than 200,000 Concurrent

Purchases without Impacting System Performance. In the past, the old system would crash if more than 40,000 users tried to access the system at one time. Now, thanks to the ability of GemFire to manage data in-memory across a cluster of nodes, the new system can easily scale to more than 200,000 concurrent users without impacting performance—even at high demand times. For example, all journeys planning information is kept in local app server in-memory rather than accessed from the legacy back-office passenger reservation system, as was the case with the old system. This has considerably reduced system latency. Provide Stable Performance to Book Approximately 150,000 Tickets per Hour, Compared to 60,000 in the Old System Due to performance issues, Indian Railways used to only be able to sell a maximum of 2,000 tickets every minute. Today, because of dynamic load balancing and improved performance, the system can manage surges in sales of more than 10,000 per minute, such as during holiday periods such as Diwali. And on an hourly basis, average online sales are now nearly two and a half times higher than they used to be, growing from 60,000 per hour in online ticket sales to 150,000 per hour. Transformed Customer Experience so Reservation Transactions Complete in Seconds Instead of 15 minutes. With the old system, it could take customers up to 15 minutes to look at travel options and book tickets. During peak times, the system was often so slow that customers couldn't finish their ticketing transactions and it was easier to actually go down to a station, wait in line and buy a ticket at the counter. The new e-ticketing system offers a vast improvement in customer experience over the old system. Now, user authorization and checkpoints are implemented from memory, streamlining the entire user log-in process. And, the new e-ticketing system offers passengers the choice of 40-50 payment gateways linked to banks when booking tickets. Further, because performance is so stable, almost 50 percent of ticket transactions take less than a second to complete. It shifted Online Purchasing From 50% of Tickets Sold to 65%, Boosting Revenue Generated from E-ticket Sales to INR600 Million Daily. The CRIS team gives GemFire credit for helping to improve online revenue and minimize demand at railway station ticket counters. In the past, 10-20 percent of Tatkal tickets were booked using the e-ticketing system in the initial 10 minutes of Tatkal booking. Today, that percentage has grown to an estimated 60-80 percent of Tatkal tickets sold in this 10 minute period. Overall, online e-ticket sales have also increased from 50% to 65% of bookings, thus boosting daily revenue by INR$600 million for Indian Railways and CRIS expects this number to grow.

State Bank of India (SBI) is working on a project to take advantage of big data to solve key issues related to borrowers and predicting defaults. As country's largest repository of financial information, SBI is looking at how big data, or analysis of large pool of seemingly unstructured data, can help foresee standard borrowers could turn willful defaulters. Willful default bothers us most and it is an area where we are putting in some kind of technology interventions. It is some kind of data mining, capturing of trends and elements of artificial intelligence. The analytical tool would study trends coming from cash flows of customers and other financial information. A borrower who fails to repay a loan despite having the capacity to do so or diverts the money or siphons it off is declared a willful defaulter. Accounting for about 20% of financial assets that exists in country's formal banking system, SBI alone has to recover Rs 9,580 crore, or 41% of the total dues of all willful defaulters. Between 2011and 2013, SBI identified about 660 willful defaulters with loans aggregating to Rs.5,700crore.SBI is also taking help of information technology to develop an automated system that would minimize the cost of zero-balance saving accounts that have been adding to higher cost but not resulting in improving bank's CASA (Current Account Savings Accounts) ratio.

"CASA is under pressure as we are opening a lot of free saving account as part of financial inclusion and now these carry a cost. IT is also known as the Direct Benefit Transfer scheme that aims bringing transparency and eliminating the stealing from the distribution funds. This scheme is dependent on.

Big Data – how airlines should use it more effectively to boost ancillary revenue. As we all know, the airline industry today is extremely interested in two major trends – Big Data and ancillary revenue. Everyone is seemingly talking Big Data, but what is it exactly and how does it help an airline to make money? And then, of course, we have the whole song-and-dance between passenger and airlines around ancillary revenues and customer experience.

A lot of airlines today are fixated on the point that the only way to make more out of their passengers is to charge them large fees for trivial and additional requests and, of course, to call them "stupid" for not reading the fine print. For an industry which practically invented the concept of customer experience as part of the pre-flight and inflight services (remember United hiring its first stewardess in 1930s, or the on-board food services, the romance and adventure of the job) as well as some of the most successful customer loyalty programs (airline miles being at the forefront), it is quite disheartening to see it struggling for survival due to runaway oil prices and employee unions. But just like any other industry sector, it's time to get creative, not combative.

Big Data, as understood and accepted by most today, is high-volume, high-velocity and high-variety information assets which can create substantial economic value and help with operations, decision-making, risk management and customer service.

AADHAAR card that will help people to get benefit of this scheme and it will also ensure the government about the number of people who are getting benefit out of this, it includes various different schemes under it like education,

scholarships etc. By keeping track of this records will help the Government to decide the funding for particular states, and they would also be able to keep track of improvement and the growth within the particular region where people are benefited of this scheme. So if these records are analyzed in a proper manner then this statistics will lead to the improvement, and Government would be able to focus on those areas which are really important, thus how big data can play a vital role there. Biggest example of Big Data and its usage in the Election campaign by the President of United States, Barak Obama, and his team focused on to the right public or we can say registered users as per their database records and based on their activities user used to receive emails and connected through phone calls. President Obama's election team was able to dynamically analyze the global state of pre-election sentiment, and deploy advertising resources and human volunteers to the areas that needed the most attention, virtually in real time. Indian Government can also use the same kind of thing, by analyzing the big data and making policies and the scheme based on those statistics which will help the people of the country as well as the growth of the country.

As there are many new projects are assigned in to the various states of the country every month for the development of the region and the country. It may be possible that Construction or the project may be deficient and may create problem in near future, as some of the projects might not be meeting the inspection criteria. So my point here is that real-time analytics can help the government and the other people alike, let me give you an example for that, The University of Texas is working on sensor technology that use wireless sensors to identify failing bridges, lower the cost of monitoring those bridges, and improve the safety of new bridges, and can report dynamic data measurements on a bridge's condition. They're working on sensors that can survive the constant vibration, weather, and even send and receive data through all the steel that normally would make radio transmission a near impossibility. So the thing would be like there would be lot data which is supposed to be filtered out to get the exact information about the bridges and their conditions, that will be bit complex but at a same time it will save lot of resources like money, human lives, etc. This data can be feed to the system and that system can help in the time of critical situation when the project needs some maintenance or any repair kind of thing. Thus it can help in that way too.

More data about students can help students in various different ways to find their interest, strength, and help to identify student weaknesses in ways that are not possible today. Impact of Big data on to the Education sector can play a really vital role, in the normal case generally school boards decide which text books students are required to follow for their particular subjects and all, but do we know which text book is actually good or bad? No, we depend upon expert opinions. But now we do have the data, as are moving from the hard cover books to the electronic books, as people are using the tablets and the other electronic book delivery mechanism like ipad or Kindle from Amazon etc. They all usually collect the information. If we take an example of an Amazon then it know how long you stay on each page, when you re-read a page, or you read chapter over and over when you have some difficulties within that. All information is kept within Amazon. Those information would be really help a lot to the author or the publisher of the book to improve the things within text books, and that can also help the Board of Education to make some worthy decision based on those statistics, and that will ultimately help the students, same way to the people of the country.

**Conclusion:**

Big data can be really helpful if the data or records are analyzed carefully and if we use that statistics in a right way, it can really help the developing country like India. This is also a very important area where big data can play a very huge role and can save tons of money. By having an access to all the medical records of the past decade or more, can help government to take a proper decision in this area. When this user data is processed in the right manner and analyzed correctly, it will result into the complete statistics on which part of the country are suffering from what kind of health related issues and disease. And that can help the government to start new services and can also lead to opening the new hospitals to serve the patients; Government can also issue some more funding for the research and development of some life saving medicines, this way government can help in health care area. To arrive to a proper decision, one need to have a faster access to information, this could be sufficed by using data warehouse techniques and big data. Data analytics can be more helpful to the government while taking policy decision. Big Data has gained tremendous prominence in the corporate world and could be therefore very helpful in government decision making process. Several areas such as predicting weather devising price for agricultural produce (eg MCXagri), deciding poverty line and number of people below poverty line could be achieved. We have a huge data pertaining to railways and citizen of India (AADHAR) card, government can make use of this data in effectively implementing policies for the right and deserving people of India. In the years to come, big data will definitely help countries in taking proper policy decisions and implementing them. Generating revenue from Government sites With the Big data technology, Governments can also earn from their running web sites, as they are providing different kind of advertisement on the government sites, and many of the advertisement may be irrelevant. For example site of Indian Railway Catering and Tourism Corporation limited (IRCTC), they deal with the tourism things, they may earn if they provide tourism relevant advertisement. With the help of big data they can provide the personalized ads,

which would be more relevant to the people who are using such sites, and that will usually increase the rate of generating revenue compare to those irrelevant advertisements.

**References**


1. Ames Manyika, Michael Chui:*Big data: The next frontier for innovation, competition, and productivity*, McKinsey & Company 2011
2. Applications of big data in government Sector, by Hitesh Parmar, Retrieved 2016, January 12, from http://www.academia.edu/5092180/ applications_of_big_data_in _government_sector
3. Chandrasekaran, N. Annual Report 2014-15 NASSCOM Retrieved 2015, June 18 from http://www.nasscom.in/sites/default/files/NASSCOM-Annual-Report-2015_0.pdf
4. Data Security Council of India, NiryatBhawan,Rao Tula, *NASSCOM Cyber Security Task Force Briefing Paper*, Prepared for the CSTF meeting on 2015 May 25.
5. Government of India, Ministry of Communications & Information Technology, Report of the Working Group on Information Technology Sector Twelfth Five Year Plan (2012–17). Retrieved2014, August20, from: http://planningcommission.gov.in/ aboutus /committee/wrkgrp12/cit/wgrep_dit.pdf
6. Hadoop apache, Retrieved 2016, January 07, from http://www-01.ibm.com/software/data/ infosphere /hadoop/index.html
7. International Conference on Big Data Management & Cloud Computing, 28th July 2013, Bhubaneswar, ISBN: 978-93-83060-09-229, Retrieved 2016, January 10, http://www.planetdata.eu/sites/default/files/presentations/ Big_Data_Tutorial_part4. pdf
8. Nik Rouda, Senior Analyst, Getting Real About Big Data: Build Versus Buy, Retrieved 2016, January 10, http://www.oracle.com/us/corporate/analystreports/esg-getting-real-bigdata-2228170.pdf
9. State/UT wise Aadhaar enrolment Ranking, India, Retrieved 2016, January 16, from : http://uidai.gov.in/images/news/state_wise_aadhaar_ enrolment_ranking_as_on_30_november_2015.pdf .
10. Zikopoulos Paul and Eiton Chris, "*Understanding Big Data*", New York: The McGraw-Hill Companies, 2012.
11. Zhao, J., Ordonez de Pablos, P., & Tennyson, R. D. (2015).Organizational Innovation and IT Governance in Emerging Economies (pp. 1-304). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-7332-8

# Unit Consumption under Fuzzy logic control of washing Machine

**S. Petricia Leema Roseline**
Dept of Computer Science
Christ College-Pune
India-411014
leema_roseline@yahoo.com

***Abstract***
**In industry different types of control techniques are used in various fields. Using Fuzzy logic, complex systems can be controlled more effectively by the designers than using traditional approaches. In this paper, one of the household item, washing machine is proposed to design the fuzzy logic controller having three input ( cloth- type, detergent, dirteness) and two output (wash time, speed). This output is further given as input to calculate unit consumption. The objective is to save output variables viz, water and power consumption.**

Key words: Fuzzy Logic, MatLab, Optimization, FIS editor, Automatic sensing

## 1. Introduction

The common house hold item washing machine to have effective control the utilization of water and electricity is vital. The controller's viz., conventional, proportional, integral and differential have proven less capability when we apply fuzzy logic control system. In industry fuzzy logic helps in inventory control (Demand and Supply), increase productivity etc. L. A. Zadeh[1] (1965) presented fuzzy sets. Mamdani[2] (1974) proposed application of fuzzy algorithm for control of simple dynamic plant. Zhen and Feng[3] (2012) suggested the design of neural network fuzzy controller in washing machine. George and Yuan[4] (1995) developed fuzzy sets and fuzzy logic. Design of an improved controller microchip for washing machine by Lohani and Hasan[5]( 2009). Virkhare and Jasutkar[6] (2014) developed Neuro-Fuzzy controller based washing machine. Fuzzy logic based control system suggested by Kumar and Haider[7](2013). Mendel[8](1995) presented fuzzy logic system for engineering. Alhanjouri and Alhaddad[9] proposed optimize wash time of washing machine using fuzzy logic. Agarwal[10] (2007) suggested fuzzy logic control of washing machines.

In this paper, fuzzy logic control of washing machine with three inputs and two output model is developed. This paper suggests the idea of controlling water and electricity using fuzzy logic. This paper suggests the procedure that can be used to get suitable wash time for different types of clothes and minimum utilization of electricity.

Fuzzy Logic

Fuzzy logic is a method to formalize the human capacity of imprecise reasoning , or approximate reasoning. Such reasoning represents the human ability to reason approximately and judge under uncertainty. In fuzzy logic all truths are partial or approximate.

Linguistic Variable

A Linguistic Variable is characterized by a quintuple ( x, U, T(x), G, M),
Where
X is a name of a variable.
U is the universe of discourse
T(x) is the term set of x, that is , set of names of linguistic values of x with each value being fuzzy number defined on U.
G is a syntactic rule for generating the names of x
M is a semantic rule for associating with each value its meaning.

Inference

The mathematical procedure to conduct deductive inference of IF-THEN rules can be implemented on computer for processing speed. There are three common methods of deductive inference for fuzzy system based on linguistic rules.
(i) Mamdani systems
(ii) Sugeno systems
(iii) Tsukamoto systems.

## 2. Structure of Fuzzy Logic System

Fuzzy inputs, outputs rule and defuzzification are given in figure 1.

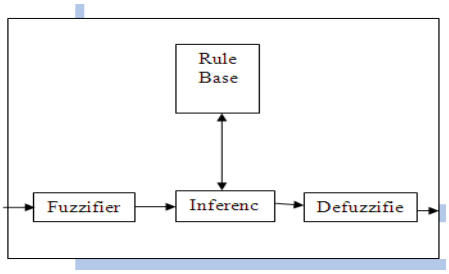


Figure1. Structure of Fuzzy Logic System

Fuzzy takes the system of information in normal language and converts it to values. The inputs are associated with membership functions are in form of words such as tall, short, shortest. Fuzzy system is determined the variables of input and output values such as big , small, biggest, etc. are being selected. Afterwards, rules are being developed. Using those rules the relationship of input –output is developed. Using fuzzified inputs outputs are being produced and hence rules are determined. Those fuzzified outputs must be converted into real values . These real values can be used in real system. This system of procedure is known as defuzzification.

**3. Washing Machine: Principle**

To know the function of washing machine we must know its components. The components are Motor, Agitator, water pump, Sensor fir door safety, Detergent drawer, controller, drain tube, mechanical programmer, and valve for inlet control of water.

**4. Proposed Design:**

Here the input parameters are type of cloth, amount of detergent and type of dirteness to achieve economical wash of the washing machine. Figure 3 describes the input and output parameters, membership function. Here there are 27 rules have been established. The rules are shown in Table 1 using centroid method in MATLAB.

Case I.

Input parameters:

1. Type of cloth
2. Detergent
3. Dirtiness

Output parameters:

1. Wash Time
2. Wash Speed

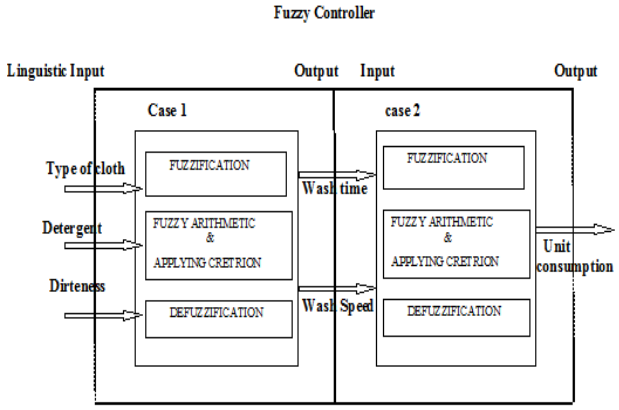
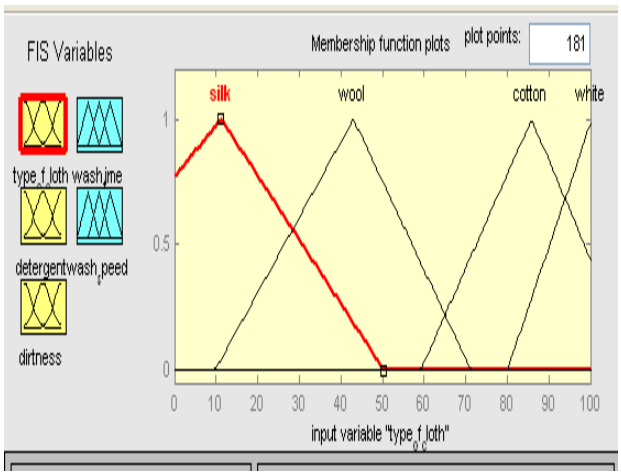


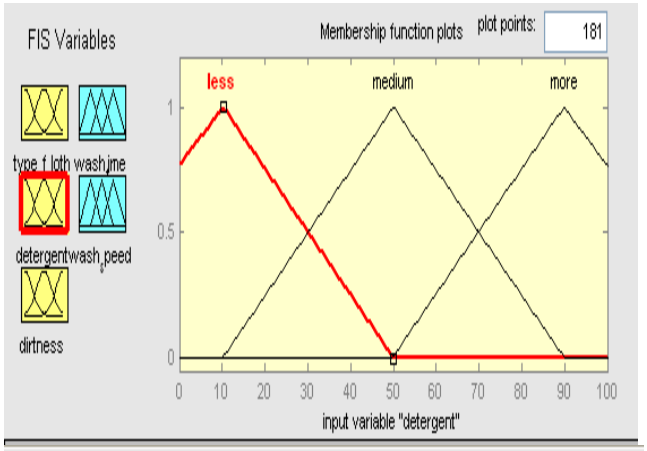
Figure. 3 Flowchart for Washing Machine

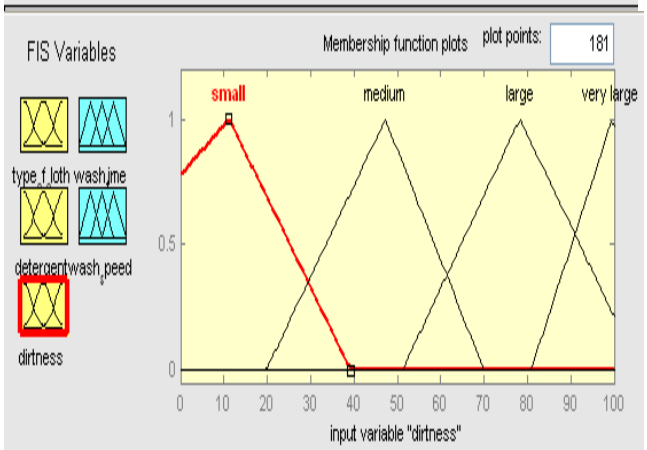
**Rules:**

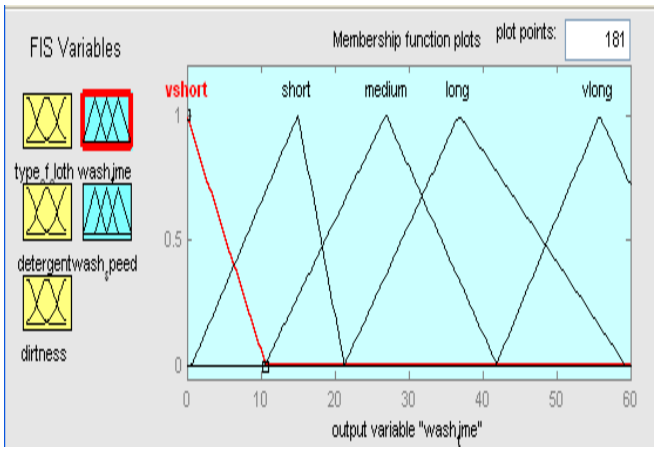
1. If (type_of_cloth is silk) and (detergent is less) and (dirtness is small) then (wash_time is vshort)(wash_speed is low)
2. If (type_of_cloth is silk) and (detergent is less) and (dirtness is medium) then (wash_time is short)(wash_speed is medium)
3. If (type_of_cloth is silk) and (detergent is less) and (dirtness is large) then (wash_time is medium)(wash_speed is medium)
4. If (type_of_cloth is silk) and (detergent is medium) and (dirtness is small) then (wash_time is medium)(wash_speed is high)
5. If (type_of_cloth is silk) and (detergent is medium) and (dirtness is medium) then (wash_time is long)(wash_speed is medium)
6. If (type_of_cloth is silk) and (detergent is medium) and (dirtness is large) then (wash_time is long)(wash_speed is high)
7. If (type_of_cloth is silk) and (detergent is more) and (dirtness is small) then (wash_time is medium)(wash_speed is medium)
8. If (type_of_cloth is silk) and (detergent is more) and (dirtness is medium) then (wash_time is long)(wash_speed is medium)
9. If (type_of_cloth is silk) and (detergent is more) and (dirtness is large) then (wash_time is vlong)(wash_speed is high)
10. If (type_of_cloth is wool) and (detergent is less) and (dirtness is small) then (wash_time is short)(wash_speed is low)
11. If (type_of_cloth is wool) and (detergent is less) and (dirtness is medium) then (wash_time is medium)(wash_speed is medium)
12. If (type_of_cloth is wool) and (detergent is less) and (dirtness is large) then (wash_time is long)(wash_speed is high)

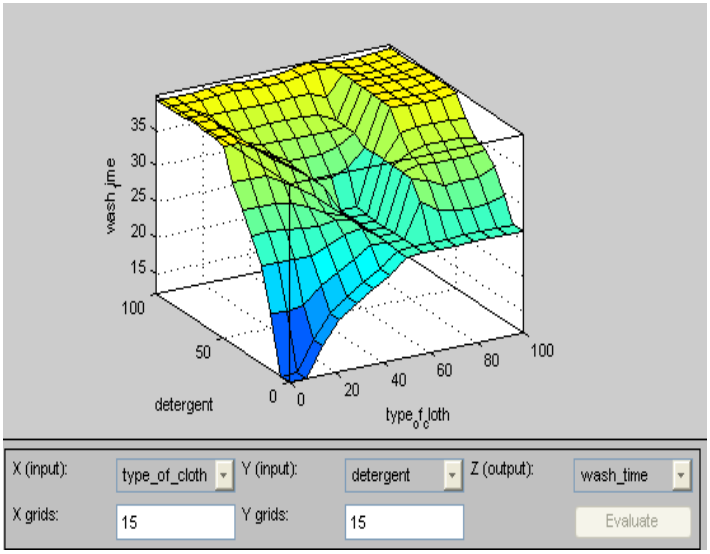
13. If (type_of_cloth is wool) and (detergent is medium) and (dirtness is small) then (wash_time is medium)(wash_speed is medium)
14. If (type_of_cloth is wool) and (detergent is medium) and (dirtness is medium) then (wash_time is medium)(wash_speed is medium)
15. If (type_of_cloth is wool) and (detergent is medium) and (dirtness is large) then (wash_time is long)(wash_speed is high)
16. If (type_of_cloth is wool) and (detergent is more) and (dirtness is small) then (wash_time is long)(wash_speed is high)
17. If (type_of_cloth is wool) and (detergent is more) and (dirtness is medium) then (wash_time is long)(wash_speed is high)
18. If (type_of_cloth is wool) and (detergent is more) and (dirtness is large) then (wash_time is vlong)(wash_speed is very high)
19. If (type_of_cloth is cotton) and (detergent is less) and (dirtness is small) then (wash_time is short)(wash_speed is low)
20. If (type_of_cloth is cotton) and (detergent is less) and (dirtness is medium) then (wash_time is medium)(wash_speed is medium)
21. If (type_of_cloth is cotton) and (detergent is less) and (dirtness is large) then (wash_time is long)(wash_speed is high)
22. If (type_of_cloth is cotton) and (detergent is medium) and (dirtness is small) then (wash_time is medium)(wash_speed is medium)
23. If (type_of_cloth is cotton) and (detergent is medium) and (dirtness is medium) then (wash_time is long)(wash_speed is high)
24. If (type_of_cloth is cotton) and (detergent is medium) and (dirtness is large) then (wash_time is vlong)(wash_speed is very high)
25. If (type_of_cloth is cotton) and (detergent is more) and (dirtness is small) then (wash_time is long)(wash_speed is high)
26. If (type_of_cloth is cotton) and (detergent is more) and (dirtness is medium) then (wash_time is long)(wash_speed is high)
27. If (type_of_cloth is cotton) and (detergent is more) and (dirtness is large) then (wash_time is vlong)(wash_speed is very high)

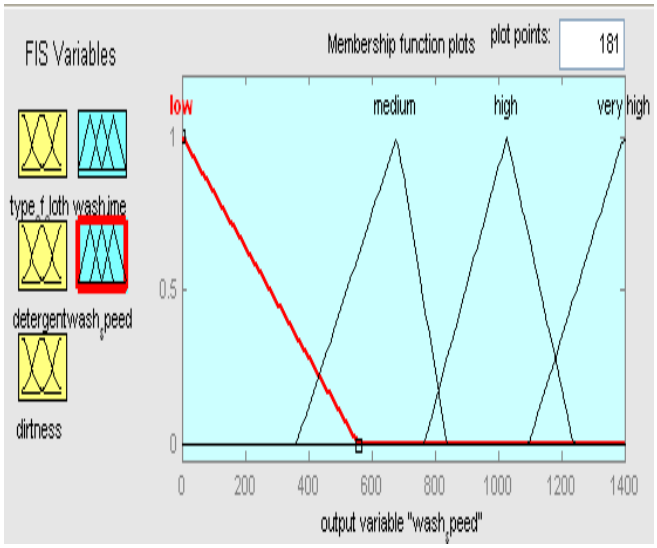


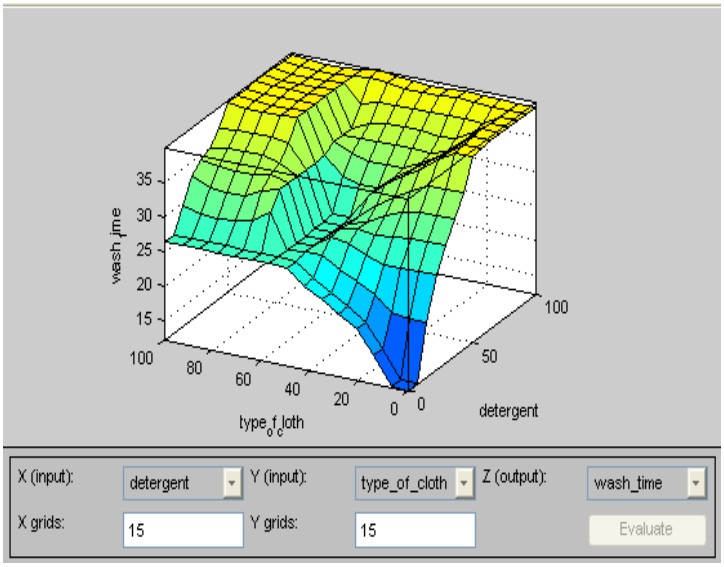






The set of rules are Linguistic inputs and Linguistic Output as shown in table 1

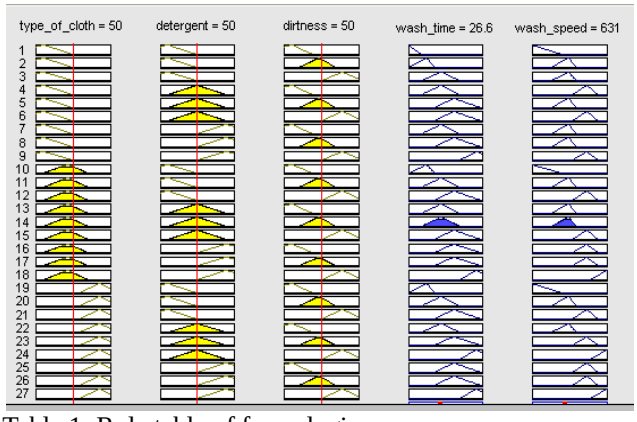
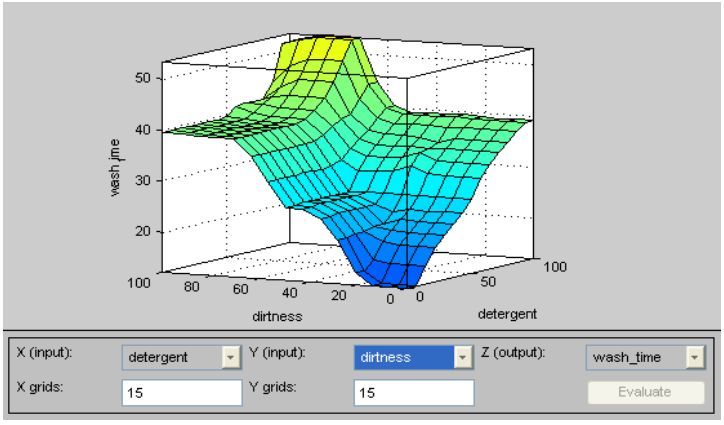
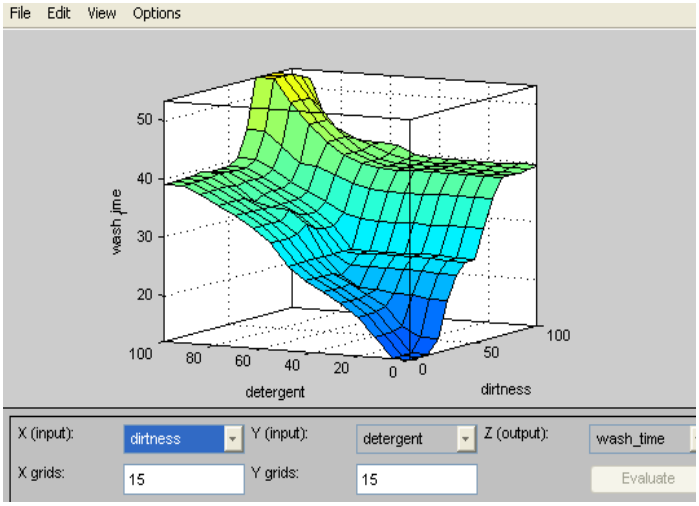
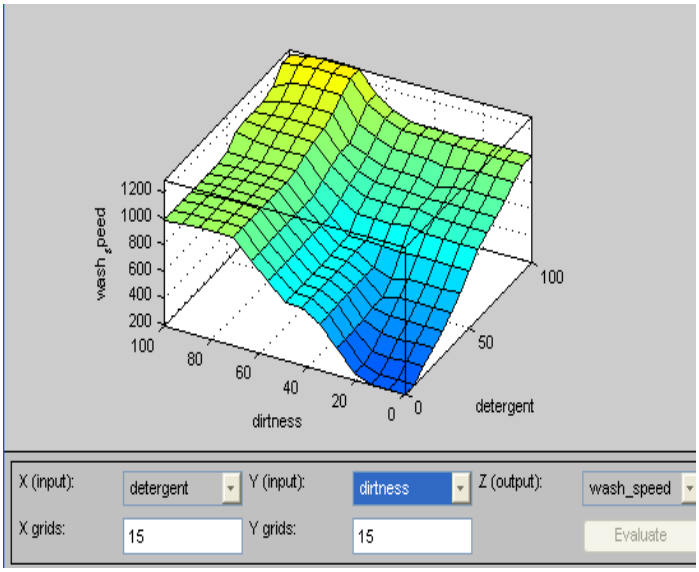
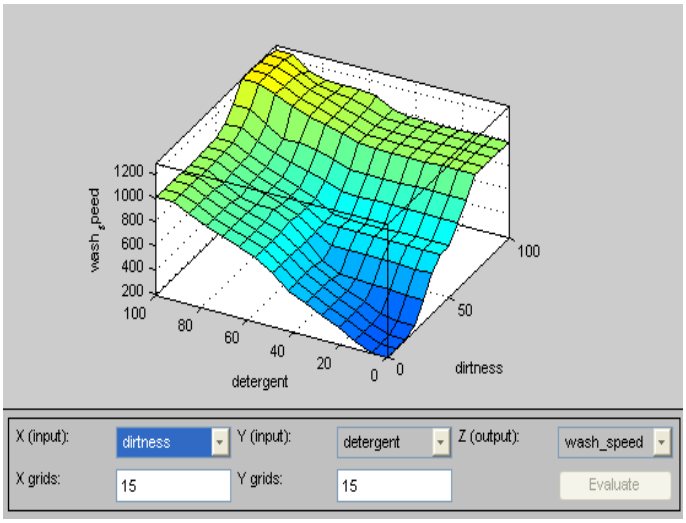



Table 1: Rule table of fuzzy logic



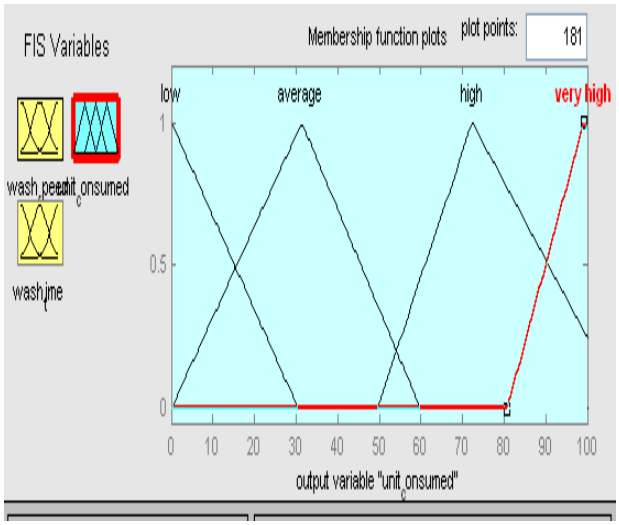



Case II.
The second part of this paper we assumed two inputs and one output
Input parameter:
1. Wash Speed
2. Wash Time

Output parameter.
1. Unit Consumed




**Rules:**

1. If (wash_speed is low) and (wash_time is very_short) then (unit_consumed is low)
2. If (wash_speed is low) and (wash_time is medium) then (unit_consumed is average
3. If (wash_speed is medium) and (wash_time is short) then (unit_consumed is average)
4. If (wash_speed is medium) and (wash_time is medium) then (unit_consumed is high)
5. If (wash_speed is high) and (wash_time is medium) then (unit_consumed is high)
6. If (wash_speed is high) and (wash_time is long) then (unit_consumed is very high)
7. If (wash_speed is very high) and (wash_time is very long) then (unit_consumed is very high)

The set of rules are Linguistic inputs and Linguistic Output as shown in table 2 using centroid method in MATLAB.

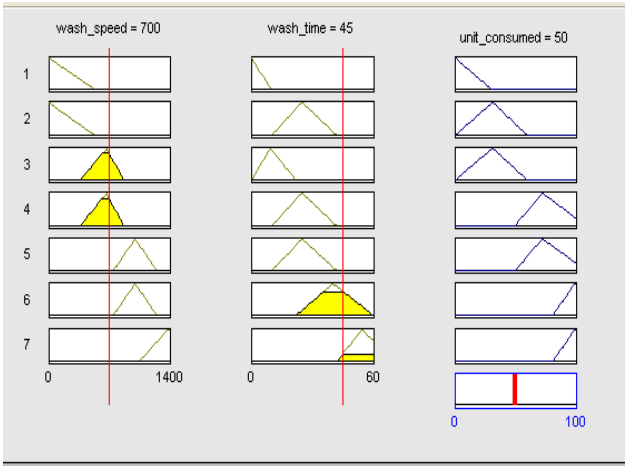
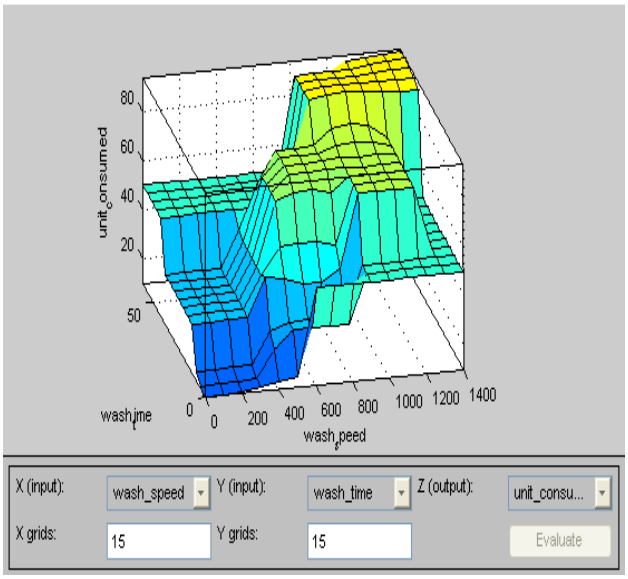



Table 2: Rule table of fuzzy logic




**5. Defuzzification: Centroid Method**

In case I and case II the defuzzification process to interpret membership degrees of fuzzy sets. Here in both cases the Centroid method is used for defuzzification for actual out put of case I in wash time and wash speed. In case II defuzzification we used Centroid method for unit consumed.

**6. Conclusion:**

In this paper fuzzy rule based inference system for the washing machine is analyzed for :

**case I** with the input parameters type of cloth , detergent and dirtiness. The out put parameter using centroid method for wash time is 26.6 m and wash speed is 631 rpm.

**case II** the input parameter for unit consumption is calculated when wash speed is 700 rpm and wash time is 45 m is 50 Units. This is quite convincing and appropriate.

**7. References**

1. Zadeh,,L.A.,.Fuzzy Sets, *Information and Control* , **8, 338-353,(1963)**

2. Mamdani, E . H., Application of Fuzzy Algorithm for Control of Simple Dynamic Plant, *Proc. IEEE*,121(12),1585-1588,(1974)

3. Zhen, A. and Feng, R. G. The Design of neural network fuzzy controller in washing machine, Int. Conference on Computing , Measurement, Control ans Sensor Network *(CMCSN),Proceeding* , Shanxi, China, **136- 139,(2012)**

4. George,J. Kilr and Bo Yuan, Fuzzy Sets and Fuzzy Logic, , *PHI*, India, **(2006)**

5. Lohani, P. and Hasan, S.R., Design of an Improved Controller Micrrochip for Washing Machine, *16th Annual Electronics Conference*, New Zealand, Dunedin: Otago University, 20-26, (2009)

6. Virkhare,N. Jasutkar, R. W., Neuro-Fuzzy Controller Based Washing Machine, International Journal of Engineering Science Invention, **3(1), 48-51, (2014)**

7. Kumar, D. and Haidr,Y. Fuzzy Logic Based Control System for Washing Machine,International Journal of Computer Science and Technology,**4(2), 198-200, (2013)**

8. Mendel,J.M. Fuzzy Logic System for Engineering *A Tutorial, proceedings of the IEEE*, 83(3),(1995)

9. Alhanjouri, M.and Alhaddad,A. Optimize Wash Time of Washing Machine Using Fuzzy Logic, Islamic University of Gaza.

10. Agarwal, M. Fuzzy Logic Controlof Washing Machine, URL:h*ttp://soft computing*.tripod.com/sample_termpater:pdf

# 3D Search Engine

**Prof. David Thomas Vidhate.**
Department of Computer Science
Christ College Pune (India)
davidvidhate@yahoo.com

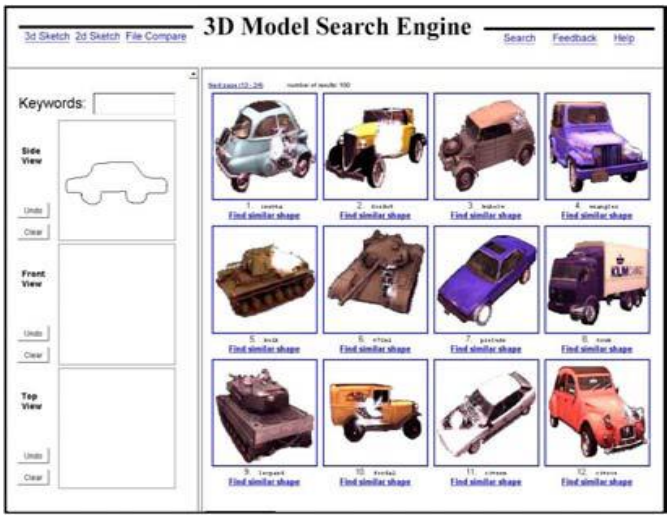
**ABSTRACT:**

**As the number of 3D models available on the Web grows, there is an increasing need for a search engine to help people. Unfortunately, traditional text-based search techniques are not always effective for 3D data. The key challenges are to develop query methods simple enough for novice users and matching algorithms robust enough to work for arbitrary polygonal models. A web-based search engine system that supports queries based on 3D sketches, 2D sketches, 3D models, and/or text keywords and also present a web-based search engine system that supports multimodel queries which include both text query and sketch query. This results in faster retrieval of the result and the percentage efficiency also increases. The net result is a growing interactive index of 3D models available on the Web (i.e., a Google for 3D models).**

**Computer science has made incredible progress in computer aided retrieval and analysis of multimedia data. Although web search is commonplace for text, images, and audio, the information revolution for 3D data is still in its infancy. Recent trends are combining to accelerate the proliferation of 3D models, leading to a time in the future when 3D models will be as ubiquitous as other multimedia data are today: (1) new scanners and interactive tools are making construction of detailed 3D models practical and cost effective, (2) inexpensive graphics hardware is becoming faster, causing an increasing demand for 3D models from a wide range of people, and (3) the web is facilitating distribution of 3D models. We need 3d Search Engine Now a days, developments are changing the way we think about 3D data. The primary challenge in computer graphics has been how to construct interesting 3D models and to how do we find them? for text, images, audio, and other media. An important question then is how people will search for 3D models. Of course, the simplest approach is to search for keywords in filenames, captions, or context. However, this approach can fail: (1) when objects are not annotated (e.g., "B19745.wrl"), (2) when objects are annotated with inspecific or derivative keywords (e.g.,"yellow.wrl" or "sarah.wrl"), (3) when all related keywords are so common that the query result contains a flood of irrelevant matches (e.g., searching for "faces" – i.e., human not polygonal), (4) when relevant keywords are unknown to the user (e.g., objects with misspelled or foreign labels), or (5) when keywords of interest were not known at the time the object was annotated. In these cases and others, a 3D search engine is needed.**

**b) How to Search For 3d Models We hypothesize that shape-based queries will be helpful for finding 3D objects. For instance, shape can combine with function to define classes of objects (e.g., round coffee tables). Shape can also be used to discriminate between similar objects (e.g., desk chairs versus lounge chairs). There are even instances where a class is defined entirely by its shape (e.g., things that roll). In these instances, "a picture is worth a thousand words." Our work investigates methods for automatic shape-based retrieval of 3D models. The challenges are two-fold. First, we must develop computational representations of 3D shape (shape descriptors) for which indices can be built and similarity queries can be answered efficiently. In this paper, we investigate combinations of 3D sketching, 2D sketching, text, and interactive refinement based on shape similarity. We have integrated these methods into a search engine that provides a publicly available index of 3D models on the Web (Figure 1.1).**

**Figure 1.1**




**It allows a user to specify a query using any combination of keywords and sketches(left). Then, for each query, it returns a ranked set of thumbnail images representing the 16 best matching 3D models (right). The user may retrieve any of the 3D models by clicking on its thumbnail, and/or he may refine the search by editing the original input or by clicking on the "Find Similar Shape" link below any thumbnail.**

## 1. SYSTEM OVERVIEW FOR 3D MODELS

Execution proceeds in four steps: crawling, indexing, querying and matching. The first two steps are performed offline, while querying and matching are performed for each user query.

### CRAWLER

The crawler is a software application that gathers (or fetches) web pages from a network, typically a bounded institutional or corporate network, but potentially the entire Internet, in a controlled and reasonably deterministic manner. But in this research a crawler which can crawl 3D image files and creates a separate collection(database). This collection is further indexed and indexed contains are stored in a separate cluster and has a separate index node. The user can enter a filename, or browse for one, and upload a 3D model file in one of several 3Dfile formats The uploaded file is first converted to only geometry in the PLY format. From this file its shape signature is computed, which is then compared to the database signatures.

The crawler stores crawled content with its meta data. **Figure 2.1**

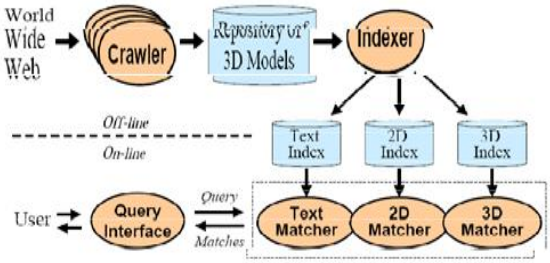



Figure 2.1: System Organization

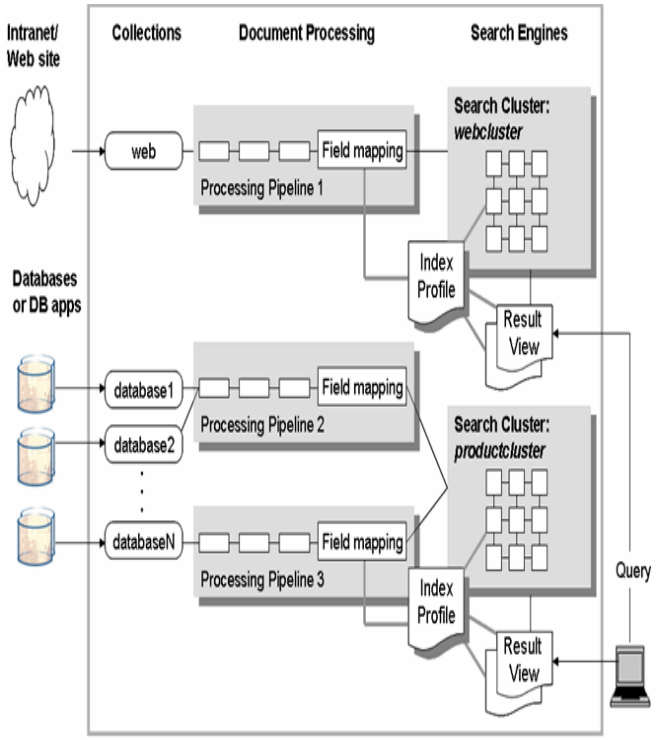
**Indexing:** Indices are computed to retrieve 3D models efficiently based on text and shape query. A special processing pipeline indexes 3D image which understand graphics system to provide collection of graphic routines which can be used to create, store, manipulate an interact with images. The indexer manipulates all the information of the images stored in the collection database. The Indexing Dispatchers job is to decide which indexing column a document is routed to. This is controlled by a routing policy. The default policy is to hash on md5 of the document identifier. The Indexing Dispatchers assigned routing policy can also be based on a routing attribute. The routing attribute named column is set per document and routes the document to a given indexing column. It can be set by a document processing stage using the content API. This routing attribute makes it possible to implement any kind of routing.

```
<process name="indexingdispatcher" description="Indexing
Dispatcher" multi="no">
<start>
<executable>indexingdispatcher</executable>
<parameters>
--cluster=webcluster --routing=column -
ORBendPointNoListen
giop:tcp:node02.myhost.com:$PORT -
ORBendPointNoPublish


giop:tcp::$PORT
</parameters>
<port base="3060" increment="1" count="1"/>
<delay>15</delay>
</start>
<stop>
<timeout>30</timeout>
<timeout_action>kill</timeout_action>
</stop>
<outfile>indexingdispatcher.scrap</outfile>
</process>
```

Indicate the document elements that should be the basis for generating the document vector. This is performed in the field specification element of the index profile, using the vectorize attribute. Multiple text fields so that selected image are generated using similarity vectors are based on text analysis of all of these fields in the index profile. The document elements are configured for automatic tokenization so that the processing pipeline for your collection(s) supports all pipelines used to index the images.

### 2. QUERYING (Query Engine):

It acts as a query engine for the 2D, 3D search, users are allowed to search interactively, now this query engine supports, query methods based on text keyword, 2D sketch and 3D sketch. The first approach is to specify shape query, this is done with simple 3D sketching tool such has TEDDY. The user creates simple 3D model with TEDDY and this images are passed to the query engine for retrieval of similar models. The pipeline understands the image as it is passed to the query engine the images satisfies similar criteria are searched in the indexer and shown on the query output area. **Figure 3.1**




The architecture of the crawler **Figure 3.1**

```
IQuery query = new Query(queryString);
query.setParameter(new
SearchParameter(BaseParameter.OFFSET, 12));
query.setParameter(new
SearchParameter(BaseParameter.HITS, 12));
IQueryResult result = searchView.search(query);
int min =
    query.getIntParameter(BaseParameter.OFFSET) + 1;
int max = result.getDocCount();
int mid = (max < (min +
query.getIntParameter(BaseParameter.HITS)))
? max: (min +
query.getIntParameter(BaseParameter.HITS));
for (int i = min; i < mid; i++) {
IDocumentSummary summary = result.getDocument(i);
System.out.println(summary.toString());}
```

Limiting the scope to a set of collections is achieved by using the filter term with a query. Filter terms can be applied

using the filter() operator in the Query Language. A reserved field name, meta collection, identifies the collection for 2D 3D images. Images are grouped into collections based on data source and document processing.

## 3. CONCLUSION:

In summary, this paper investigates issues in building a search engine for 3D models. The main research contributions are.

1. New query interfaces that integrate text, 2D sketches, 3D sketches, and 3D models.

2. A new shape descriptor based on spherical harmonics that is both discriminating and robust,

3. Results of experiments suggesting shape is useful in conjunction with text during search for 3D models. Finally, we provide a large repository of 3D models ... and a way to find the interesting ones

## 4. FUTURE WORK:

Better 2D image matching methods: our 2D sketching interface would be more effective

with better image matching algorithms. Sometimes users create query sketches with interior texture and/or details (e.g., eyes and mouth of a human face), and our search engine matches them with projected images containing only boundary outlines (e.g., just the outline of the face). For matching purposes, the interior details in sketches are "interpreted" as boundaries of holes in projected images, and unexpected results are sometimes returned to the user. Of course, this problem could be rectified somewhat by providing users with instructions or examples about how to draw their sketches. However, determining how best to present the instructions and providing image matching methods that minimize user guidance are topics for future study.

New query interfaces: it will be interesting to consider other methods for specifying shape-based queries. For instance, the following constraint-based description might be used to retrieve 3D models of a chair: "give me objects consisting of a box-shaped seat with four equal length and nearly cylindrical legs attached to the bottom side of each corner and a box-shaped back above the seat with width matching that of the seat, New matching and indexing algorithms: follow-up work should consider other types of shape matching problems. For instance, we currently compare whole objects, but it would be interesting to match partial objects as well. They could be used to find a car within a city scene or to find a Mercedes by looking for its hood ornament etc.

New Cluster and separate indexer node for the collection of the 2D, 3D image to be configured so the search operation is optimized.

## 5. References

1. J. Cho, H. Garcia-Molina, and L. Page. E_cient Crawling through URL ordering. Computer Networks and ISDN Systems. Special Issue on the Seventh International World-Wide Web Conference, Brisbane, Australia, 30(1-7):161{172, April 1998.

2. K. Hirata, Y. Hara, N. Shibata, and F. Hirabayashi. Media-based Navigation for Hypermedia Systems. In Proceedings of ACM Hypertext '93 Conference, pages 159{173, Seattle, WA, November 1993.

3. S. Lawrence and C. Giles. Searching the World-Wide Web. Science, 280(5360):98,1998.

4. R. Miller and K. Bharat. SPHINX: a framework for creating personal, site-speci_c Web crawlers . Computer Networks and ISDN Systems. Special Issue on the Seventh International World-Wide Web Conference, Brisbane, Australia, 30(1-7):119{130,April 1998.

5. Thomas Funkhouser, Patrick min, Michael Kazhdan, Joyce Chen, Alex Halderman, and David Dobkin Princeton University, "A Search Engine for 3D Models" ACM Transactions on Graphics, Vol. V, No.N, 10 202002

6. Zeleznik, R. C., Herndon, K. P., and Hughes, J. F. 1996. Sketch, "An Interface for Sketching 3DScenes", In Proceedings of SIGGRAPH 96. Computer Graphics Proceedings, Annual Conference Series. 163–170.

7. Igarashi, T.,Matsuoka, S., and Tanaka, H. 1999. Teddy: "A Sketching Interface for 3D Freeform Design". In Proceedings of SIGGRAPH 1999.Computer Graphics Proceedings, Annual Conference Series., Los Angeles,CA, 409–416

8. Microsoft Fast Esp guide.

9. (2011)"Search Engine Ranking Factors V2". SEOmoz.org[Online].http://www.seomoz.org/article/searchranking-factors.

IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

ISSN (Online): 1694-0814 www.IJCSI.org 214

# A Review On Heuristics: Search Techniques In AI

**Minal Rajendra Pawar**
Student, VPASC College, Baramati.
minalpawar886@gmail.com

## ABSTRACT:

**A process that may solve a given problem, but offers no guarantees of doing so, is called a heuristic for that problem. The aim of heuristic is to study the methods and rules of discovery and invention. Heuristic, as an adjective, means "serving to discover."**

**Heuristics are efficient cognitive processes, conscious or unconscious, that ignore part of the information. Because using heuristics saves effort. Heuristic discusses human behavior in the face of problems; this has been in fashion, presumably, since the beginning of human society, and the quintessence of such ancient discussion seems to be preserved in the WISDOM OF PROVERBS.**

***Key words:*** **heuristic, problem solving, artificial intelligence, cognitive science, Heuristic search**

**Introduction:**

**What is a Heuristics?**

The term heuristic is of Greek origin and means, "serving to find out or discover."

A search heuristic is information to guide the traversal into the direction of the set of the goals.

In a weighted state space problem graph G = (V,E, s, T,w) a heuristic h is a node evaluation function mapping V to IR+0 .

We require

• fort in T: h(t) = 0

• all other nodes v 2 V we have h(v) 6= 0

so that the goal check for u simplifies to compare h(u) with 0.

The concept of heuristic began to appear in the early 1950's A1 literature and was well known by the early 1960's. From the mid 1950's to the present, the notion of a heuristic has played a crucial role in the A1 researcher's descriptions of their work. Things that would be called a heuristic by one researcher would not be so called by others. This is because many heuristics embody a variety of different features, and the various researchers have emphasized different ones of these features as being essential to being a heuristic. That the concept of a heuristic has been, and continues to be, central in A1 is too well known to require documentation.

Heuristic is a science of problem solving behavior that focuses on plausible, provisional, useful, but fallible, mental operations for discovering solutions.

A heuristic (heuristic rule, heuristic method) is a rule of thumb, strategy, trick, simplification, or any other kind of device which drastically limits search for solutions in large problem spaces.

**Search**

Another example of heuristic making an algorithm faster occurs in certain search problems. A heuristic method can accomplish its task by using search trees.

**Use of heuristics:**

Two answers have been proposed to the question of why heuristics are useful: the accuracy-effort trade-off, and the ecological rationality of heuristics.

1. **Accuracy-Effort Trade-Off**

The classical explanation is that people save effort with heuristics, but at the cost of accuracy; heuristics trade-off some loss in accuracy for faster and more frugal cognition.

There are two interpretations of this tradeoff:

(a) Rational trade-offs.

(b) Cognitive

2. **Ecological Rationality**

"A heuristic is ecologically rational to the degree that it is adapted to the structure of the environment"

**Heuristic function** h(n):-

A heuristic function, also called simply a **heuristic**, is a function that ranks alternatives in search algorithms at each branching step based on available information to decide which branch to follow.

Heuristic function h(n)

- Admissible/Underestimate:
- Consistent/monotonic

1. **Admissible/Underestimate**
   h(n) never overestimate the actual cost from n to goal. A heuristic h is admissible if it is a lower bound estimate that underestimates the shortest remaining distance to the goal.

2. **Consistent/monotonic(desirable)**
   Many heuristics have monotone estimator functions:

   h(m)-h(n)<=w(n,m)

   Where, m is parent of n.

   This ensures

   f(n)>=f(m).

   A heuristic h is consistent, if h(u) _ h(v)+w(u, v) for all edges

   e = (u, v)

**Properties of Heuristic functions:**

1. The 2 most important properties:

   –relatively cheap to compute

   –relatively accurate estimator of the cost to reach a goal. Usually a "good" heuristic is if ½ **opt(n)<h(n)<opt(n)**

2. Examples:

   –Navigating in a network of roads from one location to another.

   –Sliding-tile puzzles.

**Conclusion:**

We began this review with the observation that the three major tools for modeling decision making—logic, statistics, and heuristics—have not been treated equally, with each suited to a particular kind of problem.

1. Heuristics can be more accurate than more complex strategies even though they process less information (less-is-more effects).

2. A heuristic is not good or bad, rational or irrational, its accuracy depends on the structure of the environment.

3. Usually, the same heuristic can be used both consciously and unconsciously, for inferences and preferences, and underlies social as well as nonsocial intelligence.

**References:**

1. Astebro T, Elhedhli S [2006]. The effectiveness of simple decision heuristics: forecasting commercial success for early-stage ventures.
2. Allen Newell and Herbert A. Simon [1976]. "Computer Science as Empirical Inquiry: Symbols and Search". Comm.

Books:

1. Artificial Intelligence Third Edition Mc Graw Hill Education Elaine Rich, Kevin Knight, Shivshankar B. Nair
2. Artificial Intelligence Vision Publication Third Edition

# Automated Boolean Matrix Data Representation Approach for Efficient Mining

**Renju Samuel and Linu Paulose**

.

**ABSTRACT:**

**In the modern world the process of identifying the required items from a large dataset leads to a large amount of time. The concept of Data Mining is normally used to isolate the required item from a large database. In this paper a new form of isolating the data from a database is used. The traditional Apriori Algorithm is used to generate the Association rules by obtaining frequent itemsets of 1 combination, 2 combinations and so on based on the minimum support count. This study aims to propose a modified Apriori algorithm in which generation of Association rules is by using a Boolean Matrix representational scheme. At first the data representation takes place via the generation of an AVL tree which is used to balance the given data. Then the Boolean Matrix gets generated from the AVL tree and then association rules get generated from the matrix. Its shows clearly that the computation time has improved for the modified algorithm when compared to the traditional Apriori Algorithm**

**Index terms:** Association Rule, AVL Tree, Boolean Matrix,Data Mining, Frequent Itemset.

## 1 - INTRODUCTION

The process analysing data from different perspectives and categorising it into useful information. It allows the users to analyse the database from differentangles. Technically data mining can be viewed as the process of finding pattern among sets of fields among relational databases.Data mining is used in business to discover pattern, trends,relationships, anomalies in order to make better decisions.Apriori algorithm is designed to operate on transactional databases. It identifies the frequent individual items in the database and extends them to larger item sets as long as those appear sufficiently often in the database.

A "bottom up" approach where frequent subsets are extended as one item at a time and groups are tested against the data is been used by Apriori algorithm. Number of items in the item sets are counted using breadth-first search and hash tree mechanisms in Apriori algorithm. Association rule which highlights the general trends in data bases can be determined by frequent sets whichresults from Apriori algorithm. The popular method for discovering interesting relations between variables in large databases is called association rule. It identifies strong rules discovered in databases using different measures. Association can be used to discover regularities between products in large scale transaction of data's in supermarkets. For example, the rule

{Milk, Coffee powder} ⇨ {Sugar}

From the sales data of a supermarket would indicate that if a customer buys milk and coffee powder together is likely to buy sugar also. Such information's can be used as the basis for decisions about business and marketing activities such as pricing and product placements. Association rule has many applications including Bioinformatics,Continuous production, Web usage mining and intrusion detection. It doesn't usually consider the order of the item either within a transaction or across the transaction.Apriori algorithm uses two rules they are

1. Find all the frequent sets, each of the frequently occurring at least as frequently as minimum support rule which is predetermined.
2. Generate association rules from the frequent item sets.

Mining association rule is converted to find the frequent item sets. But traditional Apriori algorithm has two drawbacks they are

1) Many input –output operations are needed to scan databases frequently. It needs to scan the transaction database as the same number as the elements of the frequent item sets.
2) When analysing the databases we get large number of tables, which also requires a lot of time and memory (storage space).

To overcome these drawbacks we propose modified Apriori algorithm which uses data representation schemes. The algorithm is improved through AVL-tree based scheme which provides more reduced time and condensed rule than the traditional Apriori method. Using this method it is able to get last frequent itemsets and generates association rule depending on the data representation and Boolean matrix.In Apriori algorithm we need to access the database frequently which consumes both time and memory space. In modified Apriori algorithm which is based on data representation and Boolean matrix reduces this overhead.

## 2 - RELATED WORK

Agarwal R and R Srikant suggested two new algorithms named as Apriori and Apriori Tid are used, for discovering all association rules between items in a large database.

Finally they compared these algorithms to previously known algorithms such as AIS and SETM and the new implemented algorithms are much more effective than the existing algorithms [1].

Mannila, H.H Toivenn and A. Verkamoproved that an algorithm that uses all existing information between database, and the extra step they have done is that it passes to avoid checking the coverage of duplicate itemsets . It provides a better execution time when compared to the previous results and it is simple to implement. Concluded in a way that sampling is an efficient technique for finding rules of this type and that algorithms working in the main memory can be used to obtain very good approximations [2**].**

Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan had used a way to discover patterns present in very large databases other than to verify that a certain pattern exists in that database. The result obtained produced high performance and also linear scaling in real life databases. Also they had provided a completeness property that promises all the similar patterns present in the database of certain type will be identified [3].

Hassan M. Najadat, Mohammed Al-Maolegi, Bassam Arkokhave used several groups of transaction, and various minimum support count values in the database. And they have applied the same values to both the normal and improved Apriori algorithm. Theconclusion obtained is that the new improved Apriori algorithm had improved the time consumption by 67.38%. Also another result is that it is more efficient when compared to the normal apriori algorithm [4]. t Bay and Bac Le have presented a new form of Apriori algorithm for obtaining the frequent itemsets and to generate the association rules. They scan the database only once in order to generate the transaction ids to obtain the frequent itemsets. A tree structure is used to obtain these results. They omit the duplicate itemsets in order to obtain the results of generating the association rules at a faster phase [5].

An Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z , they had used Hash function instead of normal joining method. Transactionreduction is done by generating hash function using DHP algorithm. Collision in hash table reduces the effectiveness of hash table. Hash functions are complex to be coded after 2-itemset combination.(DHP- Direct hashing and pruning) [6].

Dhilhan perera, Judy Kay, Irena Koprinska, Kalina Yacef , and Osmar Zaiane have proposed high level view of information from the database, that is together they pair the patterns of the same group. Their goal was to enable the particular groups and their identities the various operations of the group and the feedback that is provided during the operations. The tool used in order to obtain the result is TRAC. They extract patterns of the strong groups eliminating the weaker groups to get more successful results [7].

Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani had proposed a way to find all implication and redundancy rule based on confidence pruning without support pruning. Dynamic pruning done to the large data set by counting number of rows and avoid the rows containing missing values. Only confidence pruning is considered. They focused only on the reduction in the transaction size and noton efficient rule generation [8].

## 3- PROPOSED WORK

Initially the data present in the given database gets sorted out in ascending order by using the AVL tree .The next step is the data discretization of the sorted data. Here the sorted values from the AVL tree gets replaced according to the Index values provided to it. After this particular step the Hash Map table is being generated using the values obtained in the previous step. The Hash Map table consist of transaction id and the List of items present in the list. From these values the Boolean Matrix is obtained. All the places at which the value is present are termed as 1 and the place at which value is not present is termed as 0. This result is formed by using the Hash Map table. And the last step is generation of Association Rules from the Boolean Matrix.

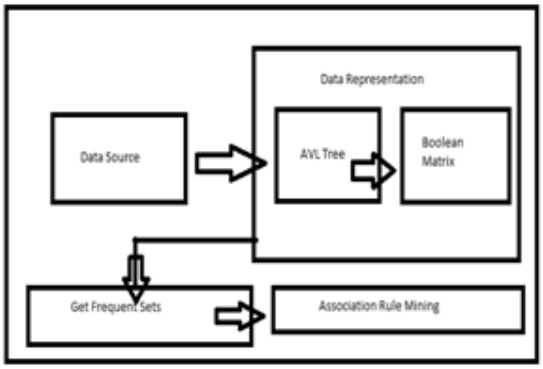



Fig. 1. Architecture Diagram

Data representation transforms source data into an AVL tree which is a balanced structure. Also it reduces the computational time for the process. Also it obtains various multi data types of the represented data from the database such as a hash table, tree and Boolean matrix.

Shown below (Fig. 2) is an example for the generation of Association rules using the modified approach.

Table 1 : Given Dataset

| Transaction ID | List Of Items |
|---|---|
| T101 | 8,14 |
| T102 | 10,12,14 |
| T103 | 8,12,14 |
| T104 | 10,14 |

| | |
|---|---|
| T105 | 8,24 |
| T106 | 12,14 |
| T107 | 8.12 |
| T108 | 12,14 |

Here the given value gets balanced by the AVL Tree from the database (Fig. 3).

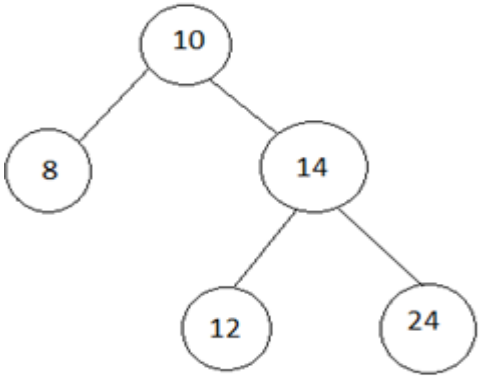



Fig. 2. Generation of AVL Tree

After the balanced values are obtained from the AVL Tree, the data discretization (Fig. 4) is done by replacing the elements present at the particular level in the database.

Table 2: Data Discretization

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 8 | 10 | 12 | 14 | 24 |

Next step is the generation of Boolean Matrix (Fig. 5), it is done by substituting the values which are obtained from the previous step, values are substituted as 1 in the place of element is present and 0 at the place where presence of element is null.

Table 3: Boolean Matrix

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| T101 | 1 | 0 | 0 | 1 | 0 |
| T102 | 0 | 1 | 1 | 0 | 1 |
| T103 | 1 | 0 | 1 | 1 | 0 |
| T104 | 0 | 1 | 0 | 1 | 0 |
| T105 | 1 | 0 | 0 | 0 | 1 |
| T106 | 0 | 0 | 1 | 0 | 1 |
| T107 | 1 | 0 | 1 | 0 | 0 |
| T108 | 0 | 0 | 1 | 1 | 0 |

Hence the Boolean Matrix is generated. Now the next step is to find the frequent itemsets from the obtained Boolean Matrix

$$\text{Computed Weight} = \frac{\text{Count 1's in Column}}{\text{Number Of Rows}} >= S$$

Here S is the minimum support count given by the user

Table 4: Generation of Frequent Itemsets

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| T101 | 1 | 0 | 0 | 1 | 0 |
| T102 | 0 | 1 | 1 | 0 | 1 |
| T103 | 1 | 0 | 1 | 1 | 0 |
| T104 | 0 | 1 | 0 | 1 | 0 |
| T105 | 1 | 0 | 0 | 0 | 1 |
| T106 | 0 | 0 | 1 | 0 | 1 |
| T107 | 1 | 0 | 1 | 0 | 0 |
| T108 | 0 | 0 | 1 | 1 | 0 |
| COUNT | 4 | 2 | 5 | 4 | 3 |
| Count / Total Number of Rows | 4/8 = 0.5 | 2/8 =.25 | 5/8 =.625 | 4/8 =.5 | 3/8 =.375 |

Here all 1's are counted in each column and it is added and divided with the total number of rows. Let us take the minimum support count (S) as 0.5,so all values which are equal to and greater than equal to 0.5 is taken into consideration. Rest all values are rejected (Fig. 6).

Hence from the following table, the values 0.5,0.625, 0.5 are selected. Its following values are taken inorder to generate the Association Rule from the table

In this example the last frequent itemset are {1, 3, and 4}

The following Rules can be generated

**Rule 1**: 1^3 →4,**Rule 2**: 1^4 →3

**Rule 3**: 3^4 →1,**Rule 4**: 1 →3^4

**Rule 5**: 3→ 1^4,**Rule 6**: 4 → 1^3

Thus the Association Rules are generated from the given dataset using the Modified Apriori Algorithm

## 4 – CONCLUSION

The following study had provided suitable information such that with the data representation scheme that is followed here reduces the computational time. It also improves the various processes such as the generation of n-frequent itemsets and also the generation of the Association Rules when compared to the normal Apriori Algorithm. The major advantage of using this type is that the entire database needs to be scanned only once inorder to generate the Association Rules. Thus it saves a lot of computational time than the normal Apriori, which has to invoke the database each and every step inorder to generate its Association Rules.

## REFERENCES


[1]. **Agarwal R and R Srikant**, **1994, "Fast Algorithms for Mining Association Rules in large databases"** : Jorge B Bocca , Matthias Jarke and Carlo Zaniolo, editors , Proceedings of the 20th International Conference on Very Large Data Bases , VLDB , Santiago , Chile

[2]. **Mannila , H.H Toivenn and A. Verkamo , 1994** : **Efficient algorithm for discovering association rule**s . AAAI workshop on Knowledge Discovery Database Seattle 181-92

[3]. **The Quest Data Mining System : Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan Srikant** , IBM Almaden Research Center San Jose, California 95120, U.S.A 1996.

[4]. **An Improved Apriori Algorithm for Association Rules , Hassan M. Najadat, Mohammed Al-Maolegi , Bassam Arkok**, International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08

[5]. **Fast Algorithm for Mining Generalized Association Rules , Bay Vo and Bac Le** , International Journal of Database Theory and Application Vol. 2, No. 3, September 2009

[6]. **Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z,** SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data

[7]. **Clustering and Sequential Pattern Mining of online Collaborative Learning Data , Dhilhan perera , Judy Kay , Irena Koprinska ,** Member , IEEE Computer, Society , Kalina Yacef , and Osmar Zaiane , Member , IEEE IEEE Transactions On Knowledge And Data Engineering , Manuscript ID 2011

[8]. **Dynamic Miss Counting Algorithm Finding Implication and Similarity Rules with Confidence Pruning,Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani ,** Data Engineering, 2000. Proceedings. 16th International Conference on29 Feb 2000-03 Mar 2000 ,501 – 511

[9]. **A study on the Performance of CT-APRIORI and CT-PRO Algorithms Using Compressed Structes for Pattern Mining ,** Journal of Global Research in Computer Science ,Volume 1 , Number 2 September 2010

[10]. **The Quest Data Mining System : Rakesh Agarwal , Manish Mehta , John Shafer and Ramakrishnan Srikanth ,** IBM Almaden Research Centre , San Jose, California 95120, U.S.A 1996

[11]. **Saravanan Suba and Chistopher.t. Article: A Study on Milestones of Association Rule Mining Algorithms in Large Databases .** International Journal of Computer Applications 47(3):12-19, June 2012. Published by Foundation of Computer Science, New York, USA

[12]. **Mining Sequential Patterns: Generalizations and Performance Improvements Ramakrishnan Srikant and Rakesh Agrawal F Srikant**, IBM Almaden Research Center 650 Harry Road, San Jose,CA95120

[13]. **Information Retrieval through Semantic Web, Gagandeep Singh, Vishal Jain ,** International Journal Of Database , Vol 8, No 5june 2012

[14]. **MiningAssociation Rules Based on Cloud Model and Application in Credit Card Marketing ,Yan-li Zhu, Yu-Fen Wang , Shun-Ping Wang , Xiao-juan Guo ,Shenzhen,** China , pp.165-168, 2010 Asia-Pacific Conference on Wearable Computing Systems, 2010



Renju Samuel is currently working as Assistant Professor , Computer Science Department in KMP College of Engineering , Kerala , India, PH- +91 9645105144. E-mail: renju.samuel@gmail.com

Linu Paulose is working as Head Of The Department in Computer Science department in KMP College Of Engineering , Kerala, India PH- +91 9947708656. E-mail: linupaulose2010@gmail.com

# Solution to Design DFA

**Prof.Nilima Shingate** [1]
Department of Computer Science
Head of Department
Christ College, Pune
Nilima.gaikwad86@gmail.com
9766794701

**Mr.Pradip Shingate** [2]
Senior Software Developer
WHIZ Technology, Pune
pradip.shingate@gmail.com
9766934468

**Abstract**

**The minimum consistent DFA problem is that of finding a DFA with as few states as possible that is consistent with a given sample (a finite collection of words, each labeled as to whether the DFA found should accept or reject). Today, it is very difficult to understand the de- signing concepts of deterministic machine. One question is arise? How it is possible to understand the concepts of deterministic machine in a very easy manner. In this paper, we have design a DFA and develop an method with suit- able examples that how DFA machine works in a simply manner. The objective of this paper to under- stand the concepts of deterministic machine in easy manner.**

Keyword - DFA, Transition Graph, Transition Table (TT)

## INTRODUCTION

Finite automata (FA) are abstract algorithms for the recognition of sequences. They are closely related to regular expressions and CFGs. Automata is also called state-transition machines. The central idea is that an automaton, at any one time, is in a unique state and can transition to some other state by reading input. The transitions are defined by a relation from state-input pairs to states. Transitions on the null string, i.e. non-reading transitions, are allowed.

The sequence of input values read is the string that is recognized. Processing begins in start state. At each step the automaton examines the text, and based on its state-transition table, goes to another state. Each time an automaton transitions on an input symbol, that symbol is discarded so that the next symbol may be processed. Whenever the automaton is in a designated final state, the input read so far is said to have been accepted. If, on the other hand, input appears for which there is no defined transition, the automaton is said to reject the input. Processing continues until the input is rejected or there is no more input. Finite automata are either deterministic (DFA) or nondeterministic (NFA).

## I. MYTHOLOGY

Automata theory is a branch of theoretical computer science, DFA is known as Deterministic Finite Automata. A finite state machine accepts or rejects finite strings of symbols and gives a unique computation for each input string[15]. McCulloch and Pitts were among the first researchers to introduce a concept similar to finite automaton in 1943.[5]

A DFA is defined as an abstract mathematical concept, but due to the deterministic nature of a DFA, it is implementa- ble in hardware and software for solving various specific problems[15]. For example, a DFA can model software that decides whether or not online user-input such as email addresses are valid.

Finite Automata (M) is defined as a set of five tuples $(Q, \Sigma, \delta, q_0, F)$

Where

Q= a finite, non-empty set of states
Σ= a finite, non-empty set of inputs
δ is the state-transition function: $\delta: Q \times \Sigma \rightarrow Q$

$q_0$ is the initial state

F is the set of final states, a subset of Q.

δ can be represents using either of three approach given below

•*Transition Graph.*

• *Transition Function.*

• *Transition Table.*

We had used the transition table as the approach to represent δ.

## III. ALGORITHM

By applying this Algorithm we can design Deterministic Finite Automata that accept strings over input symbol a, b having atmost x number of a & y number of b
Algorithm to draw TG
Deterministic Finite Automata M= (Q, Σ, δ, $q_0$, F)
Where
Danish Ather, et al., JNIC, Vol. 1, No. 2, pp. 30-33, 2013 31
Q= {q11,q12,q13,….q21,q22,……qij}
Σ= {a,b}
δ:QXΣ→Q{Represented by Transition Graph }
Q0 = qij when i=j and i=j=1. i.e. q11
F={q11,q12,q13,….q21,q22,……qij}
Let Q be the set of states in Deterministic Finite Automata such that Q={q11,q12,q13,….q21,q22,……qij}
Where i = 1 to x+1
j= 1 to y+1
Input Symbol Σ={a,b}
q11 is the initial state.
Design a directed transition graph having (x+1)*(y+1) states and mark all states as final states.
Label each node as q11,q12,q13,….q21,q22,……qij
Where i = 1 to x+1; j= 1 to y+1; x= na & y=nb
FOR i= 1 to x
do
FOR j= 1 to y
do
if i=j=1 then qij ε Q0 (Initial State)
else there exist a edge E such that δ(qij,a)→qij+1
done inner loop
done outer loop
FOR i= 1 to x
do
FOR j= 1 to y
do
if i=j=1 then qij ε Q0 (Initial State)
else there exist a edge E such that δ(qji,b)→qj+1,i
if i=x+1 and j=y+1 then there exist δ(qij, a)→qij and δ(qij, a)→qij.
done inner loop
done outer loop
q11 being the initial state
DFA "M" will strings over input symbol a, b having atmost x number of a & y number of b if all the input is consumed and halting state is the final state.

## IV.IMPLEMENTATION

**Example 1:**

**Design a DFA,over a language {a,b,c} which start with 'ab' and end with 'bc'.**

Let the resultant DFA is M = (Q, Σ, δ, $q_0$, F)
Where

Q = {$q_0$, $q_1$, $q_2$, $q_3$, $q_4$}
Σ = {a, b, c}
δ = Q X Σ → Q
$q_0$ = { $q_0$}
F = {$q_3$}

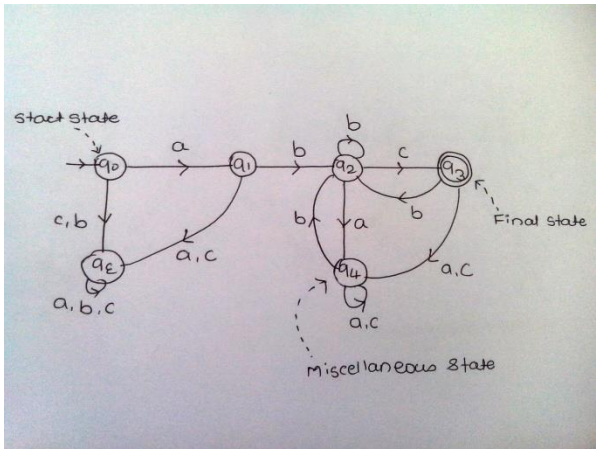



**Figure 1**: **DFA,over a language {a,b,c} which start with 'ab' and end with 'bc'**

**Transition Table:**

| M | a | b | c |
|---|---|---|---|
| $q_0$ | $q_1$ | $q_E$ | $q_E$ |
| $q_1$ | $q_E$ | $q_2$ | $q_E$ |
| $q_2$ | $q_3$ | $q_2$ | $q_6$ |
| $q_3$ | $q_6$ | $q_2$ | $q_4$ |
| $q_4$ | $q_7$ | $q_5$ | $q_4$ |
| $q_5$ | $q_7$ | $q_7$ | $q_4$ |
| $q_6$ | $q_6$ | $q_2$ | $q_6$ |
| $q_7$ | $q_7$ | $q_7$ | $q_4$ |
| $q_E$ | $q_E$ | $q_E$ | $q_E$ |

**Example 2:**

**Design a DFA,over a language {p,q,r} which start with 'pq' having sub string 'qpr' and end with 'rq'.**

Let the resultant DFA is M = (Q, Σ, δ, $q_0$, F)
Where
Q = {$q_0$, $q_1$, $q_2$, $q_3$, $q_4$, $q_5$, $q_6$, $q_7$}
Σ = {p, q, r}
δ = Q X Σ → Q
$q_0$ = { $q_0$}
F = {$q_5$}

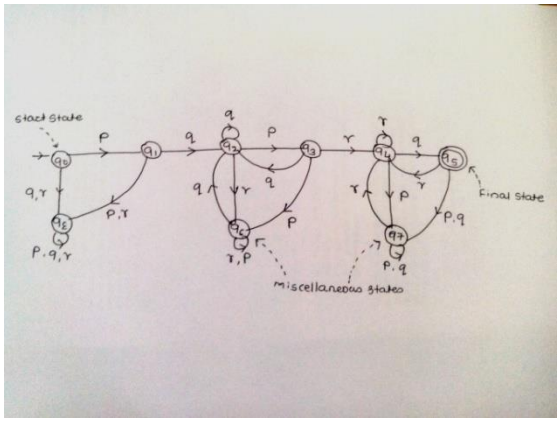



**Figure 2**: **DFA,over a language {p,q,r} which start with 'pq' having sub string 'qpr' and end with 'rq'.**

**Transition Table:**

| M | a | b | c |
|---|---|---|---|
| $q_0$ | $q_1$ | $q_E$ | $q_E$ |
| $q_1$ | $q_E$ | $q_2$ | $q_E$ |
| $q_2$ | $q_4$ | $q_2$ | $q_3$ |
| $q_3$ | $q_4$ | $q_2$ | $q_4$ |
| $q_4$ | $q_4$ | $q_2$ | $q_4$ |
| $q_E$ | $q_E$ | $q_E$ | $q_E$ |

## V.RESULT ANALYSIS AND DISCUSSION

Whenever we are having miscellaneous states, the placement of the all input symbol from that state is as follow :

**RULE or CLUE:**

The miscellaneous states input symbol placement is depends upon the previous state i.e. find out the input symbol of previous state which is having a self loop, Except that input symbol, give the self loop of all remaining input symbols to the miscellaneous state and place that input symbol which is having self loop to previous state.

**In the FIRST example - DFA,over a language {a,b,c} which start with 'ab' and end with 'bc'**.

**$q_4$** is miscellanies state. And **$q_2$** is previous state having self loop of 'b'.

**According to RULE or CLUE:**

Placement of {a,b,c} from **$q_4$** are as follows:
i) Placement of b from **$q_4$** to **$q_2$**.
ii) Rest all input symbol having self loop to the **$q_4$** i.e. {c,a} having self loop on **$q_4$**.

**In the SCECOND example - DFA,over a language {p,q,r} which start with 'pq' having sub string 'qpr' and end with 'rq'.**

**$q_6$** and **$q_7$** is miscellanies state.

And $\mathbf{q_2}$ and $\mathbf{q_4}$ is previous state having self loop of q and r respectively.

**According to RULE or CLUE:**

Placement of {p,q,r} from $\mathbf{q_6}$ are as follows:
i) Placement of q from $\mathbf{q_6}$ to $\mathbf{q_2}$.
ii) Rest all input symbol having self loop to the $\mathbf{q_6}$ i.e. {r,p} having self loop on $\mathbf{q_6}$.

Placement of {p,q,r} from $\mathbf{q_7}$ are as follows:
i) Placement of r from $\mathbf{q_7}$ to $\mathbf{q_4}$.
ii) Rest all input symbol having self loop to the $\mathbf{q_7}$ i.e. {p,q} having self loop on $\mathbf{q_7}$.

## VI. CONCLUSION

This research will definitely enhance the teaching learning environment of theory of computation and helps students to design DFA.

## VII.REFERENCES

[1.] Hopcroft, John E.; Motwani, Rajeev; Ullman, Jeffrey D.(2001). Introduction to Automata Theory, Languages, and Computation (2 ed.). Addison Wesley. ISBN 0-201-44124-1. Retrieved 19 November 2012.

[2.] Lawson, Mark V. (2004). Finite automata. Chapman and Hall/CRC. ISBN 1-58488-255-7. Zbl 1086.68074.

[3.] McCulloch, W. S.; Pitts, E. (1943). "A logical calculus of the ideas imminent in nervous activity". Bulletin of Mathematical Biophysics: 541–544.

[4.] Rabin, M. O.; Scott, D. (1959). "Finite automata and their decision problems.". IBM J. Res. Develop.: 114–125.

[5.] Sakarovitch, Jacques (2009). Elements of automata theory. Translated from the French by Reuben Thomas. Cambridge: Cambridge University Press. ISBN 978-0-521-84425-3.
Zbl 1188.68177.

[6.] Sipser, Michael (1997). Introduction to the Theory of Computation. Boston: PWS. ISBN 0-534-94728-X.. Section 1.1: Finite
Automata, pp. 31–47. Subsection "Decidable Problems Concerning Regular Languages" of section 4.1: Decidable Languages, pp. 152–155.4.4 DFA can accept only regular language

[7.] C. Allauzen and M. Mohri, Finitely subsequential transducers, International J. Foundations Comp. Sci. 14 (2003), 983-994

[8.] M.-P. Béal and O. Carton, Determinization of transducers over finite and infinite words, Theoret. Comput. Sci. 289 (2002),
225-251

[9.] Bruggemann-Klein, Regular expressions into finite automata, Lecture Notes in Computer Science 583 (1992), 87-98

[10.]J. Carroll and D. Long, Theory of Finite Automata, Prentice- Hall, Englewood Cliffs, NJ, 1989.

[11.]M. V. Lawson, Finite Automata, CRC Press, Boca Raton, FL, 2003.

[12.]D. Perrin, Finite automata, in Handbook of Theoretical Computer Science, Volume B (editor J. Van Leeuwen), Elsevier, Amsterdam, 1990, 1-57.
[13.]D. Perrin and J. E. Pin, Infinite Words, Elsevier, Amsterdam, 2004.

[14.]Ch. Reutenauer, Subsequential functions: characterizations, minimization, examples, Lecture Notes in Computer Science 464 (1990), 62-79.

[15.]J. Sakarovitch, Kleene's theorem revisited, Lecture Notes in Computer Science 281(1987), 39-50.

[16.]E. Roche and Y. Schabes (editors), Finite-State Language Processing, The MIT Press, 1997.

# COMPLEX EVENT PROCESSING AND DATA MINING FOR SMART CITIES

**Rupali Memane**
**9923061379**

**ABSTRACT**

**Complex Event Processing (CEP) is emerging as a new paradigm for continuous processing of streaming data in order to detect relevant information and provide support for timely reactions. The main role of a CEP engine is to detect the occurrence of event patterns on the incoming Streaming data. However, the problem of discovering the event patterns, although strongly related to the data mining field, has not been studied from the perspective of CEP applications. This paper presents the first steps towards defining a framework that would allow seamless integration of CEP and data mining method. We present the smart cities scenarios as a good working-field for experimentation. A concrete use case is discussed and preliminary results are presented for real-live data that has been collected.**

## INTRODUCTION

The avalanche of data which information systems had to face in the last years influenced their evolution and characteristics. Continuous, on-time processing of incoming data streams imposed particular requirements [1],which traditional Database Management Systems (DBMS) were not able to fulfill. Consequently, due to the market needs, new tools have been developed, able to process multiple data sources, often streams, in a timely fashion in order to extract relevant information. Grouped under the domain of *event processing* (or, according to [2]

information flow processing domain), two main types of such systems have emerged: Data Stream Management Systems (DSMS) and Complex Event Processing (CEP) systems.

The term *event processing* here refers to a broad study area. In [3] the term of *event processing* is coined to "any form of computing that performs operations on events". The key concept is that of an **event** which can represent anything that happens or is observed as happening (e.g. a mouse click, a sensor reading, water level increase, a river flood, spring coming, etc.). A common characteristic of event processing applications is to continuously receive such events from different **event sources** (e.g. sensors, software modules, blogs, etc.). The central module processing the events, called the CEP **engine**, detects **event patterns** from the incoming data streams and outputs the detected or predicted complex events which can be further used by other **event consumers**, or it can return as an input to the CEP engine. The event patter's role is to specify how the incoming events should be processed in order to extract relevant information. The language used to define these patterns should have the ability of specifying complex relationships among events flowing into the CEP engine. The typical approach in defining patterns of events is to manually specify them. This is done either by domain experts, capable of providing the definition of event patters or by using other tools externally of the CEP systems in order to discover these patterns and then encode them in the event processing language (EPL). However, we see the integration of machine learning algorithms with the CEP system, as a solution for direct support in definition of event patterns. Although massive amount of research has been conducted in the areas such as pattern recognition and multisensor data fusion, the systems developed for many of the CEP applications do not provide a seamless integration with such techniques, but rather consider the human component responsible for defining the complex events patters that should be monitored and detected. Therefore, an important improvement for applying machine learning algorithms in event-based application is to develop a framework that would allow easy integration of existing algorithms with event processing techniques. A first step in achieving such integration is choosing a scenario for running experiments. One example is the smart cities scenarios, as there can identified many data sources and use cases for data mining and CEP. In this paper we are proposing such a scenario, identify the data sources and run preliminary experiments for analysing the data. Future steps are discussed in the direction of using CEP engines with the patterns discovered and defining a framework for an easier integration of data mining and CEP.

## 2 SMART CITIES SCENARIO

The high level requirements for making a city smarter, as envisioned by IMB in the larger Smarter Planet1 program, refer to collaboration and coordination between city agencies managing different domains (e.g. water management, transportation, buildings, etc.) in order to be able to optimize the limited resources and to efficiently and effectively deliver city services. Moreover, different technologies may enable smarter cities, such as: communication channels (e-mail, instant messaging, etc.), business rules, data sharing (data models, accessibility) and integration of different sources of data [4]. In another study [5], the classification of cities as smart is made based on 6 criteria: economy, people, governance, mobility, environment and living. Out of these, we focus on smart mobility, which refers to transport (accessibility, modern transport systems) and availability of ICT infrastructure.

### 2.2 Description of the Data Sources Used

After receiving the data through the data sources APIs, custom built adapters are used for storing data in a uniform data structure, which allows us to integrate all the sources for generating the input dataset for data mining. As illustrated in Figure 1, for the data storage functionality we have used the QMiner infrastructure which is based on tightly integrated and scalable custom software modules. The data mining algorithm applied is for learning association-rules

Depending on the how often the sources were updated, different time intervals were used for crawling data as can be observed in **Error! Reference source not found;**

Table 1: *Time intervals for data collection*

| Source | Update time interval |
|---|---|
| Tfl Road Disruptions | 5 minutes |
| Current Weather Conditions[4] | 30 minutes |
| Events (from Eventful.com) | Once per week |

Table 1: *Time intervals for data collection*

The road disruption events are identified with a unique id and have the following properties: start and end time, location details, time of last update, type, severity and category. The category property is described in **Table 2,** as it has predefined values which are used in the analysis of the results, while for the rest of properties more details can be found in [6].

The total number of road events registered is 3090. The type of the events indicates if the event has happened during the period for which the data was crawled, or is it scheduled to happen in the future (caused by road works or other planned cultural events).

Table 2: *Categories of road events and their frequency*

| Category | Interpretation | Percentage |
|---|---|---|
| Works | planned and emergency road works | 7% |
| Accident | road traffic accidents | 30% |
| Signal Failure | automatic traffic signals failure | 6% |
| Breakdown | vehicle breakdown | 19% |
| Incident | emergency incident | 4% |
| Event | cultural events | <1% |
| Hazard | dangerous structures, fire, flooding, ice and spillages | 6% |
| Other Cause | abnormal loads, unexplained congestion, etc. | 27% |

Unfortunately the precipitation information, which could presumably have an influence on traffic disruptions, was not available from this data source.

The cultural events are described by time, location, performers and can be of more categories (e.g. music, concert, arts, sports, etc.). The total number of events happening collected is 5931, in 1707 different locations. The top 10 most frequent categories of events are presented in **Table 3** (one event can have more categories).

## 3 DATA PREPROCESSING

The data preprocessing consisted of two main steps: (1) data integration and (2) generation of the set of instances. In the data integration step, the traffic events were correlated with cultural events and weather data based on time and location. We have also defined nearby events, as follows: if an event is situated at predefined maxim distance from the current event and starts with a predefined maximum time interval before or after the current event,

Table 3: *Frequent categories of cultural events*

| Category | Freq. | Category | Freq. |
|---|---|---|---|
| Music | 2652 | Sales | 215 |
| Performing Arts | 1138 | Festivals-Parades | 200 |
| Other | 887 | Family-fun-kids | 157 |
| Singles social | 279 | Outdoors-recreation | 124 |
| Sports | 351 | attractions | 100 |

Table 4: *Instance Attributes*

| Attribute | Type | Distinct / [min-max] | Missing Values |
|---|---|---|---|
| category | Nominal | 8 | No |
| severity | Nominal | 3 | No |
| day | Nominal | 7 | No |
| time | Nominal | 4 | No |
| duration | Nominal | 3 | 27% |
| traffic events nearby | Nominal | 2 | No |
| weather | Nominal | 12 | 2% |
| Temperature (°C) | Numeric | [10-30] | 2% |
| wind speed (kph) | Numeric | [0-32] | 2% |
| events nearby | Nominal | 2 | |
| cultural events categories (25) | Nominal | 2 | No |

The maximum distance and time interval have been tested with different values, as it will be explained in the next section. In the second step, we have created a set of instances around the road events, where each instance has 37 attributes, described in **Table 4**. As there can be more categories of cultural events, each is represented as an attribute which has value *t* if category is present and *f* otherwise. An example of an instance is illustrated in

## 4 PRELIMINARY RESULTS

First, analysis was done on each attribute. The analysis with respect to day of the week and time of the day is illustrated in **Figure 3**, and presents the expected correlation.

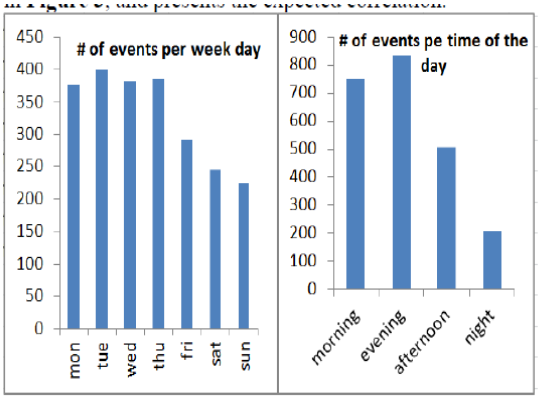



Figure 3: *Number of events per week day (left) and time of the day (right)*

Another aspect of importance is the presence of nearby events (road or cultural) for an instance of road event. This has been calculated for three sets of parameters, and the results in **Table 5** show the number of instances where nearby traffic events (#tfl), respectively cultural events (#evt) exist.

Table 5: *Number of instances that have nearby events for different constraints on distance and time difference*

| Distance (m) | Time diff. (mins) | #tfl | #evt |
|---|---|---|---|
| 500 | 60 | 9 | 104 |
| 1000 | 60 | 31 | 213 |
| 1000 | 240 | 96 | 487 |

Possible correlations between the attributes of our dataset have been studied using association rules. The algorithm used for discovering such rules is the Apriori algorithm, implemented in Weka. We choose the dataset with the constraints of 1000 meter in distance and 60 minutes time difference. As Weka crashed when running the algorithm on all the attributes we first try removing all the categories of cultural events from the attributes, reducing the number of attributes to 10. However no relevant rules were found.

As our interest was in the relation of different traffic events categories (listed in **Table 2**) with nearby cultural

events, we have reduced the dataset to 213 instances (for which the constraints on distance and time where 1000 meters, respectively 60 minutes), which had at least one nearby cultural event. Although the rules obtained are not necessarily related to traffic events, they do illustrated normal relations, such as: cultural events are more often in the evening (rule 1) or that some cultural events categories are related (rule 2)

Rule 1: Weather = Clear, music = t, performing arts = t

(23) ==> Time=Evening (21) [conf:(0.91)]

Rule 2: singles_social = t. performing_arts = t (33) ==>

music = t (28) [conf:(0.85)]

## 5 CONCLUSIONS AND FUTURE WORK

A number of conclusions can be drawn after our preliminary experiments as follows. Data must be collected for a longer time period, allowing thus generation of cleaner and more precise datasets. Although the 27% of missing values for the duration attribute (see **Table 4**) is not very high, it does influence the calculation of the nearby events. In our experiments we used an approximation of average duration, for being able to obtain other nearby events. A larger dataset would allow for extraction of data of better quality, which can then be used more successfully for association-rule learning. Determining the nearby events may be done differently for different categories of cultural events; (e.g. a big music concert affects a larger area then a sales event) a more thoroughly study for determining the parameters involved is needed. There is no clear evidence of the influence of cultural events over road events. However, common sense tells us that there should be. Therefore we shall continue our study once we gather more data.

As future steps we consider integration of more data sources, and first we will focus our attention on social media. Next we will also consider visualization techniques that can provide a faster insight into the data, and then proceed with data mining methods. Finally, once the event pattern obtained, we will research into connecting them to a CEP engine.

## References

[1] Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. ACM SIGMOD Record. 34, 42-47 (2005).

[2] Cugola, G., Margara, A.: Processing Flows of Information : From Data Stream to Complex Event Processing. ACM Computing Surveys.

[3] Niblett, P.: Event Processing In Action. (2010).

[4] Wang, Q., Meegan, J., Freund, T., Li, F.T., Cosgrove, M.: Smarter City: The Event Driven Realization of City-Wide Collaboration. 2010 International Conference on Management of e-Commerce and e- Government. 195-199 (2010).

[5] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers, E.: Smart cities Ranking of European medium-sized cities. , Vienna, Austria(2007).

[6] Transport for London, Live Traffic Disruptions – Data Dictionary, (last accessed 1st September 2012), http://www.tfl.gov.uk/assets/downloads/businessandpartners/data-dictionary-live-traffic-disruptions.pdf

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume11, Issue 1, 2009.

# A survey on Web mining: Data mining approach

**Nilesh Magar [1] , Aniket Nagane[2]**
[1 2] MAEER's Arts, Commerce and Science College, Pune(India)
[1] nileshraghav@gmail.com, [2] aniket.nagane@gmail.com

## ABSTRACT:

**The Web has become very important and versatile tool for almost all application in day today life. Internet is a hug source of information. Unfortunately information from many sources is not available in the form which could be mined properly. Web mining deals with extracting the interesting patterns and developing useful abstractions from diversified sources. Paper deals with the comparative study of preprocessing tasks, applications and challenge of web mining.**
**Key words:** web mining, knowledge discovery

## I INTRODUCTION

Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late [2]. Internet can be considered as a huge semi structured database, presenting all the problems implicit in semi-structured data. Extracting the structure of every HTML document is a challenging issue given the absence of predefined standard and schema [3]. But an internet is continuously becoming a central part of social, cultural, political, educational, academic and commercial life as it contains a wide range of information and applications.

Web mining can be broadly defined as the discovery and analysis of useful information from the internet[4]. Data can be collected from various sources like the server side, client side, proxy servers, or it can be obtained from an organization's database. This data is mostly large and hyperlinked. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and Meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are[4].

1) Unlabeled;

2) Distributed;

3) Heterogeneous (mixed media);

4) Semi structured;

5) Time varying;

6) High dimensional.

Web mining tasks can be summarized as follow:

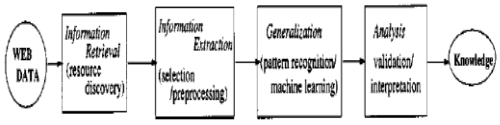



Fig. 1. Web mining subtasks.

**Information Retrieval:** This process deal with the retrieval of relevant information from web data. The IR process mainly includes document representation, indexing, and searching for documents [4]. While retrieving information, methodologies used should ensure that non-relevant information should not extracted or very few amount of non relevant information should be retrieved.

**Information Selection/Extraction and Preprocessing:** once relevant information is retrieved from web data, important task is to extract knowledge and other required information without human interaction. The major methods of information retrieval involve writing wrappers (hand coding) which map the documents to some data model[4]

**Generalization:** Here in this phase pattern recognition and machine learning methods are used on the extracted

data. A major problem while web mining is the labeling problem: hug amount of data is available on the web but it is unlabeled. Data mining techniques require proper labeled data as positive (yes) or negative (no) An

approach to solve this problem is based on the fact that the web is much more than just a linked collection of documents, it is an interactive medium[4].

**Analysis:** Analysis is actual phase of Knowledge discovery in databases. In this phase humans play an important role as web is an interactive medium. In this phase validation and or interpretation of the mined patterns is performed for which appropriate tools like,Webviz system are required to use to understand, visualize, and interpret these patterns[4].

## II WEB MINING CATEGORIES:

Web mining may be of three types, namely, WCM, web structure mining (WSM), and web usage mining (WUM) [4].

**1) WCM**: WCM deals with the discovery of useful information from the web contents/data/documents/services.

However, web contents are not only text, but encompass a very broad range of data such as audio, video, symbolic, metadata, and hyperlinked data. Out of these, research at present is mostly centered around text and hypertext contents.

**2) WSM**: WSM refers to mining the structure of hyperlinks within the web itself (inter document structure unlike

WCM, which refers to intra document structure). Here, structure represents the graph of the links in a site or between sites.

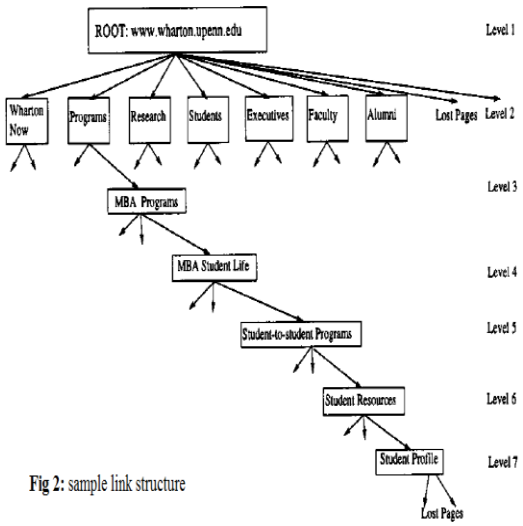



**Fig 2:** sample link structure

**3) WUM:** While content mining and structure mining utilize the real or primary data on the web, usage mining mines secondary data generated by the users' interaction with the web.Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web.WUM works on user profiles, user access patterns, and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites. WUM plays a key role in personalizing space, which is the need of the hour. To satisfy all the users with the same tool is extremely difficult, and, instead, we need to learn user access patterns, their path patterns at an individual level, and also as a whole at web sites.

## III APPLICATIONS OF WEB MINING:

Web mining may benefit those organizations that want to utilize the web as a knowledge base for supporting decision-making. The most dominant application area for WM is related to Internet based e-commerce (business-to-consumer) and Web-based customer relationship management (CRM) an integral part of E-business today.

Few of the web mining applications are discussed here in brief:

**Market analysis and a management** (market basket analysis, cross-selling, market segmentation, etc.): By analyzing web data organizations can take the decision about their marketing policies. Organizations can easily identify various segments of market and can decide policies for the promotional activity as well as direct sell of their products.

**Risk analysis and management** (forecasting, customer retention, etc.) : Organizations can also benefited from web mining to identify the various risk issues in the market and accordingly they can decide their marketing policies to cope up with or to overcome from that issues. Organization can determine current market situations (forecast) and accordingly manage those risks.

**Fraud detection and management** (e-commerce, etc.): As there is a growth in e-commerce, everyday online fraud is also becoming a major issue. Web mining can also help in detection of fraud transaction, illegal online gambling, hacking, virus spreading, child pornography distribution and cyber terrorism.

**E-Learning**: Web mining can be used for enhancing the learning process in e-learning environments. Applications of web mining to e-learning are usually web-usage based

applications. the use of machine learning techniques and web usage mining to enhance web-based learning environments for the educator to better evaluate the learning process and for the learners to help them in their learning task[6].

**E-Government**: The process by organizations that interact with citizens for satisfying user (or group of users) preferences leads to better social services. The major characteristics of e-government systems are related to the use of technology to deliver services electronically focusing on citizens needs by providing adequate information and enhanced services to citizens to support political conduct of government. Empowered by web mining methods e-government systems may provide customized services to citizens resulting to user satisfaction, quality of services, support in citizens decision making, and finally leads to social benefits. Such social benefits much rely on the organization's willingness, knowledge and ability to move on the level of using web mining[6].

**E-Politics and E-Democracy:** E-politics provides political information and politics "on demand" to the citizens improving political transparency and democracy, benefiting parties, candidates, citizens and the society. Election campaigners, parties, members of parliament and members of local governments on the web are part of e-politics. Despite the importance of e-politics in democracy there is limited web mining methods to meet citizen needs.

Business Intelligence, Customer Relationship Management, Bio-Informatics, and Knowledge Management are also candidate applications of web mining.

## IV CHALLENGES IN WEB MINING:

Web mining is not like traditional mining in which we extracts patterns from well structured/ well formatted data. Here we have discussed some basic problems of web mining.

**Quality of keyword - based searches**: The quality of keyword-based searches suffers from several inadequacies such as a search often returns many answers, especially if the keywords posed include words from popular categories such as sports, politics, or entertainment [5]. For example, depending on the context, an apple could be a fruit, juice, company or computer and a search can miss many highly related pages that do not explicitly contain the posed keywords and, a search for the term data mining can miss many highly regarded machine learning or statistical data analysis pages.

**Working with Unstructured/ Semi Structured Data**: Internet can be considered as a huge unstructured/ semi structured database, Working with such data is challenging job as extracting the structure of every HTML document is a challenging issue given the absence of predefined standard and schemas.

## V CONCLUSION:

Web mining is a new and promising research issue to help users in gaining insight into overwhelming information on the Web. In the field of science and technology management, web mining has more important value and meaning, how to extract and use of the rich internet information resources is the key making the technological workers and decision-makers to victory. This paper aims to explore the preprocessing tasks, applications and challenge of web mining in the science and technology management, and forecast the potential of combining the present and future applications, hope to bring enlightenment for related people.

## REFERENCES:

*1 Pawan Lingras, Rough set clustering of Web mining, 0-7803-7280-8/02/$10.00 c 2002 IEEE*

*2 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996) (© AAAI)*

*3 Subashini S, Singhania University, Mahesh T. R, Professor, Dept. of CSE, TJIT, Bangalore, Rajasthan,Web Mining: Prominent Applications and Future Directions, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555Vol. 2, No.4, August 2012*

*4 Sankar K. Pal, Fellow, IEEE, Varun Talwar, Student Member, IEEE, and Pabitra Mitra, Student Member, IEEE, Web Mining in Soft Computing Framework:Relevance, State of the Art and Future Directions, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 5, SEPTEMBER 2002 1163*

*5 T. Sunil kuma[1]r, Dr. K. Suvarchala[2], 1 Department of CSE, Krishnamurthy Inst. Of Tech. and Engineering, India, 2Departmentof CSE & Informatics, Bankathlal Bhadruka College of IT,India, OSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727Volume 7, Issue 3(Nov. -Dec. 2012), PP 24-29*

*6 Georgios Lappas, Technological Educational Institution of Western Macedonia, An Overview of Web Mining in Societal Benefit Areas*

# A Review On Heuristics: Search Techniques In AI

**Minal Rajendra Pawar**
Student, VPASC College, Baramati.
minalpawar886@gmail.com

## ABSTRACT:

**A process that may solve a given problem, but offers no guarantees of doing so, is called a heuristic for that problem. The aim of heuristic is to study the methods and rules of discovery and invention. Heuristic, as an adjective, means "serving to discover."**

**Heuristics are efficient cognitive processes, conscious or unconscious, that ignore part of the information. Because using heuristics saves effort. Heuristic discusses human behavior in the face of problems; this has been in fashion, presumably, since the beginning of human society, and the quintessence of such ancient discussion seems to be preserved in the WISDOM OF PROVERBS.**

***Key words:*** **heuristic, problem solving, artificial intelligence, cognitive science, Heuristic search**

**Introduction:**

**1. What is a Heuristics?**

The term heuristic is of Greek origin and means, "serving to find out or discover."

A search heuristic is information to guide the traversal into the direction of the set of the goals.
In a weighted state space problem graph G = (V,E, s, T,w) a heuristic h is a node evaluation function mapping V to IR+0 .
We require
• fort in T: h(t) = 0
• all other nodes v 2 V we have h(v) 6= 0
so that the goal check for u simplifies to compare h(u) with 0.
The concept of heuristic began to appear in the early 1950's A1 literature and was well known by the early 1960's. From the mid 1950's to the present, the notion of a heuristic has played a crucial role in the A1 researcher's descriptions of their work. Things that would be called a heuristic by one researcher would not be so called by others. This is because many heuristics embody a variety of different features, and the various researchers have emphasized different ones of these features as being essential to being a heuristic. That the concept of a heuristic has been, and continues to be, central in A1 is too well known to require documentation.
Heuristic is a science of problem solving behavior that focuses on plausible, provisional, useful, but fallible, mental operations for discovering solutions.

A heuristic (heuristic rule, heuristic method) is a rule of thumb, strategy, trick, simplification, or any other kind of device which drastically limits search for solutions in large problem spaces.

Search
Another example of heuristic making an algorithm faster occurs in certain search problems. A heuristic method can accomplish its task by using search trees.

**Use of heuristics:**

Two answers have been proposed to the question of why heuristics are useful: the accuracy-effort trade-off, and the ecological rationality of heuristics.

**Accuracy-Effort Trade-Off**

The classical explanation is that people save effort with heuristics, but at the cost of accuracy; heuristics trade-off some loss in accuracy for faster and more frugal cognition.

There are two interpretations of this tradeoff:
(a) Rational trade-offs.
(b) Cognitive

**2. Ecological Rationality**

"A heuristic is ecologically rational to the degree that it is adapted to the structure of the environment"

**Heuristic function** h(n):-
A heuristic function, also called simply a **heuristic**, is a function that ranks alternatives in search algorithms at each branching step based on available information to decide which branch to follow.
Heuristic function h(n)

- Admissible/Underestimate:
- Consistent/monotonic

**Admissible/Underestimate**

h(n) never overestimate the actual cost from n to goal. A heuristic h is admissible if it is a lower bound estimate that underestimates the shortest remaining distance to the goal.

**Consistent/monotonic(desirable)**

Many heuristics have monotone estimator functions:
h(m)-h(n)<=w(n,m)
Where, m is parent of n.
This ensures
f(n)>=f(m).
A heuristic h is consistent, if h(u) _ h(v)+w(u, v) for all edges
e = (u, v)

**Properties of Heuristic functions:**

The 2 most important properties:
–relatively cheap to compute
–relatively accurate estimator of the cost to reach a goal. Usually a "good" heuristic is if ½ **opt(n)<h(n)<opt(n)**

Examples:
–Navigating in a network of roads from one location to another.
–Sliding-tile puzzles.

**3. Conclusion:**

We began this review with the observation that the three major tools for modeling decision making—logic, statistics, and heuristics—have not been treated equally, with each suited to a particular kind of problem.

Heuristics can be more accurate than more complex strategies even though they process less information (less-is-more effects).

A heuristic is not good or bad, rational or irrational, its accuracy depends on the structure of the environment.

Usually, the same heuristic can be used both consciously and unconsciously, for inferences and preferences, and underlies social as well as nonsocial intelligence.

**4. References:**

1. Astebro T, Elhedhli S [2006]. The effectiveness of simple decision heuristics: forecasting commercial success for early-stage ventures.

2. Allen Newell and Herbert A. Simon [1976]. "Computer Science as Empirical Inquiry: Symbols and Search". Comm.

Books:

3. Artificial Intelligence Third Edition Mc Graw Hill Education Elaine Rich, Kevin Knight, Shivshankar B. Nair
4. Artificial Intelligence Vision Publication Third Edition

# Automated Boolean Matrix Data Representation Approach for Efficient Mining

*Renju Samuel and Linu Paulose*

.

**ABSTRACT:**

**In the modern world the process of identifying the required items from a large dataset leads to a large amount of time. The concept of Data Mining is normally used to isolate the required item from a large database. In this paper a new form of isolating the data from a database is used. The traditional Apriori Algorithm is used to generate the Association rules by obtaining frequent itemsets of 1 combination, 2 combinations and so on based on the minimum support count. This study aims to propose a modified Apriori algorithm in which generation of Association rules is by using a Boolean Matrix representational scheme. At first the data representation takes place via the generation of an AVL tree which is used to balance the given data. Then the Boolean Matrix gets generated from the AVL tree and then association rules get generated from the matrix. Its shows clearly that the computation time has improved for the modified algorithm when compared to the traditional Apriori Algorithm**

**Index terms:** Association Rule, AVL Tree, Boolean Matrix,Data Mining, Frequent Itemset.

## 1 – INTRODUCTION

The process analysing data from different perspectives and categorising it into useful information. It allows the users to analyse the database from differentangles. Technically data mining can be viewed as the process of finding pattern among sets of fields among relational databases.Data mining is used in business to discover pattern, trends,relationships, anomalies in order to make better decisions.Apriori algorithm is designed to operate on transactional databases. It identifies the frequent individual items in the database and extends them to larger item sets as long as those appear sufficiently often in the database.

A "bottom up" approach where frequent subsets are extended as one item at a time and groups are tested against the data is been used by Apriori algorithm. Number of items in the item sets are counted using breadth-first search and hash tree mechanisms in Apriori algorithm. Association rule which highlights the general trends in data bases can be determined by frequent sets whichresults from Apriori algorithm. The popular method for discovering interesting relations between variables in large databases is called association rule. It identifies strong rules discovered in databases using different measures. Association can be used to discover regularities between products in large scale transaction of data's in supermarkets. For example, the rule

{Milk, Coffee powder} ⇨ {Sugar}

From the sales data of a supermarket would indicate that if a customer buys milk and coffee powder together is likely to buy sugar also. Such information's can be used as the basis for decisions about business and marketing activities such as pricing and product placements. Association rule has many applications including Bioinformatics,Continuous production, Web usage mining and intrusion detection. It doesn't usually consider the order of the item either within a transaction or across the transaction.Apriori algorithm uses two rules they are

1. Find all the frequent sets, each of the frequently occurring at least as frequently as minimum support rule which is predetermined.
2. Generate association rules from the frequent item sets.

Mining association rule is converted to find the frequent item sets. But traditional Apriori algorithm has two drawbacks they are

3) Many input –output operations are needed to scan databases frequently. It needs to scan the transaction database as the same number as the elements of the frequent item sets.
4) When analysing the databases we get large number of tables, which also requires a lot of time and memory (storage space).

To overcome these drawbacks we propose modified Apriori algorithm which uses data representation schemes. The algorithm is improved through AVL-tree based scheme which provides more reduced time and condensed rule than the traditional Apriori method. Using this method it is able to get last frequent itemsets and generates association rule depending on the data representation and Boolean matrix.In Apriori algorithm we need to access the database frequently which consumes both time and memory space. In modified Apriori algorithm which is based on data representation and Boolean matrix reduces this overhead.

## 2 - RELATED WORK

Agarwal R and R Srikant suggested two new algorithms named as Apriori and Apriori Tid are used, for discovering all association rules between items in a large database. Finally they compared these algorithms to previously known algorithms such as AIS and SETM and the new implemented algorithms are much more effective than the existing algorithms [1].

Mannila, H.H Toivenn and A. Verkamoproved that an algorithm that uses all existing information between database, and the extra step they have done is that it passes to avoid checking the coverage of duplicate itemsets . It provides a better execution time when compared to the previous results and it is simple to implement. Concluded in a way that sampling is an efficient technique for finding rules of this type and that algorithms working in the main memory can be used to obtain very good approximations [2**].**

Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan had used a way to discover patterns present in very large databases other than to verify that a certain pattern exists in that database. The result obtained produced high performance and also linear scaling in real life databases. Also they had provided a completeness property that promises all the similar patterns present in the database of certain type will be identified [3].

Hassan M. Najadat, Mohammed Al-Maolegi, Bassam Arkokhave used several groups of transaction, and various minimum support count values in the database. And they have applied the same values to both the normal and improved Apriori algorithm. Theconclusion obtained is that the new improved Apriori algorithm had improved the time consumption by 67.38%. Also another result is that it is more efficient when compared to the normal apriori algorithm [4]. t Bay and Bac Le have presented a new form of Apriori algorithm for obtaining the frequent itemsets and to generate the association rules. They scan the database only once in order to generate the transaction ids to obtain the frequent itemsets. A tree structure is used to obtain these results. They omit the duplicate itemsets in order to obtain the results of generating the association rules at a faster phase [5].

An Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z , they had used Hash function instead of normal joining method. Transactionreduction is done by generating hash function using DHP algorithm. Collision in hash table reduces the effectiveness of hash table. Hash functions are complex to be coded after 2-itemset combination.(DHP- Direct hashing and pruning) [6].

Dhilhan perera, Judy Kay, Irena Koprinska, Kalina Yacef , and Osmar Zaiane have proposed high level view of information from the database, that is together they pair the patterns of the same group. Their goal was to enable the particular groups and their identities the various operations of the group and the feedback that is provided during the operations. The tool used in order to obtain the result is TRAC. They extract patterns of the strong groups eliminating the weaker groups to get more successful results [7].

Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani had proposed a way to find all implication and redundancy rule based on confidence pruning without support pruning. Dynamic pruning done to the large data set by counting number of rows and avoid the rows containing missing values. Only confidence pruning is considered. They focused only on the reduction in the transaction size and noton efficient rule generation [8].

## 3- PROPOSED WORK

Initially the data present in the given database gets sorted out in ascending order by using the AVL tree .The next step is the data discretization of the sorted data. Here the sorted values from the AVL tree gets replaced according to the Index values provided to it. After this particular step the Hash Map table is being generated using the values obtained in the previous step. The Hash Map table consist of transaction id and the List of items present in the list. From these values the Boolean Matrix is obtained. All the places at which the value is present are termed as 1 and the place at which value is not present is termed as 0. This result is formed by using the Hash Map table. And the last step is generation of Association Rules from the Boolean Matrix.

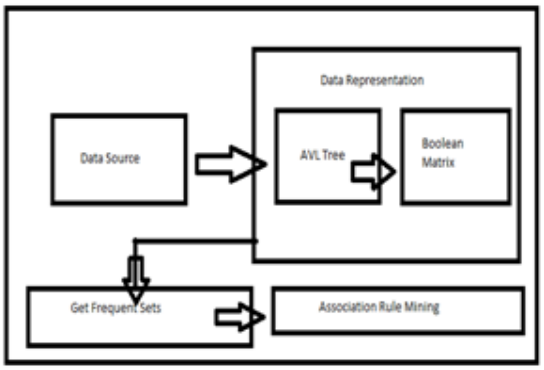



Fig. 1. Architecture Diagram

Data representation transforms source data into an AVL tree which is a balanced structure. Also it reduces the computational time for the process. Also it obtains various multi data types of the represented data from the database such as a hash table, tree and Boolean matrix.

Shown below (Fig. 2) is an example for the generation of Association rules using the modified approach

.

Table 1 : Given Dataset

| Transaction ID | List Of Items |
|---|---|
| T101 | 8,14 |
| T102 | 10,12,14 |
| T103 | 8,12,14 |
| T104 | 10,14 |
| T105 | 8,24 |
| T106 | 12,14 |
| T107 | 8.12 |
| T108 | 12,14 |

Here the given value gets balanced by the AVL Tree from the database (Fig. 3).

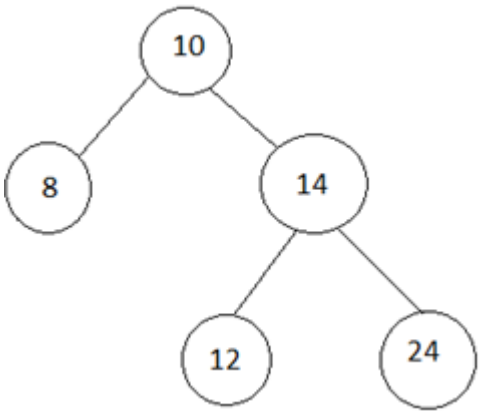



Fig. 2. Generation of AVL Tree

After the balanced values are obtained from the AVL Tree, the data discretization (Fig. 4) is done by replacing the elements present at the particular level in the database.

Table 2: Data Discretization

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 8 | 10 | 12 | 14 | 24 |

Next step is the generation of Boolean Matrix (Fig. 5), it is done by substituting the values which are obtained from the previous step, values are substituted as 1 in the place of element is present and 0 at the place where presence of element is null.

Table 3: Boolean Matrix

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| T101 | 1 | 0 | 0 | 1 | 0 |
| T102 | 0 | 1 | 1 | 0 | 1 |
| T103 | 1 | 0 | 1 | 1 | 0 |
| T104 | 0 | 1 | 0 | 1 | 0 |
| T105 | 1 | 0 | 0 | 0 | 1 |
| T106 | 0 | 0 | 1 | 0 | 1 |
| T107 | 1 | 0 | 1 | 0 | 0 |
| T108 | 0 | 0 | 1 | 1 | 0 |

Hence the Boolean Matrix is generated. Now the next step is to find the frequent itemsets from the obtained Boolean Matrix

$$\text{Computed Weight} = \frac{\text{Count 1's in Column}}{\text{Number Of Rows}} >= S$$

Here S is the minimum support count given by the user

Table 4: Generation of Frequent Itemsets

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| T101 | 1 | 0 | 0 | 1 | 0 |
| T102 | 0 | 1 | 1 | 0 | 1 |
| T103 | 1 | 0 | 1 | 1 | 0 |
| T104 | 0 | 1 | 0 | 1 | 0 |
| T105 | 1 | 0 | 0 | 0 | 1 |
| T106 | 0 | 0 | 1 | 0 | 1 |
| T107 | 1 | 0 | 1 | 0 | 0 |
| T108 | 0 | 0 | 1 | 1 | 0 |
| COUNT | 4 | 2 | 5 | 4 | 3 |
| Count / Total Number of Rows | 4/8 = 0.5 | 2/8 =.25 | 5/8 =.625 | 4/8 =.5 | 3/8 =.375 |

Here all 1's are counted in each column and it is added and divided with the total number of rows. Let us take the minimum support count (S) as 0.5,so all values which are equal to and greater than equal to 0.5 is taken into consideration. Rest all values are rejected (Fig. 6).
Hence from the following table, the values 0.5,0.625, 0.5 are selected. Its following values are taken inorder to generate the Association Rule from the table
In this example the last frequent itemset are {1, 3, and 4}
The following Rules can be generated
**Rule 1**: 1^3 →4,**Rule 2**: 1^4 →3
**Rule 3**: 3^4 →1,**Rule 4**: 1 →3^4
**Rule 5**: 3→ 1^4,**Rule 6**: 4 → 1^3
Thus the Association Rules are generated from the given dataset using the Modified Apriori Algorithm

## 4 – CONCLUSION

The following study had provided suitable information such that with the data representation scheme that is followed here reduces the computational time. It also improves the various

processes such as the generation of n-frequent itemsets and also the generation of the Association Rules when compared to the normal Apriori Algorithm. The major advantage of using this type is that the entire database needs to be scanned only once inorder to generate the Association Rules. Thus it saves a lot of computational time than the normal Apriori, which has to invoke the database each and every step inorder to generate its Association Rules.

**REFERENCES**


[15]. **Agarwal R and R Srikant**, **1994, "Fast Algorithms for Mining Association Rules in large databases"** : Jorge B Bocca , Matthias Jarke and Carlo Zaniolo, editors , Proceedings of the 20th International Conference on Very Large Data Bases , VLDB , Santiago , Chile

[16]. **Mannila , H.H Toivenn and A. Verkamo , 1994** : **Efficient algorithm for discovering association rule**s . AAAI workshop on Knowledge Discovery Database Seattle 181-92

[17]. **The Quest Data Mining System : Rakesh Agrawal and Manish Mehta and John Shafer and Ramakrishnan Srikant** , IBM Almaden Research Center San Jose, California 95120, U.S.A 1996.

[18]. **An Improved Apriori Algorithm for Association Rules , Hassan M. Najadat, Mohammed Al-Maolegi , Bassam Arkok**, International Research Journal of Computer Science and Application Vol. 1, No. 1, June 2013, PP: 01 – 08

[19]. **Fast Algorithm for Mining Generalized Association Rules , Bay Vo and Bac Le** , International Journal of Database Theory and Application Vol. 2, No. 3, September 2009

[20]. **Effective Hash Based Algorithm for Mining Association Rules by Hao Y, Ye Li and Cheng Z,** SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data

[21]. **Clustering and Sequential Pattern Mining of online Collaborative Learning Data , Dhilhan perera , Judy Kay , Irena Koprinska ,** Member , IEEE Computer, Society , Kalina Yacef , and Osmar Zaiane , Member , IEEE IEEE Transactions On Knowledge And Data Engineering , Manuscript ID 2011

[22]. **Dynamic Miss Counting Algorithm Finding Implication and Similarity Rules with Confidence Pruning,Shinji Funjiwara, Jeffrey D. Ullman and Rajeev Motwani ,** Data Engineering, 2000. Proceedings. 16th International Conference on29 Feb 2000-03 Mar 2000 ,501 – 511

[23]. **A study on the Performance of CT-APRIORI and CT-PRO Algorithms Using Compressed Structes for Pattern Mining ,** Journal of Global Research in Computer Science ,Volume 1 , Number 2 September 2010

[24]. **The Quest Data Mining System : Rakesh Agarwal , Manish Mehta , John Shafer and Ramakrishnan Srikanth ,** IBM Almaden Research Centre , San Jose, California 95120, U.S.A 1996

[25]. **Saravanan Suba and Chistopher.t. Article: A Study on Milestones of Association Rule Mining Algorithms in Large Databases .** International Journal of Computer Applications 47(3):12-19, June 2012. Published by Foundation of Computer Science, New York, USA

[26]. **Mining Sequential Patterns: Generalizations and Performance Improvements Ramakrishnan Srikant and Rakesh Agrawal F Srikant**, IBM Almaden Research Center 650 Harry Road, San Jose,CA95120

[27]. **Information Retrieval through Semantic Web, Gagandeep Singh, Vishal Jain ,** International Journal Of Database , Vol 8, No 5june 2012

[28]. **MiningAssociation Rules Based on Cloud Model and Application in Credit Card Marketing ,Yan-li Zhu, Yu-Fen Wang , Shun-Ping Wang , Xiao-juan Guo ,Shenzhen,** China , pp.165-168, 2010 Asia-Pacific Conference on Wearable Computing Systems, 2010



Renju Samuel is currently working as Assistant Professor , Computer Science Department in KMP College of Engineering , Kerala , India, PH- +91 9645105144. E-mail: renju.samuel@gmail.com

Linu Paulose is working as Head Of The Department in Computer Science department in KMP College Of Engineering , Kerala, India PH- +91 9947708656. E-mail: linupaulose2010@gmail.com

# Collaborative Multiparty Data Mining

**Deepak Joshy**
Program Analyst
L&T Technologies, Chennai
deepakjoshy@gmail.com

## ABSTRACT:

***Collaborative pattern mining is a pattern mining technique used in a distributed environment where each site participate collaboratively with one another for the discovery of patterns. Each individual site is capable of mining patterns and can communicate with other sites in the distributed environment. Currently existing work have failed to tackle various problems using this technique like effective data exchange and cross-sitepattern pruning mechanisms. This paperadvocates an effective method to perform pattern pruning in a distributive manner with the help of a data structure called Bloom Filter. With its help cross site pattern pruning can be performed with minimized data transfer between the sitesas well as increased speed and with the capability of mining Local, Global and Inter-database Patterns.***

***Keywords— Distributed data mining, Bloom Filter, Frequent item set, Distributed associative rule mining***

## I. INTRODUCTION

Pattern mining is the process of discovering patterns or relations that are present among data in very large dataset or database. The discovery of the patterns can help to understand the data better and we can make use of the knowledge discovered for future purposes.For many applications data is collected from various distributed sources and these sources may be distributed over different locations [3,4]. The amount of data will be huge and lot of processing and data transfer will be required in order to discover the patterns.

From an association rule mining point of view ,past researches have made significant efforts to discover a variety of patterns such as frequent item-sets, temporal, spatial, and/or sequential association rules, closed patterns or sequential patterns. Common challenges that are faced in this area are: (1) identification of patterns from continuous volume of data or from large databases [13](L-Patterns) and (2) discovery new patterns by unifying multiple databases into a single view[4,5] ( G-pattern mining ). In distributed databases, common goal is discover G-patterns, usually done by unifying multiple databases into a single view. Collective data mining [6] is a typical research work done in this area. A common way is find promising local candidates and sending them to a central site for synthesizing [7,8].In situations where two or more companies/organizations agreed to share data for research formutual gain but they are not willing to disclose all their sensitive data, so there should be some mechanism to share the essential data without compromising the data security. Hence there should be some efficient way to exchange data effectively along with a mechanism to monitor the data being shared.

Patterns that can be discovered in a distributed environment are of three kind 1) Local Patterns (L-Pattern), 2) Global Patterns (G- Pattern), and 3) Inter-database Patterns (I- Pattern). Local Patterns [9] are those patterns that are local to a certain distributed site. Global patterns are those patterns that are present in all the distributed sites and inter-database patterns are the patterns that are present in more than one site but not all. Local and Global patterns can be found out with ease with already existing method but in-order to find Inter-database patterns no efficient methods currently exists.

## II. RELATED WORK

Mining of data in a distributed environment can be done in 3 ways, 1) Sequentially 2) Paralleland 3) Collaboratively.In sequential way, mining is performed in a sequential manner starting from the first site, discovering patterns as we move along to other sites one after the other. When the final site is reached the global patterns will discovered.This method is incapable of finding out inter-database patterns.In the parallel approach, there will a master site which will collect data from all the distributed sites and performs mining and returns the result to the user. The disadvantage with this approach is that there is the need of central site and in case when data mining is carried out for different companies with different requirements, there must be multiple master sites controlled by each company/organization for their own purposes.

In the collaborative approach the distributive sites are capable of finding out patterns by their own. They will exchange message to communicate with other distributed sites. This is an effective technique to find out inter-database patterns and there is no need for a central site.The existing work using this technique is very less and their efficiency is also poor. The information exchange is done in raw formats and hence the data transfer is more and cause for a bottleneck in the network. The challenges that we face using this method are absence of a system framework capable of mining 3 types of pattern, data

exchange method to pass information between the distributed sites, type of mining procedure that should be used and pattern pruning mechanism. This paper proposes a method to exchange pattern information with other distributed sites effectivelywhich is both fast and requires less data transfer also helps to perform pattern pruning between the sites easily.

## III. PROPOSED WORK

In collaborative pattern mining[15], pattern discovery is performed in a distributed manner with each distributed site carrying out pattern pruning in collaboration with its peers by using some pattern pruning mechanism. The pattern pruning mechanism this paper proposes is based on a data structure called Bloom Filter[10,11,12]. In this paper, frequent pattern mining[2] is used to perform mining and the using the patterns that are obtained at a distributed site pattern pruning is performed at other sites with the help of Bloom Filter.

Bloom Filter is a probabilistic space efficient data structure that is capable of finding out whether an element is present in a set or not. It contains a m-bit array and k Hash Functions $H_1(.)$, $H_2(.)$ ... $H_k(.)$.The strength of a BF test whether a given element is a member of a set in a very effective way[7,14]. Initially each element in the m-bit array is initialized as 0. Each hash function will take an element as input and the output will be a number between 0 and m indicating the position in the m-bit bit array and the value in the corresponding bit position will be changed to 1.

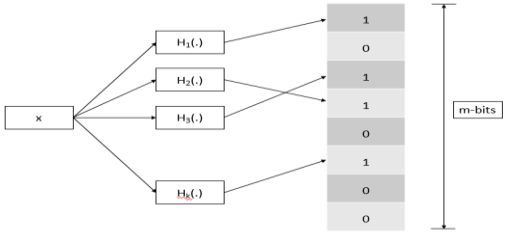



**Figure 1 : Bloom Filter Architecture**

If we want to add an element 'x' to the Bloom Filter, we use the k-Hash functions to hash the element and find k-positions in the array to be made as '1'. In order to search for a particular element say 'y' in the bloom filter, we again uses the k-hash functions to find k positions and checks in the m-bit array whether these k positions have value '1'. If all the k positions have the value '1' then we can say that the element could be present in the set. If at least one of the k position has value as '0', we can conclude that the element 'y' is not present in the set. False positives are possible but false negatives will not occur, hence we optimize the size of the bloom filter and the number of hash functions to reduce the false positive rate according to our application.

Another feature of Bloom filters is that there is a clear tradeoff between the size of the filter and the rate of false positives. After the insertion of *n* keys into a boom filter of size *m* making use of *k* hash functions, the probability of a bit being still 0 is :

$$p_0 = \left(1-\frac{1}{m}\right)^{kn} \approx 1-e^{-\frac{kn}{m}}. \qquad (1)$$

We assume that we use hash function that spread elements evenly across the space {1-m}.
The probability of a false positive is:

$$p_{err} = (1-p_0)^k = \left(1-\left(1-\frac{1}{m}\right)^{kn}\right)^k \approx \left(1-e^{-\frac{kn}{m}}\right)^k \quad (2)$$

In (2)$p_{err}$ is minimized for $k = \frac{m}{n}\ln 2$ hash functions. In practice however, only a small number of hash functions are used. The reason is that the computational overhead of each hash additional function is constant while the incremental benefit of adding a new hash function decreases after a certain threshold (see **Error! Reference source not found.**2).

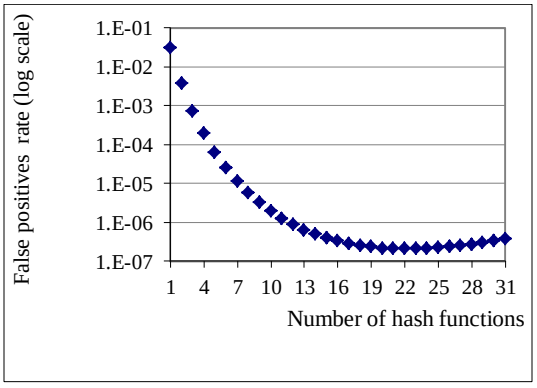



**Figure 2: False positive rate as a function of the number of hash functions used. The size of the Bloom filter is 32 bits per entry (m/n=32). In this case using 22 hash functions minimizes the false positive rate. Note however that adding a hash function does**

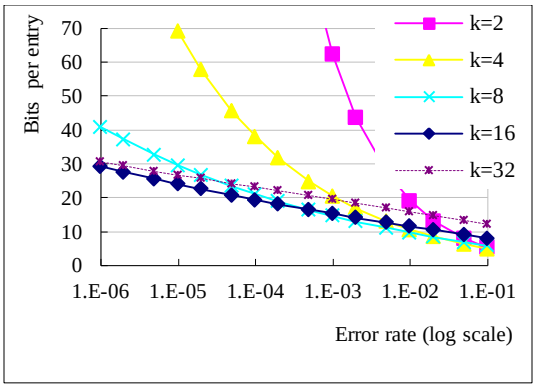



**Figure 3: Size of Bloom filter (bits/entry) as a function of the error rate desired. Different lines represent different numbers of hash keys used. Note that, for the error rates considered, using 32 keys does not bring significant benefits over using on**

(2) is the base formula for engineering Bloom filters. It allows, for example, computing minimal memory requirements (filter size) and number of hash functions given the maximum acceptable false positives rate and number of elements in the set (as we detail in **Error! Reference source not found.**3).

$$\frac{m}{n} = \frac{-k}{\ln\left(1 - e^{\frac{\ln p_{err}}{k}}\right)} \text{ (bits per entry)} \quad (3)$$

The user will give a query according to which patterns are to discover. The client machine will distribute the query to each distributed site. And at each site the pattern mining takes place.

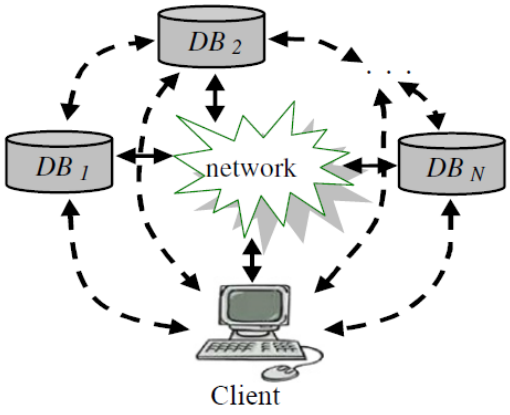



**Figure 3: Mining activities are carried out in distributed sites with bloom filter exchanged between sites for cross site pruning**

We make use of apriori[2] algorithm to discover the frequent patterns at the distributed site. At a given site the apriori algorithm will generate patterns with the help of data present at that distributed site, after the fining out the length-l patterns, a bloom filter will be created using those patterns and it will be broadcasted to other sites. At the receiving site, using the bloom filter pattern pruning will be carried out eliminating un-wanted patterns. In this way the sites will collaborative with one another in pruning and a finally a global pattern will be discovered at each site. Along with the global patterns relevant inter database patterns are also generated at that site.

*1. At site $S_i$, use the Apriori mining approach to generate a complete set of length-l patterns.*
*2. Using the frequent length-l patterns in site $Site_i$ construct a bloom filter ($BF_i$-l), and broadcast $BF_i$-l to other distributed sites.*
*3. After a distributed site $S_i$ receives the bloom filters $BF_j$-l from other sites, it can query $BF_j$-l and prune out length-l patterns in $S_i$ and then grow length-(l+1) patterns.*
*4. Set l←l+1 and repeat Steps 2 to 4 until no more frequent patterns can be discovered from any sites*

The patterns that are discovered at the distributed sites will be send to the client site where the user can view the patterns. This method is asynchronous and hence each distributed site can work without synchronizing with other sites.Notice that bloom filter cannot encode support values of the patterns.

## IV. CONCLUSION

In this paper, our aim is to mine patterns in a distributive environment and to discover Local, Global and Inter-database patterns. We have seen that the current researches focuses only on mining L- and G- Patterns. There is no effective method to mine all three types of patterns. Using the bloom filter technique we can effectively mine three types of patterns (L, G and I-patterns) in a collaborative manner with minimized data transfer between the sites. This mining technique has very little privacy concerns and requires low computational costs and memory consumption. Experimental comparisons demonstrated that this outperforms other simple methods. The problem addressed in this paper mainly focuses on frequent item-set mining. However, this method can be extended to handle other types of patterns as well such as constrained frequent item-sets, closed frequent patterns and sequential patterns.

## V. REFERENCES

[1] R. Agrawal, J.C. Shafer, Parallel mining of association rules, IEEE Transactions onKnowledge and Data Engineering 8 (6) (December 1996) 962–969.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, Proc. of VLDBConference, 1994.

[3] F. Provost, Distributed data mining: scaling up and beyond. In Kargupta, H.,Chan, P., eds.: Advances in Distributed and Parallel Knowledge Discovery, MIT/AAAI Press, 2000.

[4] S. Zhang, C. Zhang, X. Wu, Knowledge discovery in multiple database, Springer,2004.

[5]M. Ashrafi, D. Taniar, K. Smith, ODAM: an optimized distributed association rulemining algorithm, IEEE Distributed Systems Online 5 (3) (2004).

[6]H. Kargupta, B.H. Park, D. Hershberger, E. Johnson, Collective data mining: a newperspective toward distributed data mining, Advances in Distributed and ParallelKnowledge Discovery, MIT/AAAI Press, Cambridge, MA, 1999.

[7] D. Cheung, V. Ng, A. Fu, Y. Fu, Efficient mining of association rules in distributeddatabases, IEEE Trans. on Knowledge and Data Engineering 8 (1996).

[8] X. Wu, S. Zhang, Synthesizing high-frequency rules from different datasources, IEEE Transactions on Knowledge and Data Engineering 15 (2)(2003) 353–367.

[9] X. Zhu, R. Jin, Y. Breitbart, G. Agrawal, MMIS-07, 08: mining multiple informationsources workshop report, ACM SIGKDD Explorations 10 (2) (2008) 61–65.

[10] A. Border, M. Mitzenmacher, Network applications of bloom filters: a survey, Proc.of the 40th Annual Allerton Conf. on Communication, Control, and Computing,Urbana-Champaign, Illinois, 2002, pp. 636–646.

[11] B. Chazelle, J. Kilian, R. Rubinfeld, A. Tal, The Bloomier filter: an efficient datastructure for static support lookup tables, Proc. of the 5th ACM-SIAM Symposiumon Discrete Algorithms, 2004, pp. 30–39.

[12] S. Cohen, Y. Matias, Spectral bloom filters, Proc. of SIGMOD Conference, 2003,pp. 241–252.

[13]S. Zhang, M. Zaki, Mining multiple data sources: local pattern analysis, DataMining and Knowledge Discovery 12 (2–3) (2006) 121–125.

[14] X. Gong, W. Qian, Y. Yan, A. Zhou, Bloom filter-based XML packets filtering formillions of path queries, Proc. of ICDE Conference, 2005, pp. 890–901.

[15]Xingquan Zhu, Bin Li, Xindong Wu, Dan He, Chengqi Zhang, CLAP: Collaborative pattern mining for distributed information systems, Decision Support Systems, Volume 52, Issue 1, December 2011, pp. 40–51

# Cryptography and computer Security

**Prof.Sofeen Sayyed.**
Arihant College of Arts,Commerce and Science,Camp-Pune-01
Sayyed_sofeen@yahoomail.com
9923199033.

**Abstract:**

**To expose students to current and classical computer security literature and to,**

- **Give students an understanding of what it means to do research in computer science and specifically in computer security**
- **Teach students to critically evaluate research papers**
- **Tackle some challenging projects and write paper suitable for publication.**

Keywords:

Cipher text, encryption, decryption, public key, protocols, etc.

## Introduction:

The art of protecting information by transforming it (*encrypting* it) into an unreadable format, called cipher text. Only those who possess a secret *key* can decipher (or *decrypt*) the message into plain text. Encrypted messages can sometimes be broken by cryptanalysis, also called *code breaking*, although modern cryptography techniques are virtually unbreakable.

As the Internet and other forms of electronic communication become more prevalent, electronic security is becoming increasingly important. Cryptography is used to protect e-mail messages, credit card information, and corporate data. One of the most popular cryptography systems used on the Internet is Privacy because it's effective and free.

Cryptography systems can be broadly classified into symmetric-key systems that use a single key that both the sender and recipient have, and key systems that use two keys, a public key known to everyone and a private key that only the recipient of messages uses.

Modern cryptography concerns itself with the following four objectives:

1) **Confidentiality**: The information cannot be understood by anyone for whom it was unintended.
2) **Integrity**: The information cannot be altered in storage or transit between sender and intended receiver without the alteration being detected.
3) **Non-repudiation:** The creator/sender of the information cannot deny at a later stage his or her intentions in the creation or transmission of the information.
4) **Authentication**: The sender and receiver can confirm each other's identity and the origin/destination of the information.

Procedures and protocols that meet some or all of the above criteria are known as cryptosystems. Cryptosystems are often thought to refer only to mathematical procedures and computer programs; however, they also include the regulation of human behavior, such as choosing hard-to-guess passwords, logging off unused systems, and not discussing sensitive procedures with outsiders.

## History:

The word is derived from the Greek word *kryptos*, meaning hidden. The origin of cryptography is usually dated from about 2000 BC, with the Egyptian practice of hieroglyphics. These consisted of complex pictograms, the full meaning of which was only known to an elite few. The first known use of a modern cipher was by Julius Caesar (100 BC to 44 BC), who did not trust his messengers when communicating with his governors and officers. For this reason, he created a system in which each character in his messages was replaced by a character three positions ahead of it in the Roman alphabet.

In recent times, cryptography has turned into a battleground of some of the world's best mathematicians and computer scientists. The ability to securely store and

transfer sensitive information has proved a critical factor in success in war and business.

A glimpse of Classical cryptography:

Ever since writing was invented, people have been interested not only in using it to communicate with their partners, but also in trying to conceal the content of their message from their adversaries. This leads to cryptography (or cryptology), the science of secret communication. The basic situation is that one party, say King Arthur, wants to send a message to King Bela. There is, however, a danger that the evil Caesar Caligula intercepts the message and learns things that he is not supposed to know about. The message, understandable even for Caligula, is called the plain text. To protect its content, King Arthur encrypts his message. When King Bela receives it, he must decrypt it in order to be able to read it. For the Kings to be able to encrypt and decrypt the message, they must know something that the Caesar does not know: this information is the key. Many cryptosystems have been used in history; most of them, in fact, turn out to be insecure, especially if the adversary can use powerful computing tools to break it. Perhaps the simplest method is substitution code: we replace each letter of the alphabet by another letter. The key is the table that contains for each letter the letter to be sub-situated for it. While a message encrypted this way looks totally scrambled, substitution codes are in fact easy to break.

Number Theory in Cryptography Terminology:

Two parties Alice and Bob want to communicate securely such that a third party Eve who intercepts messages cannot learn the content of the messages.

Alice and Bob share a secret. Only they know a secret key K that is used to encrypt and decrypt messages. Given a message M, Alice encodes it (possibly with padding) into m, and then sends the cipher text encrypt (m; K) to Bob. Then Bob uses K to decrypt it and obtains decrypt encrypt (m; K) =m.

How to verify a password—without learning it?

In a bank, a cash machine works by name and password. This system is safe as long as the password is kept in secret. But there is one week point in security: the computer of the bank must store the password, and the administrator of this computer may learn it and later misuse it. Complexity theory provides a scheme where the bank can verify that the customer does indeed know the password—without storing the password itself!

Suppose that the password is a 100-digit prime number p (this is, of course, too long for everyday use, but it illustrates the idea best). When the customer chooses the password, he chooses another prime q with 101 digits, forms the product N =pq of the two primes, and tells the bank the number N. When the teller is used, the customer tells his name and the password p. The computer of the bank checks whether or not p is a divisor of N; if so, it accepts p as a proper password. The division of a 200 digit number by a 100 digit number is a trivial task for a computer. Let us assume that the system administrator learns the number N stored along with the files of our customer. To use this in order to impersonate the customer, he has to find a 100-digit number that is a divisor of N; but this is essentially the same problem as finding the prime factorization of N , and this is hopelessly difficult. So—even though all the necessary information is contained in the number N —the computational complexity of the factoring problem protects the password of the customer!

Public key cryptography:

RSA: an example of a Public Key Cryptosystem

Named after Rivest, Shamir, Adelman (1976 at MIT).

Cryptographic systems used in real life are more complex than those described in the previous section—but they are based on similar principles. In this section we sketch the math behind the most commonly used system, the RSA code.

Let Alice generate two 100-digit prime numbers, p and q and computes their product m = pq. Then she generates two 200-digit numbers d and e such that ( p−1)( q− 1) is a divisor (ed – 1).

The numbers m and e she publishes on her web site, or in the phone book, but the prime factors p

And q and the number d remain her closely guarded secrets. The number d is called her private key, and the number e, her public key (the number p and q she may even forget—they will not be needed to operate the system, just to set it up. Suppose first that Bob wants to send a message to Alice. He writes the message as a number x (we have seen before how to do so). This number x must be a non-negative integer less than m (if the message is longer, he can just break it up into smaller chunks).The next step is the trickiest: Bob computes the remainder of x e modulo m. Since both x and e are huge integers (200 digits), the number x ,d has more than 200digits – we could not even write it down, let alone compute it!

RSA involves a public key and private key. The public key can be known to everyone, it is used to encrypt messages. Messages encrypted using the public key can only be decrypted with the private key. The keys for the RSA algorithm are generated the following way:

1. Choose two different large random prime numbers $p$ and $q$
2. Calculate $n = pq$
    - $n$ is the modulus for the public key and the private keys

3. Calculate the totient: $\phi(n) = (p-1)(q-1)$.
4. Choose an integer $e$ such that $1 < e < \phi(n)$, and $e$ is coprime to $\phi(n)$ such that $e$ and $\phi(n)$ share no factors other than 1; gcd($e$, $\phi(n)$) = 1.
5. Compute $d$ to satisfy the congruence relation $de \equiv 1 \pmod{\phi(n)}$ **ie:** $de = 1 + k\phi(n)$ for some integer $k$.
   - $d$ is kept as the private key exponent

The **public key** is made of the modulus $n$ and the public (or encryption) exponent $e$.
The **private key** is made of the modulus $n$ and the private (or decryption) exponent $d$ which must be kept secret.

- For efficiency a different form of the **private key** can be stored:
  - $p$ and $q$: the primes from the key generation,
  - $d \mod (p-1)$ and $d \mod (q-1)$.
  - $q^{-1} \mod (p)$.

Encrypting messages

Alice gives her public key ($n$ & $e$) to Bob and keeps her private key secret. Bob wants to send message **M** to Alice.

First he turns **M** into a number smaller than $n$ by using an agreed-upon reversible protocol known as a padding scheme. He then computes the ciphertext $c$ corresponding to:

$$c = m^e \mod n$$

This can be done quickly using the method of exponentiation by squaring. Bob then sends $c$ to Alice.

**Decrypting messages**

Alice can recover $m$ from $c$ by using her private key $d$ in the following procedure:

$$m = c^d \mod n$$

Given $m$, she can recover the original message **M**.

The decryption procedure works because first

$$c^d \equiv (m^e)^d \equiv m^{ed} \pmod{n}.$$

Now, since

$$ed \equiv 1 \pmod{p-1}$$ and
$$ed \equiv 1 \pmod{q-1}$$

Fermat's little theorem yields

$$m^{ed} \equiv m \pmod{p}$$ and
$$m^{ed} \equiv m \pmod{q}.$$

Since $p$ and $q$ are distinct prime numbers, applying the Chinese remainder theorem to these two congruences yields

$$m^{ed} \equiv m \pmod{pq}.$$

Thus,

$$c^d \equiv m \pmod{n}.$$

**A working example**

Here is an example of RSA encryption and decryption. The parameters used here are artificially small, but you can also use OpenSSL to generate and examine a real keypair.

1. Choose two random prime numbers
2. : $p = 61$ and $q = 53$
3. Compute $n = pq$
4. : $n = 61 * 53 = 3233$
5. Compute the totient $\phi(n) = (p-1)(q-1)$
6. : $\phi(n) = (61-1)(53-1) = 3120$
7. Choose $e > 1$ coprime to 3120
8. : $e = 17$
9. Choose $d$ to satisfy $de \equiv 1 \pmod{\phi(n)}$
10. : $d = 2753$
11. :17 * 2753 = 46801 = 1 + 15 * 3120

The **public key** is ($n = 3233$, $e = 17$). For a padded message $m$ the encryption function is:

$$c = m^e \mod n = m^{17} \mod 3233$$

.

The **private key** is ($n = 3233$, $d = 2753$). The decryption function is:

$$m = c^d \mod n = c^{2753} \mod 32$$

.

For example, to encrypt $m = 123$, we calculate

$$c = 123^{17} \mod 3233 = 855$$

To decrypt $c = 855$, we calculate

$$m = 855^{2753} \mod 3233 = 123$$

..

**What is Computer Security?**

Computer Security is the protection of computing systems and the data that they store or access.It is important to

- Enabling people to carry out their jobs, education, and research
- Supporting critical business process
- Protecting personal and sensitive information

Example: The lock on the door is the 10%. You remembering to lock the lock, checking to see if the door is closed, ensuring others do not prop the door open, keeping control of the keys, etc. is the 90%. You need both parts for effective security.

**Security Objectives**

- Learn **"good computing security practices."**
- Incorporate these practices into your everyday routine. Encourage others to do so as well.
- Report anything unusual - Notify the appropriate contacts if you become aware of a suspected security incident.

The Internet can be a hazardous place:
How many attacks to computers on campus do you think take place every day?

- Thousands of attacks **per minute** bombard our campus network.
- An unprotected computer can become infected or compromised **within a few seconds** after it is connected to the network.

Many cyber security threats are largely avoidable.

**Some key steps that everyone can take include:**

- Use good, cryptic passwords that can't be easily guessed - and keep your passwords secret
- Make sure your computer's operating system and applications are protected with all necessary security "patches" and updates
- Make sure your computer is protected with up-to-date antivirus and anti-spyware software
- Don't click on unknown or unsolicited links or attachments, and don't download unknown files or programs onto your computer
- Remember that information and passwords sent via standard, unencrypted wireless are especially easy for hackers to intercept
    - To help reduce the risk, look for "https" in the URL before you enter any sensitive information or a password. (The "***s***" stands for "secure".)
    - Also avoid standard, unencrypted e-mail and unencrypted Instant Messaging (IM) if you're concerned about privacy

CONCLUSION:

Computer and network security, or cybersecurity, are critical issues. But merely protecting the systems that hold data about citizens, corporations, and government agencies it is not enough. The infrastructure of networks, routers, domain name servers, and switches that glue these systems together must not fail, or computers will no longer be able to communicate accurately or reliably. Given the magnitude of securing cyberspace, a reflection on what we are trying to do seems in order. Several questions arise, such as what exactly the infrastructure is, what threats it must be secured against, and how protection can be provided on a cost-effective basis.

REFERENCES:

1)Journal of Discrete Mathematical Sciences and Cryptography --Taylor & Francis

2) Discrete Mathematics-Lecture Notes, Yale University, Spring 1999 L. Lov´asz and K. Vesztergomb

3)Applied Cryptography—Bruce S.
4)Introduction To Cryptography Security- S.k. Kataria & Sons

# CIPHER TEXT CONVERSION WITH ENCRYPTED POLYNOMIAL EQUATION

**Dr. Kavita Khobragade[1], Shounak Joshi[2]**
[1]Associate Professor, Computer Science Department. Fergusson College, Pune. kavitanand@rediffmail.com
[2] T.Y. Student, Fergusson College, Pune.
shounakjoshi16@gmail.com

**Abstract:**

**The modern era is dominated by paperless messages and transactions. Due to this there is a great need of interchanging of data through internet. For this the data has to be secured. Cryptography is a technique used for converting simple message into non-readable form. The cryptography can be achieved by using Substitution and transposition cipher technique. The most frequently used technique is "Caesar Cipher". In this research paper we are discussing the method used to convert message text into encrypted polynomial equation. In our algorithm, the basic technique (Caesar Cipher) is used to replace the simple original text into encrypted format and then convert the text into a random polynomial equation by considering alphabetic serial order. As an output, each string in the plain text will get converted into a random polynomial equation and this particular equation will be sent to the recipient. Our technique is more secure as there is no need send any key to decrypt the text. Therefore our algorithm is robust. So that encrypted data may not be hacked easily by the intruders and security is preserved.**

***Keywords: Cryptography, encryption, decryption, Caesar Cipher, Polynomial equation.***

## I. Introduction

In today's computing world, everyone is connected with the web of computers, servers, mobiles and many computing devices. All the work we need to do like, sharing data, banking transaction, chatting, Shopping etc. To achieve these goals we may need to share some private information which should not be shared with any 3rd person other than sender and receiver. But while doing this data may get hacked by intruders (i.e. 3rd person)[1, 2, 4]. So, to keep the data secure we convert our messages or data in some non-readable form. For this we use the technique called Cryptography.

The word Cryptography is derived from the Greek word Kryptos which means "hidden or secret". The cryptography is art and science of converting the plain text into some non-readable form. In the language of cryptography, the message one intends to send is called the plaintext while the message that is actually sent is called the cipher text. So, we need to have some technique to create the cipher text from plain text. For achieving cryptography we already have some existing techniques. One the easiest and widely used technique is "Substitution Cipher"[2].

In Substitution Cipher method, each letter or group of letters are replaced by another letter or group of letters to disguise the plain text. One of well know substitution cipher technique is "Caesar Cipher"[2, 4, 5, 6].

In Caesar Cipher every letter will replace with the third character from that alphabet[3]. For example A will be replaced by D, B will be replaced with E etc. The following table gives the brief idea about this:

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| D | E | F | G | H | I | J | K | L | M |
| K | L | M | N | O | P | Q | R | S | T |
| N | O | P | Q | R | S | T | U | V | W |
| U | V | W | X | Y | Z | | | | |
| X | Y | Z | A | B | C | | | | |

**Example**:
1. Suppose the word is "HELLO".
Plain Text: HELLO
H->K, E->H, L->O, L->O, O->R
Cipher Text: KHOOR

But the major drawback of Caesar Cipher is easy decryption. The algorithm is very easy since, we are replacing every character with the next 3rd character from its position in alphabetic order. So it can be hacked easily by any intruder. To overcome the drawback of Caesar Cipher a new algorithm was developed known as "Modified Caesar Cipher"[3].

In Modified Caesar Cipher method of cryptography we can replace the alphabets in any

fashion. For example A can be replaced with any other character in English alphabet set (from B through Z).

Thus, for each alphabet we have 25 possibilities of replacement [2].
For example, suppose we are replacing letters as follows:

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| Q | W | E | R | T | Y | U | I | O | P |
| K | L | M | N | O | P | Q | R | S | T |
| A | S | D | F | G | H | J | K | L | Z |
| U | V | W | X | Y | Z | | | | |
| X | C | V | B | N | M | | | | |

**Example**:
Let we consider the word: HELLO to be sent as message.
Plain Text: HELLO
H->I, E->T, L->D,L->D, O->G
Cipher Text: ITDDG.

We will take another example with another key (i.e. using another alphabet arrangement).
Suppose the key will be,

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| M | N | B | V | C | X | Z | L | K | J |
| K | L | M | N | O | P | Q | R | S | T |
| H | G | F | D | S | A | P | O | I | U |
| U | V | W | X | Y | Z | | | | |
| Y | T | R | E | W | Q | | | | |

Let we consider, the message as, HELLO only,
Plain Text: HELLO
H->L, E->C, L->G, L->G, O->S
Cipher Text: LCGGS

The drawback of this algorithm is we need to send the key along with the message. So the intruder can decrypt the message easily by using the key. Therefore we need to develop a technique in which we will send the message with key but still the data can be sent securely.

## II. Literature Review and Objectives

P.Subhasri et.al used reverse Ceaser cipher algorithm for cloud computing [9]. Imran et. Al, used the Ceaser cipher method. They used three methods in which each letter in the plaintext is replaced by some fixed number of position down the alphabet. The result is a data which is encrypted and be decrypted to its readable form [1].
The objectives of this research are

1. To develop a new encryption technique that will strengthen the security of the basic Caesar Cipher encryption.
2. To develop a technique in which no extra things will need to decrypt the message.

## III. Proposed Algorithm For Encryption

The techniques which are developed from the basic substitution Cipher, use some key to decrypt the text again. But the proposed system will not need any extra key to send with message and still it will provide better security. We consider a standard table for this.
**Step 1**: Accept the plain text.
**Step 2:** Encrypt the text using Caesar Cipher algorithm.
**Step 3**: To encrypt the text into polynomial form, referring the following table:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A | B | C | D | E | F | G | H | I |
| 2 | J | K | L | M | N | O | P | Q | R |
| 3 | S | T | U | V | W | X | Y | Z | |

A. Process the text character bye character; if the character is an alphabet then the variable of equation is ‘x’ and if it is a digit then variable is ‘y’.
B. To represent the character in equation form then we will print the row number as multiplier and column number as power of the variable.
C. If the character is any digit then the power represents the digit.
D. Append the representations of each character to form a polynomial equation of the word.

**Step 4**: Repeat step 3 for each word in the message till the end of text.
**Step 5**: Send the encrypted text.
**Step 6**: Stop

**Examples of Encryption**

1. Plain Text: HELLO
   Step 1: Encrypt the plain text by using Caesar Cipher.
   H->K, E->H, L->O, L->O, O->R
   Cipher Text: KHOOR
   Step 2: convert this encrypted text into equation,
   K->2x^2, H->x^8, O->2x^6, O->2x^6, R->2x^9
   Final Cipher Text (After getting equation): 2x^2+x^8+2x^6+2x^6+2x^9
2. Plan Text: INDIA16

Step 1: Encrypt the plain text by using Caesar Cipher.
I->L, N->Q, D->G, I->L, A->D, 1, 6
Cipher Text: LQGLD16
Step 2: convert this encrypted text into equation,
L->2x^3, Q->2x^8, G->x^7, L->2x^3, D->x^4, 1->y^1, 6->y^6
Final Cipher Text: 2x^3+2x^8+x^7+2x^3+x^4+y^1+y^6

3. Plan Text: ZING20
Step 1: Encrypt the plain text by using Caesar Cipher.
Z->C, I->L, N->Q, G->J, 2, 0
Cipher Text: CLQJ20
Step 2: convert this encrypted text into equation,
C->x^3, L->2x^3, Q->2x^8, J->2x, 2->y^2, 0->y^0 = 1
Final Cipher Text: x^3+2x^3+2x^8+2x+y^2+1

## IV. Proposed Algorithm For Decryption:

**Step 1**: Take each equation from received text in consideration.
**Step 2**: If variable is 'x' then,
Consider the multiplier as row number and power as column number and represent the alphabet from the column and row numbers.
Else if variable is 'y' then,
Consider the power and the power represents the digit itself. Decrypt the digit.
**Step 3**: Decode every equation by using step 2.
**Step 4**: Decode the text we got in step 3 by reversing the Caesar Cipher encryption (i.e. character - 3).
**Step 5**: Stop

**Examples of Decryption:**

1.
Step 1: Cipher Text: x^3+2x^3+2x^8+2x+y^2+1
Step 2: x^3->C, 2x^3->L, 2x^8->Q, 2x->J, y^2->2, 1=y^0-> 0
Cipher Text after 2nd step: CLQJ20
Step 3: Decrypt this this Cipher text by using Caesar cipher.
C->Z, L->I, Q->N, J->G, 2, 0
Final Decrypted Text (plain text): ZING20

2.
Step 1: Cipher Text: 2x^2+x^8+2x^6+2x^6+2x^9
Step 2: 2x^2->k,x^8->H, 2x^6->O, 2x^6->O, 2x^9->R
Cipher Text after 2nd step: KHOOR
Step 3: Decrypt this this Cipher text by using Caesar cipher.
K->H, H->E, O->L, O->L, R->O
Final Decrypted Text (plain text): HELLO

## V. Conclusion

Caesar Cipher method is simple to encode and decode. That's why the transmitted data can be hacked easily. In modified cipher method we need to send a key to decode the data so by getting the key, one may get successful in decrypting the data. Therefore by generating and sending polynomial expression for the message solves the problem of hacking. The data can be sent securely. In this paper we tried to make the encryption process more secure by applying Caesar cipher method followed by our algorithm to form the equation of words. This combination provides more secure and dependable cipher.

**References:**

[1] Enas Ismael Imran and Farah abdulameerabdulkareem "Enhancement Caesar Cipher for Better security", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 01-05.
[2] Dr. Kavita Khobragade and Rahul Patil, Computer Networks – II, Nirali Prakashan, 1st Edition, December 2015.
[3] https://en.wikipedia.org/wiki/Caesar_cipher
[4] Atul Kahate , *Cryptography and Network Security*, 2nd edition, McGraw-Hill, 2009.
[5] William Stalling, "Network Security Essentials (Applications and Standards)", Pearson Education,2004.
[6] Stallings W, *Cryptography and Network Security*, 2nd edition, Prentice Hall, 1999.
[8] William Stallings, *Cryptography and Network Security*, 3rd edition, Pearson Education, 2003.
[9] P.Subhasri, Dr. A. Padmapriya, " Implementation of Reverse Caesar Cipher Algorithm for Cloud Computing", *International Journal for Advance Research in Engineering and Technology*, Vol. 1, Issue VI,pp: 43-46, July 2013 [ISSN 2320-6802].

# Scope of Visual Basic for Applications in Data Analysis and Automation

**Mrs. Leena Sanu**[1]
Dept. of Computer Science
Christ College
Email:leena31_thomas@yahoo.com

**Mr.Sanu Poulose**[2]
Data Analyst
HR Analytics and Reporting
Vodafone Shared Services

**Abstract:**

**VBA, which stands for *Visual Basic for Applications,* is a programming language developed by Microsoft. Excel, along with the other members of Microsoft Office 2003, includes the VBA language. MS Excel an application which is always available in any organization or institution. It is very effectively used in data collection & analysis in all domains. Within organizations, teams like Reporting or Business intelligence or Analytics who create reports out of source files have to face a lot of challenges with data. The Objective of this paper is to discuss how these challenges can be overcomed by automation, using VBA which proves to be an effective tool in terms of cost and time.**

**Keywords:** *VBA, objects,* Sheets, Ranges, Workbook, Module, Class Module, User Form, Excel Functions

**Introduction:** VBA is the tool that people use to develop programs that control Excel. You're probably aware that people use Excel for thousands of different tasks. Within organizations, teams like Reporting or Business intelligence or Analytics who create reports out of Source files have challenges with data like

a. Collating data from multiple files into one file
b. Breaking data into different files from a single file.
c. Cleansing of Source files
d. Formatting of data
e. Connecting with databases like MS Access
f. Data Visualization dynamically using Graphs & Charts
g. Sending Report attachment to multiple recipients via Outlook and many more.

**Overview:** Visual Basic for Applications (VBA) uses objects to control Microsoft Excel. Whether you are working with the Excel application (Application object), workbooks (Workbook object), or individual cells (Cell object), you do everything by manipulating an object. All of the objects either contain other objects or are part of a larger object. For example, the Workbook object contains Worksheets objects, which in turn contain Cell objects.

When working with an object, either a property is being set or read, or the object is told

To perform an action, called a method. If you wanted to select a worksheet named Sheet2 in the active workbook, you would use the Select method of the Worksheet object,

Eg. Worksheets("Sheet2").Select. Visual Basic for Applications (VBA) sees Excel as a series of objects that have attributes to describe them and actions those objects as mentioned below.

a. Properties: Describes some aspect of object eg. Name
b. Methods: Performs an action
c. Events: eg. A workbook is opened.
d. Collections: Set of related objects like Range, Sheets, Workbooks, and Shapes etc.

**Scope of VBA in Data Analysis & Automation:** There are various ways by which your activities will be done easily with the help of macros used in VBA. Tasks like:

*Automating repetitive operations*: If you need to perform the same action on, say, 12 different Excel workbooks, you can record a macro while you perform the task on the first workbook and then let the macro repeat your action on the other workbooks. *Creating a custom command*: Do you often issue the same sequence of Excel menu commands? If so, save

yourself a few seconds by developing a macro that combines these commands into a single custom command, which you can execute with a single keystroke or button click We will now discuss the various applications of VBA using the VBA Syntax.

1. **Auto fill of Data (Jan to Dec)**

***Code***

```
Sub AutofillListJanToDec()
With Range ("E1")
.Value = "Jan"
.AutoFill .Resize (12)
End With
End Sub
```

**Explanation**

You can fill any range in Excel with sequential data like Months as shown in the example. In this example 12 cells are filled with month names from Jan to Dec.

2. **Creating a date series (Entering Date series for dec)**

***Code***

```
Sub FillDataSeriesMonths()
    With Range ("C1")
    .Value = DateSerial(2012,
    12, 1)
    .Resize(31).DataSeries
    xlColumns,
    xlChronological, xlDay, 1
    End With
End Sub
```

**Explanation**

You can fill any range in Excel with series data like dates as shown in the example.In this example 31 cells are filled with dates from 1st Dec 2012 to 31st Dec 2012.

3. **Select Case Statements & using For Condition**

***Code***

```
 For i = 1 To 7
Select Case i
    Case 1 To 4
    j = j + 1
    k = k - 2
Case Else
    j = j - 1
    k = k + 2
End Select
    Cells(i,
j).Resize(1,k).Interior.Color = 0
Next
```

**Explanation**

This is an example of how Select Case logic is applied on excel data to get the desired result

4. **Applying Advanced Filter & Formulae**

***Code***

```
data.AdvancedFilter xlFilterCopy, ,
copyto, True
.Offset(, 1).Value = "Total Sales"
With .CurrentRegion
With .Offset(1,
1).Resize(.Rows.Count - 1, 1)
.Value = "=SUMIF(" & CriAddress
& _
",rc[-1]," & SumAddress & ")"
.Value = .Value
End With
End With
```

**Explanation**

You can apply filter on huge data on multiple criteria's and paste it in separate workbook and save it in the respective folder. This facilitates in saving a lot of time.

eg. In HR Domain if I have all the employee records for 100 countries in one sheet and the number of records are 50,000. With a click of a button using VBA you can create 100 folders and renaming it with respective country name and filtering down the data for the respective country and saving the respective country details in that country folder.

5. **Creating Customized Functions**

***Code***

```
 Function GetLastRow(rng As Range)
As Long
Dim Lastcell As Range

With rng
Set Lastcell = _
.Find("*", .Resize(1, 1), , , xlByRows,
xlPrevious)
```

```
End With
If Lastcell Is Nothing Then
GetLastRow = 1
Else
GetLastRow = Lastcell.Row
End If
End Function
```

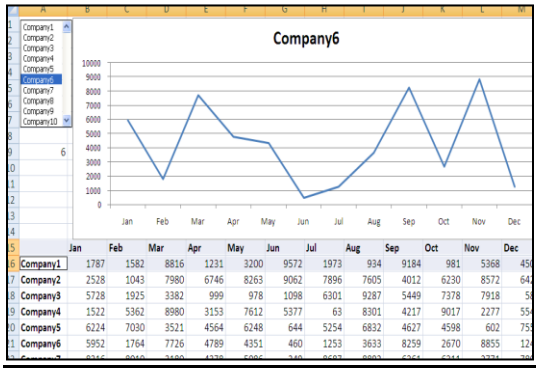
**Explanation**

We can create a customized function as per our requirement and use it in excel.

Example shown is a customized function **GetLastRow** which gives the last row number in the selected range.

If you go to excel and in any cell put= **GetLastRow** & Range, you will get the last row number in the selected range.

5. **Dynamic Dashboards using Graphs:**



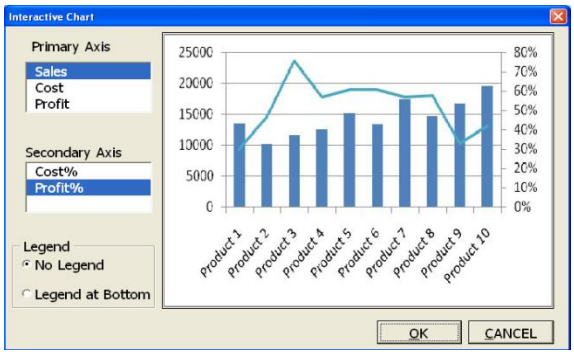

***Code***

```
Sub CompanyDashboard()
Dim sht As Worksheet
Dim cho As ChartObject
Dim cht As Chart
Dim rng As Range

Set sht = Sheets("Company Dashboard")
Set cho = sht.ChartObjects("Chart 3")
Set cht = cho.Chart

With Range("Heading")
Set rng = Union(.Offset(0, 0), .Offset(Range("ChtLink").Value))
End With

cht.SetSourceData rng
End Sub
```

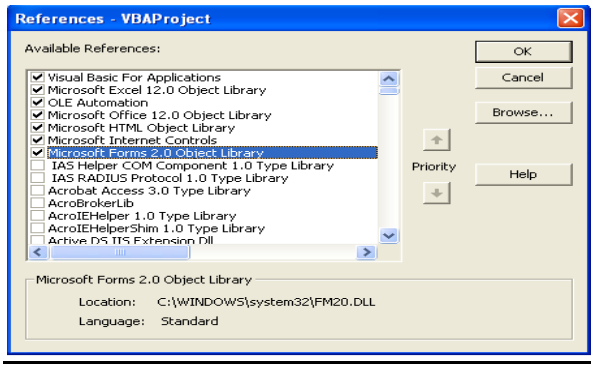
**Explanation**

We can create dynamic dashboards like these examples where if the user clicks on any Company in the List Box it will dynamically update the respective company graph.Thus data visualization in reports can be achieved using VBA

6. **Dynamically display chart in User form:** Multiple Graphs can be generated giving user to select multiple options as shown in the below example.



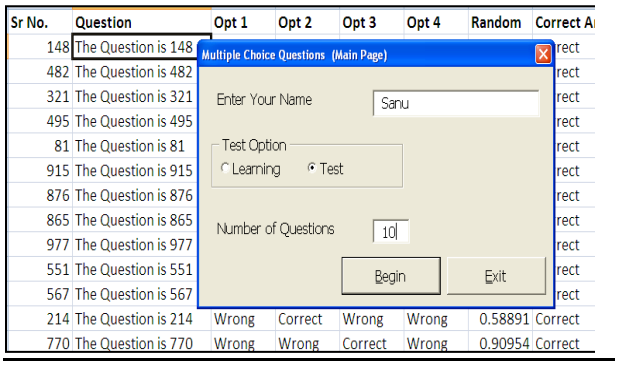

7. **Import Table Data from Html Page to Excel:**

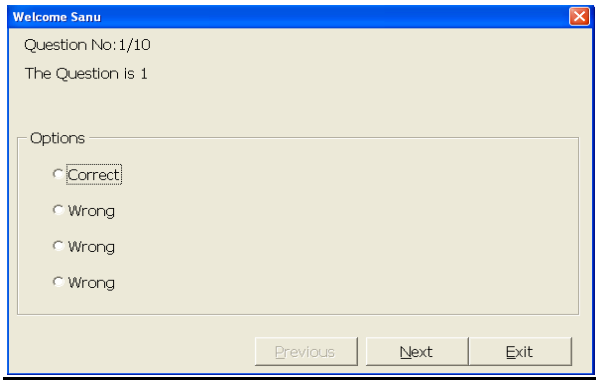



**Output in Excel from HTML doc**

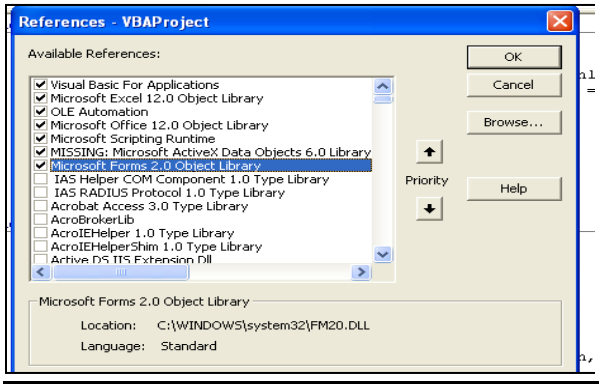
| | A | B | C | D |
|---|---|---|---|---|
| 1 | Index | Current | Pt. Change | % Change |
| 2 | | | | |
| 3 | S&P BSE SENSEX | 20399.4 | 205.02 | 1.02 |
| 4 | | | | |
| 5 | | | | |
| 6 | S&P BSE 100 | 6058.14 | 72.68 | 1.21 |
| 7 | | | | |
| 8 | S&P BSE MID CAP | 6141.05 | 64.26 | 1.06 |
| 9 | | | | |
| 10 | S&P BSE SMALL CAP | 5956.17 | 64.2 | 1.09 |
| 11 | | | | |
| 12 | S&P BSE 200 | 2416.32 | 28.06 | 1.17 |
| 13 | | | | |

## 8. Learning Modules




## 9. Data Consolidation Using MSAccess.




***Code***

```
Public Conn As New ADODB.Connection
Public Rst As New ADODB.Recordset
With Conn
.Provider = "Microsoft.ACE.OLEDB.12.0"
.Open "D:\Data.accdb"
End With
Rst.Open "Select Distinct Region from tblCSVData", Conn, adOpenStatic, adLockReadOnly
```

**Explanation**

We can connect access to excel using the code shown. Also shown is how we can run a sql query in access and export in excel.
This is for huge data which needs database and is maintained in MS Access. By VBA we can connect Excel to Access and run all sql queries and fetch the required data into excel.This saves a lot of time and energy with huge data source files.

**Conclusion:** VBA is the approach for data analysts in the data analysis field. VBA Techniques will have a bigger impact in the way of working for data analysts by automation of repetitive tasks and exploring the benefits which will save time& energy and adding value to the organization. Learning VBA will also upgrade the skill the data analyst and give a sense of achievement and will set the first step into the field of Analytics.VBA techniques should be used in data analysis and automation to get accurate outputs in the most efficient way.

**References:**

[1] Microsoft Office Excel 2003 Programming Inside Out / Curtis Frye, Wayne S. Freeze, Felicia K.Buckingham.p.cm.Includes index.ISBN 0-7356-1985-9
[2] www.tutorialspoint.com/**vba**/**vba**_overview.htm
[3]https://en.wikipedia.org/wiki/Visual_Basic_for_Applications
[4] www.excel-**vba**.com/excel-**vba**-contents.htm
[5] https://msdn.microsoft.com/en-us/library/.../ee814737(v=office.14).aspx
[6] excel**vba**tutor.com/**vba**_tutorial.html

# A Review on Forward and Backward Reasoning in AI

**Kiran Umesh Kadam**
Student, VPASC College, Baramati.
kadamkiran0089@gmail.com

**Abstract**

**Nowadays, more and more students all over the world need expert systems, especially in academic sectors. They need advice in order to be successful in their studies, and this advice must be provided by the academic management. There are two reasoning strategies in expert system, which have become the major practical application of artificial intelligence research: forward chaining and backward chaining. The first one starts from the available facts and attempts to draw conclusions about the goal. The second strategy starts from expectations of what the goal is, and then attempts to find evidence to support these hypotheses. The aim of this paper is to make a comparative study to identify which reasoning strategy system (forward chaining or backward chaining) is more applicable when making evaluations in expert management, especially in the academic field.**

**Keywords –**
Artificial Intelligence, Expert system, inference engine, knowledge base, reasoning method

## I. INTRODUCTION

In the academic field, some students need the best advice in order to improve their situation, and this advice must be provided by the academic management. In addition, as modern society has become more specialized, the need to have expert systems is increasing rapidly. These needs are visible in all areas of life .These days, human provide most expert advice. They have initiative and intelligence to adapt to a new model automatically and made decisions based on new data rapidly. However, humans also take a long time to learn things and can only perform one complete task at a time. For some of the reasons mentioned above, many companies have attempted to a mass and collate the knowledge of experts about a special problem area. Some other reasons are as follows: the knowledge of people who have particular expertise can be stored before they retire, so the company does not lose their knowledge and the implementation of expert systems lightens the load on specialists. Such systems can be created to solve routine problems more easily and quickly. The aim of this paper is to make a comparative study between forward-chaining and backward-chaining to select which of them is more applicable to the management expert system. The comparison will be based on the adequacy of the reasoning strategy, which means to what extent the given strategy can be used to represent all knowledge required for the management expert system. The basic description of an educational management expert system is the use of expert system techniques in management to be used in the academic sector.

## II.FORWARD REASONING:

An inference engine using forward chaining searches the inference rules until it finds one where the IF clause is known to be true. When found it can conclude, or infer, the THEN clause, resulting in the addition of new information to its dataset. In other words, it starts with some facts and applies rules to find all possible conclusions. Therefore, it is also known as **Data Driven Approach** The standard definition of a forward-chaining system is that the system operates by repeating the following sequence of operations .

1. Examine the rules to find one who's If part is satisfied by the current contents of Working Memory.
2. Fire the rule by adding to Working Memory the facts that are specified in the rules Then part. (The Then part may perform other actions as well, but that can be ignored for now.)

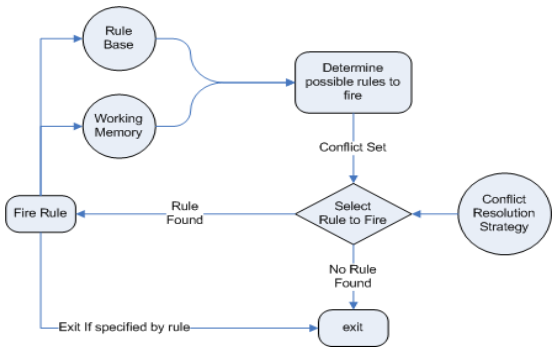



Fig.- Forward Reasoning algorithm

## III.BACKWORD REASONING:

An inference engine using backward chaining would search the inference rules until it finds one which has a

THEN clause that matches a desired goal. If the IF clause of that inference rule is not known to be true, then it is added to the list of goals (in order for goal to be confirmed it must also provide data that confirms this new rule). In other words, this approach starts with the desired conclusion and works backward to find supporting facts. Therefore, it is also known as **Goal-Driven Approach**. Backward-chaining systems try to satisfy the goals in the goal stack. They do this by finding rules that can conclude the information needed by the goal, and trying to make the If parts of those rules satisfied. In more detail, the standard backward-chaining control cycle.

1. Check the conclusions of the rules to find all rules that can satisfy the top goal on the stack.
2. Process these rules one at a time:
   **a.** Evaluate the conditions in the rules If part one at a time:

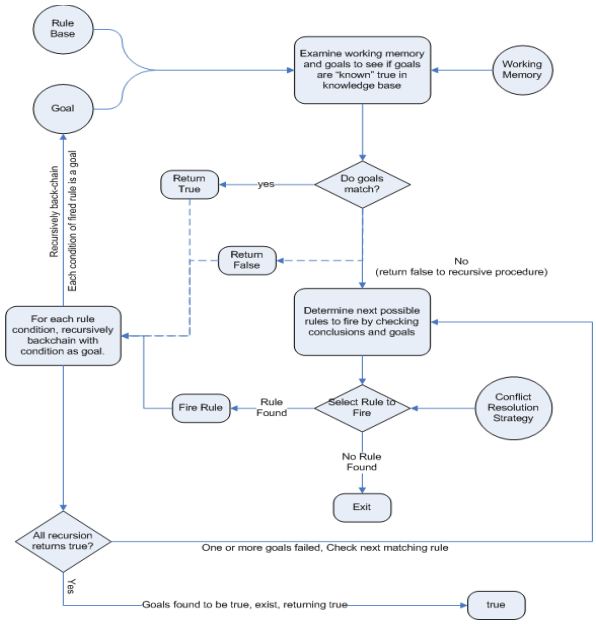



Fig.- Backward Reasoning algorithm

i. If the condition is currently unknown (that is, if there is not enough information currently known to determine whether the condition is satisfied) push a goal to make that condition known, and recursively invoke the system.
ii. If the condition is known to be unsatisfied, continue with the loop at Step 2.
iii. If it was not possible to determine whether the condition was satisfied, continue with the loop at Step 2.

**b.**If all the conditions in the selected rule are satisfied, add to Working Memory the facts specified in the Then part of the rule, pop the goal off the stack, and return from this invocation of the system.The system terminates with success when the goal stack is empty.

## IV. COMPARISON BETWEEN FORWARD REASONING AND BACKWARD REASONING

| **Forward-reasoning** | **Backward-reasoning** |
|---|---|
| Starts with the initial facts. | Starts with some hypothesis or goal. |
| Asks many questions. | Asks few questions. |
| Tests all the rules. | Tests some rules. |
| Slow, because it tests all the rules. | Fast, because it tests fewer rules. |
| Provides a huge amount of information from just a small amount of data. | Provides a small amount of information from just a small amount of data. |
| Attempts to infer everything possible from the available information. | Searches only that part of the knowledge base that is relevant to the current problem. |
| Primarily data-driven | Goal-driven |
| Uses input; searches rules for answer | Begins with a hypothesis; seek information until the hypothesis is accepted or rejected. |

| Top-down reasoning | Bottom-up reasoning |
|---|---|
| Works forward to findconclusions from facts | Works backward to find facts that support the hypothesis |
| Tends to be breadth-first | Tends to be depth-first |
| Suitable for problems that startfrom data collection, e.g. planning, monitoring, control | Suitable for problems that start from a hypothesis, e.g. diagnosis |
| Non-focused because it infers allconclusions, may answer unrelated questions | Focused; questions all focused to prove the goal and search as only thepart of KB that is related to the problem |
| Explanation not facilitated | Explanation facilitated. |
| All data is available | Data must be acquired interactive(i.e. on demand) |
| A small number of initial states but a high number of conclusions | A small number of initial goals and alarge number of rules match the facts |
| Forming a goal is difficult | Easy to form a goal |

**V.CONCLUSION**

To conclude which of the reasoning strategies is better for the GAES, we will consider the recommendations of for choosing either the forward or backward-chaining, and we will apply each of these recommendations to the GAES. Unfortunately, the only way to know how a rule will behave is to profile it and understand the business case. Many people make the mistake of thinking a rule engine will magically solve their problems. Writing high performance rules is not easy or intuitive. The best way to build efficient applications using rule engines is to take the time to learn how each approach works and use both techniques. Although it increases the learning curve, the choice between forward and backward chaining isn't something that can be summed up in 2-3 sentences. It's crucial to consider the intent of the rule, size of the dataset and performance requirements.

**VI. REFERENCES**

**Reference Paper-**

1. Artificial Intelligence : A Knowladge Based Approach ;Morris W. Firebaugh, 1988
2. Artificial intelligence / Elaine Rich, Kevin Knight. New York : McGraw-Hill, c1991
3. Catalogue of artificial intelligence techniques / ed. Alan Bundy Berlin : Springer-Verlag, c1990

**Books:**

5. Artificial Intelligence Third Edition McGraw Hill Education Elaine Rich,Kevin Knight, Shivshankar B. Nair.
6. Artificial Intelligence Vision Publication Third Edition

# Future and possibilities in Wireless Communication

**Dr Prashant P. Mulay**
Asst. Prof., Dept. of Computer Science,
Annasaheb Magar College, Hadapsar, Pune-411028.
ppm1575@gmail.com

***ABSTRACT:***

**In this paper, the author discussed the basics of wireless communication technology in detail and also how Wi-Fi can be replaced by Li-Fi. Wi-Fi is useful for general wireless coverage within buildings while Li-Fi is ideal for high density wireless data coverage in confined areas where there are no obstacles. Li-Fi is a wireless optical networking technology that uses light emitting diodes (LEDs) for transmission of data. The term Li-Fi refers to visible light communication (VLC) technology that uses light as medium to deliver high-speed communication in a manner similar to Wi-Fi. Li-Fi provides better bandwidth, efficiency, availability and security than Wi-Fi and has already achieved high speeds in the laboratory. In the present paper the author has discussed various aspects such as standards, operations of physical and MAC layer, modulation techniques used to transfer data in the visual light communication technology. He also listed out its advantages, applications and its future scope.**

***Keywords***: ***LED; VLC; Wi-Fi; Bandwidth.***

## INTRODUCTION

Information communication technology (ICT) facilitates mobility of data. Mobility and speed are key issues in data communication. Wi-Fi with RF technology provides both but it is facing an issue of spectrum availability. When number of users increases the speed decreases dramatically. To tackle the problem of speed on internet, Professor Harold Hass, suggested an idea called "Data through illumination" which is based on visual light communication. He used fiber optics to send data through LED light bulbs. Light modulation certainly is not a new concept, but Haas is looking to move things forward and enable connectivity through simple LED bulbs. With Li-Fi, we can connect to the internet simply by being within range of an LED beam, or we could conceivably transmit data using street lights or our car headlights.

Visible Light Communication systems use Light Emitting Diodes as a medium to high-speed communication in a similar manner as Wi-Fi. It is typically implemented using white LED bulbs at the downlink transmitter. By varying the current through the LED at a very high speed, we can vary the output at very high speeds. This is the principle of the Li-Fi. The working of the Li-Fi is itself very simple—if the LED is ON, the signal transmitted is a digital 1 whereas if it is OFF, the signal transmitted is a digital 0. By varying the rate at which the LEDs flicker, we can encode various data and transmit it. In this paper researcher studied the various aspects of wireless communications.

OBJECTIVE

The researcher has initiated this study with following objectives-

1. To study radio communication and its limitations.
2. To study various aspects of visual light communication
3. To study the utility of Li-Fi and its applications.

## CURRENT SCENARIO IN COMMUNICATION

### Radio Frequency Communication

A wireless system uses radio frequencies to "connect" users and is capable of operating over a much larger geographical area than a hard-line (wired) system.

Wireless systems (radios) transmit data and voice information using a specific radio frequency (RF) to other radios tuned to the same frequency. Common radio messages are transmitted over the RF band between 0.05 MHz and 900 MHz. Most public safety communications radios (portable, mobile, base station, and repeaters) transmit frequencies between 30 MHz and 900 MHz which are dedicated to public service use. Cell phones and systems, such as global positioning receivers, call boxes, electronic signs, irrigation systems, and mobile command units, that transmit information from remote locations, transmit in the microwave band between 1 GHz and 20 GHz.

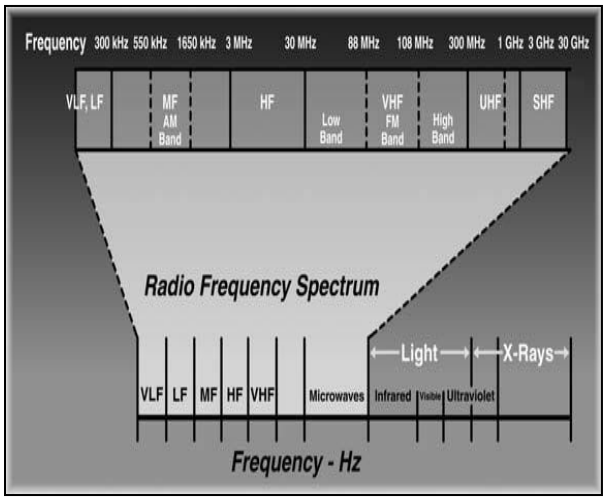



Figure1: Radio Frequency Spectrum

Within the HF spectrum, groups of frequencies are allocated to specific radio services — aviation, maritime, military, government, broadcast, or amateur. Frequencies are further regulated according to transmission type: emergency, broadcast, voice, Morse code, facsimile, and data. Frequency allocations are governed by international treaty and national licensing authorities. When we modulate radio waves to carry information, we refer to them as carriers. To convey information, a carrier must be varied so that its properties — its amplitude, frequency, or phase (the measurement of a complete wave cycle) — are changed, or modulated, by the information signal. There are two basic modes of propagation: ground waves and sky waves. As their names imply, ground waves travel along the surface of the earth, while sky waves "bounce" back to earth. Sky waves make beyond line-of-sight (BLOS) communications possible.

## Visual Light Communication (VLC):

### Overview

VLC refers to any use of the visible light portion of the electromagnetic spectrum to transmit information. It is certified by the IEEE 802.15 with the final standard being approved in 2011. The standard of VLC specifies VLC consisting of mobile-to-mobile (M2M), fixed-to-mobile (F2M) and infrastructure-to-mobile (I2M) communications. The main purpose of VLC is to focus on short-range mobile to mobile and fixed to mobile communications at high speed.Data rates are supported from some 100 kbps up to 100 Mbps using various modulation schemes.

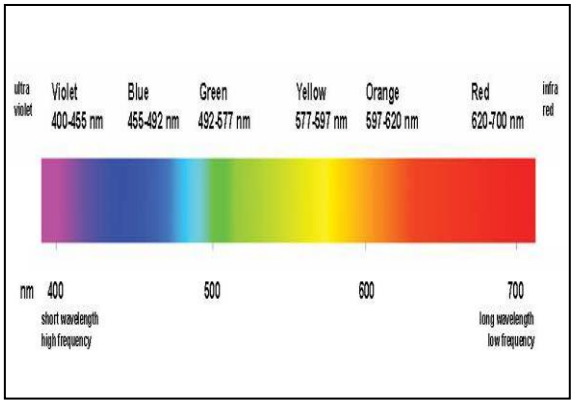



Figure 2 : Visible Light Spectrum
Source:
http://nextgenlite.com/images/VisibleLightSpectrumGradientForWeb.jpg

VLC is a data communication medium, which uses visible light between 400 THz (780 nm) and 800 THz (375 nm). It uses fast pulses of light to transmit information wirelessly.

The main components of this communication system are a white LED, Which acts as a communication source and a silicon photodiode serving as the receiving element. The LED can be switched on and off to generate digital strings of 1s and 0s. By modulating the LED light with the data signal, the LED illumination can be used as a communication source. A data rate of greater than 100 Mbps is possible by using high speed LEDs with appropriate multiplexing techniques. VLC data rate can be increased by parallel data transmission using LED arrays where each LED transmits a different data stream.

LEDs have many advantages both in communication and lighting purposes.

• Longer life time (on average, LEDs last three to five times longer than fluorescent and 20 to 30 times longer than incandescent light) .
• High brightness (3watt LED can produce illumination as a 45watt incandescent light bulb) .
• Lower energy consumption (and environmentally friendly).
• Fast switching (due to an LEDs' short response time, it can modulate easily for the purpose of communication. by quickly switching LEDs on and off (brighter and dimmer).

### Standards

The IEEE802.15.7 standard defines physical layer (PHY) & media access control (MAC) layer for short-range optical wireless communications using light wavelengths from 10,000 nm to 190 nm. The standard is capable of delivering data rates sufficient to support audio and video multimedia services and also considers mobility of the optical link, compatibility with various light infrastructures, impairments due to noise and interference from sources like ambient light. It also accommodates optical communications for cameras where transmitting devices incorporate light emitting sources and receivers are digital cameras with a lens and image sensor. The standard adheres to applicable eye safety regulations. The MAC layer supports 3 multi-access technologies: peer-to-peer, star configuration and broadcast mode. It also handles physical layer management issues such as addressing, collision avoidance and data acknowledgement protocols. The physical layer is divided into 3 types: PHY I, II, III and employ a combination of different modulation schemes.

The standard provides (i) access to several hundred THz of unlicensed spectrum; (ii) immunity to electromagnetic interference and noninterference with Radio Frequency (RF) systems; and (iii) for visible light systems, additional security by allowing the user to see the communication channel.

### Modulation Techniques:

***On-off keying (OOK):*** As the name suggests the data is conveyed by turning the LED off and on. In its simplest form a digital '1' is represented by the light 'on' state and a digital '0' is represented by the light 'off' state. The beauty of this

method is that it is really simple to generate and decode. The 802.15.7 standard uses Manchester Coding to ensure the period of positive pulses is the same as the negative ones but this also doubles the bandwidth required for OOK transmission. Alternatively, for higher bit rates run length limited (RLL) coding is used which is more spectrally efficient. Dimming is supported by adding an OOK extension which adjusts the aggregate output to the correct level.

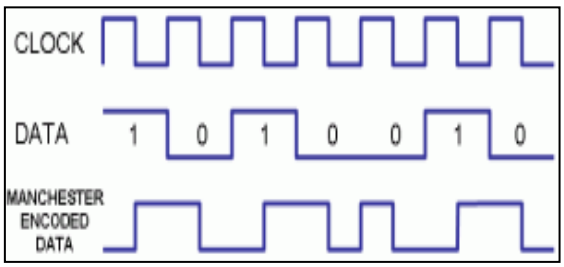



Figure 3 :The OOK modulation scheme uses Manchester Coding

***Variable pulse position modulation (VPPM):*** Pulse position modulation (PPM) encodes the data using the position of the pulse within a set time period. The duration of the period containing the pulse must be long enough to allow different positions to be identified, e.g. a '0' is represented by a positive pulse at the beginning of the period followed by a negative pulse, and a '1' is represented by a negative pulse at the beginning of the period followed by a positive pulse. VPPM is similar to PPM but it allows the pulse width to be controlled for light dimming support as shown.

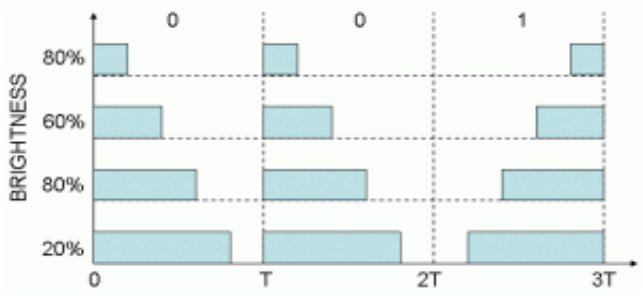



Figure 4 :Variable Pulse Position Modulation (VPPM) supports dimming

***Colour shift keying (CSK):*** This can be used if the illumination system uses RGB type LEDs. By combining the different colours of light, the output data can be carried by the colour itself and so the intensity of the output can be near constant. The xy chromaticity diagram shows the colour space and associated wavelengths in blue text (units are nm). Mixing of the red, green & blue primary sources produces the different colours which are coded as information bits. The disadvantage of this system is the complexity of both the transmitter and the receiver.

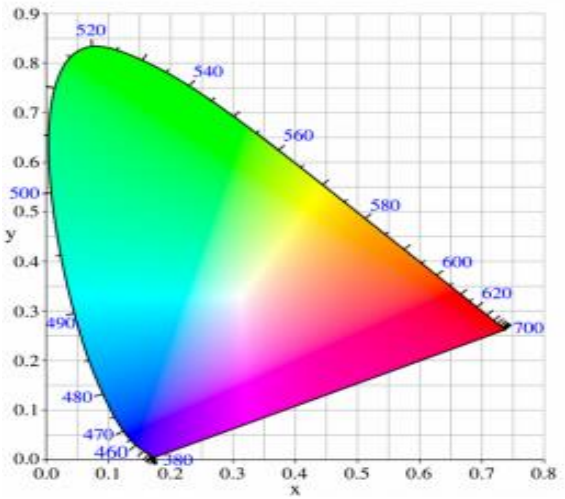



Figure 5 RGB LEDs can combine different wavelengths for CSK

## ADVANTAGES of VLC

The basic advantage of VLC is its bandwidth which much larger than the radio frequency bandwidth. With a larger bandwidth it is possible to accommodate more users and potentially achieve higher transfer rates because each user can be given a larger portion of the bandwidth to transfer information.

Visible light systems have over other communication systems is its abundance. Light sources are everywhere, and can be more efficiently used by increasing its simultaneous functionality by transmitting data in addition to lighting an area.

Li-Fi can achieve about 1000 times the data density of Wi-Fi because visible light can be well contained in a tight illumination area whereas RF tends to spread out and cause interference.

Very high data rates can be achieved due to low interference, high device bandwidths and high intensity optical output.

RF transmission and propagation in water is extremely difficult but Li-Fi works well in this environment. The transmission of light avoids the use of radio frequencies which can dangerously interfere with electronic circuitry in certain environments.

## APPLICATIONS of VLC

There are many applications for VLC. These include:
***RF Spectrum Relief***: Excess capacity demands of cellular networks can be off-loaded to Li-Fi networks where available. This is especially effective on the downlink where bottlenecks tend to occur.
***Smart Lighting***: Any private or public lighting including street lamps can be used to provide Li-Fi hotspots and the same communications and sensor infrastructure can be used to monitor and control lighting and data.
***Hazardous Environments***: Li-Fi provides a safe alternative to electromagnetic interference from radio frequency communications in environments such as mines and petrochemical plants.
***Hospital & Healthcare***: Li-Fi emits no electromagnetic interference and so does not interfere with medical instruments, nor is it interfered with by MRI scanners.
***Aviation***: Li-Fi can be used to reduce weight and cabling and add flexibility to seating layouts in aircraft passenger cabins where LED lights are already deployed. In-flight entertainment (IFE) systems can also be supported and integrated with passengers‟ own mobile devices.
***Underwater Communications***: Due to strong signal absorption in water, RF use is impractical. Acoustic waves have extremely low bandwidth and disturb marine life. Li-Fi provides a solution for short-range communications.
***Vehicles & Transportation***: LED headlights and taillights are being introduced. Street lamps, signage and traffic signals are also moving to LED. This can be used for vehicle-to-vehicle and vehicle-to-roadside communications. This can be applied for road safety and traffic management.

## CONCLUSION

Visible light was chosen for a variety of reasons, but primarily because it will not add to the cluttering of the radio frequency spectrum, which is heavily regulated by the FCC, and also because it will avoid the issue of interference in sensitive settings such as hospitals and airplanes. Limitations and drawbacks that we have to consider include noise from ambient light and the line-of-sight of the system. The main problem is that light cannot pass through objects, so if the receiver is inadvertently blocked in any way, then the signal will immediately be cut out. If the light signal is blocked one could switch back over to radio waves. If the intensity of ambient light is greater than that of the light from our system, the signal-to-noise ratio (SNR) is low, which will distort transmitted data.

VLC is particularly suitable for many popular internet “content consumption” applications such as video and audio downloads, live streaming, etc. These applications place heavy demands on the downlink bandwidth, but require minimal uplink capacity.

## REFERENCES

[1] http://nextgenlite.com/images/VisibleLightSpectrumGradientForWeb.jpg
[2] http://www.lificonsortium.org
[3] https://standards.ieee.org/develop/project/802.15.7.html
[4] IEEE: IEEE 802.15 WPAN Visual Light Communication Study Group (IGvlc) http://www.ieee802.org/15/pub/IGvlc.html, 2008
[5] Le-Minh, H., Zeng, L., O'Brien, D.C., Bouchet, O., Randel, S., Walewski, J, Borges, J.A.R., Langer, K.-D., grubor, J.G., Lee, K., and Won, E.T.: ‘Shortrange
[6] R. Abu-alhiga and H. Haas, “Subcarrier-Index Modulation OFDM,” in Proc. of the International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Tokyo, Japan, Sep. 13–16, 2009.
[7] Radio Communications: In the Digital Age, Volume One: HF Technology, Edition 2, Harris Corporation, RF Communications Division
[8] RF and Microwave Wireless Systems. Kai Chang , John Wiley & Sons, Inc., 2000.
[9] Visible light communication consortium, www.vlcc.net, 2008
[10] Visible Light Communications: challenges and Possibilities Dominic C. O'Brien, Lubin Zeng1, Hoa Le-Minh, Grahame Faulkner, Joachim W. Walewski, Sebastian Randel, University of Oxford (UK); Siemens AG, Corporate Technology, Information and Communications, Munich
[11] Visible Light Communications’, Wireless World Research Forum, 2007

# Fuzzy Model Evaluation for Intensify Multimedia Data Security in Cloud Enviornment using Homomorphic Cryptosystem

Anjuna D Raj[*]
Faculty of Dept of CSE
Christ College
Pune, India
anjunadevaraj@ gmail.com

Bhasisankar .B[**]
Network Administrator
EZWare Technologies Pvt ltd.
Kalyani nagar, Pune
bhasisankar@gmail.com

***Abstract***

**Cloud computing is a set of technologies that provide efficient and effective usage of information technology resources. The main feature of cloud is, it is an Internet based technology where quality services are provided to users including data and software, on remote servers. As it is an emerging field it leads too many privacy control weakness and vulnerabilities in its private storage. Cloud computing poses privacy concerns because the service provider can access the data that is in the cloud at any time. It could accidentally or deliberately alter or even delete information. Many cloud providers can share information with third parties if necessary for purposes of law and order even without a warrant. That is permitted in their privacy policies which users have to agree to before they start using cloud services. Security is a common factor which is facing by PaaS and IaaS cloud service models. Solutions to this include policy and legislation as well as end users' choices for how data is stored. Users can encrypt data that is processed or stored within the cloud to prevent unauthorized access[1]. Here we introduce a new concept Cryptosystem to secure our multimedia data from unauthorized usage. In cryptosystem combination of AES and RSA are used to encryption and decryption since AES is a better choice where it is still rigid to various cracking Techniques. The power of any Cryptosystem algorithm is mainly pointed to the number of keys used and the type of the key length which can be vary from 128-bit to 256-bit key.**

*Keywords*--***Cryptosystem, Cloud computing, Paas, Iaas, RSA, AES, Fuzzy***

## *Introduction*

Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. The new data storage in "Cloud" brings about many challenging design issues which have deep influence on the security and performance of the overall system. One of the biggest concerns with data storage in cloud is data integrity proof at untrusted servers. Prior to cloud computing, buyers of commercial software products did not demand the level of security from vendors that they do today. Software that was purchased and installed locally within the enterprise provided various security features for the buyer to configure in order to secure the application. The vendors would make it easy to integrate with enterprise security data stores such as Active Directory and provide single sign-on (SSO)[3] capabilities and other features so that buyers could configure the software to meet their security requirements. These commercial software products were run within the buyer's perimeter behind the corporate firewall. With cloud computing, vendors have a greater responsibility to secure the software on behalf of the cloud consumers. Since consumers are giving up control and often allowing their data to live outside of their firewall, they are now requiring that the vendors comply with various regulations. Building enterprise software in the cloud today requires a heavy focus on security.

The Advanced Encryption Standard (AES) is being developed to replace DES [DES]. The AES will be a new Federal Information Processing Standard (FIPS) Publication that will specify a cryptographic algorithm for use by U.S. Government organizations. However,

the AES will also be widely used by organizations, institutions, and individuals outside of the U.S. Government. A user of RSA creates and then publishes a public key based on two large prime numbers, along with an auxiliary value. The prime numbers must be kept secret. Anyone can use the public key to encrypt a message, but with currently published methods, if the public key is large enough, only someone with knowledge of the prime numbers can feasibly decode the message. After decryption the system checks the strength of the security by using fussy logic in matlab.

system study

### Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put 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. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

### A.1.Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

### A.2.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 lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

### A.3.Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

## SYSTEM ANALYSIS

The cloud computing is a cost-effective, service availability, flexible and on demand service delivery platform for providing business through the internet.The cloud can leak confidential data, modify the data, or return inconsistent data to different users. This may happen due to bugs, crashes, operator errors, or misconfigurations. Furthermore, malicious security breaches can be much harder to detect or more damaging than accidental ones: external adversaries may penetrate the cloud storage provider, or employees of the service provider may commit an insider attack. These concerns have prevented security conscious enterprises and consumers from using the cloud despite its benefits. Data integrity and privacy are the main issues in cloud storage environment. The requested multimedia data is transferred securely from cloud server to client. Here the Security is provided with combinations of cryptographic algorithms.

1).RSA (Riverst, Shamir, Adleman)

2). AES (Advanced Encryption Standard)

Stages to Encrypt the multimedia message Encryption of Image

*The encryption algorithm is divided into two stages i.e. Divide the image into block of fixed size then shuffling these blocks by applying the matrix transformation and at last applying bit rotation algorithm.[3]*

*The working of the encryption algorithm is explained in the following steps:*

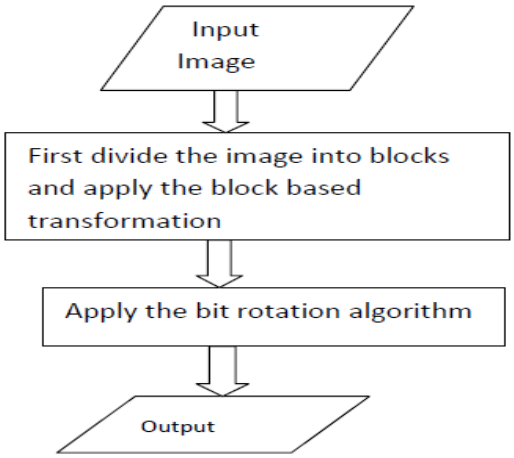



Fig 1: Diagram of Image Transformation technique

1) Step 1: Input Image.

2) Step 2: Divide the input image into equal size blocks, after that Shuffle these blocks by matrix transformation.

3) Step 3: Apply bit rotation algorithm.

4) Step 4: Output encrypted image.

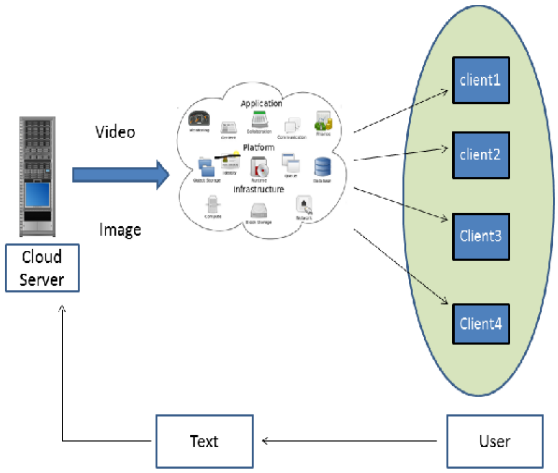



*Fig 2. Architecture diagram of proposed system*

### *Storing and Retrieving of multimedia data from cloud*

- In the First step, pick any text file, image and then it will be uploaded in the cloud computing work.
- It will check the integrity of data using data integrity process.
- The data files text file, images are encrypted for a secured storage.

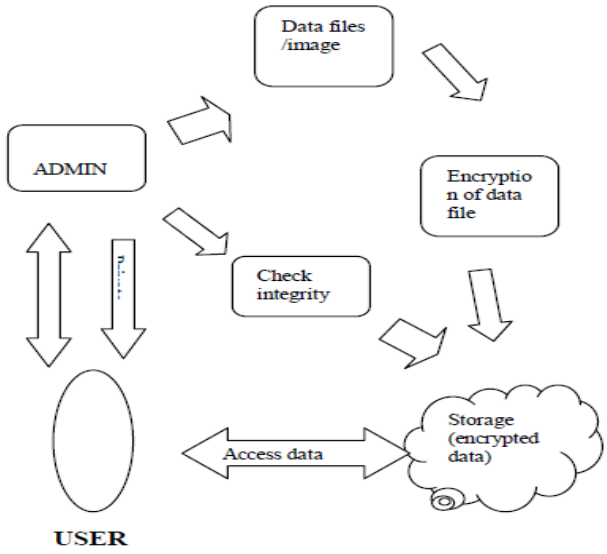



Fig 3. Storing and retrieving data from cloud

## SYSTEM DESIGN

Module 1: Cloud setup and data transformation.

Business-IT Corporate consist of different wings like Designing, Business analysis, HR team, Financial Analysis. Software testing, Software maintenance and Communication and Reporting centers. Each wing consists of number systems having different data accessibility and they are handling different software and tools. This all have to be installed in their system in accordance with their type of working environment. Once they start up with cloud no need to install and configure each part at each time, everything is there in the cloud server. The designed cloud storage can be used within the intranet by using any local network or by WiFi. The System Admin have the privilege to give the accessible areas to each of the employees as per their wings. Each of the employees are provided with their own username and password for access their system. Thus we can achieve the security to each of their work and to their PC. There are two types of user, a user might be a registered one or be a visitor. Authorized one in the sense that one who have login id or a registered one.

A visitor may not be registered they may be just a viewer of our site. For a registered one they can access our site in the sense of getting our services like download documents, image or any type of information, giving feedback for rate our services that leads to the popularity of the organization and giving their suggestion for the product and place orders. If he/her is a valid user then the request is forwarded to the Cloud server. The cloud server with the help of database server validates the user and provides services**.**

The users of the Drive are as follows:

**Employee:**

- Have only the right to access
- Can view the basic things of an organization like view notices memo and application submission like leave request facilities management.
- They are not allowed to make changes.

**Team Lead/Project Head:**

- They have some more privilege than employees
- Application acceptance and rejection, forwarding to higher authorities
- Instructing the employees and lead them for the completion of system development/ maintenance/ testing.

**Team Manager:**

- Verify the work of the employees and motivate them to complete the work before the deadline.

**Human Resource(HR) Team:**

- They are performing basic read and write functions
- Can have the privilege to communicate with the clients for understanding their requirements.
- Selecting young and new talents to the organization.

**Finance and Account Team:**

- This people are doing all the financial and economical activities of the organization like sending tender details to the clients.
- Checking monthly profit and evaluating the economical status by surveying continues statistical reporting

Module 2: Encryption and Decryption in **RSA**

RSA scheme is a block cipher in which the plain text and cipher text are integer between 0 and n-1 for some n. A typical size for n is 1024 bits or 309 decimal digits. That is n is less than $2^{1024}$. RSA scheme has since that time reigned supreme as the most widely accepted and implemented general purpose approach to public key encryption. The RSA algorithm without digital signature is not secured. A user of RSA creates and then publishes the product of two large prime numbers along with an auxiliary value, as their public key. The prime factors must be kept secret. Anyone can use the public key to encrypt a message, but with currently published methods, if the public key is large enough, only someone with knowledge of the prime factors can feasibly decode the message.
In RSA, encryption keys are public, while the decryption keys are not, so only the person with the correct decryption key can decipher an encrypted message. Everyone has their own encryption and decryption keys. The keys must be made in such a way that the decryption key may not be easily deduced from the public encryption key. Security of RSA Brute force attack.

Key generate in RSA algorithm

**Step 1:**

Select two large prime

numbers randomly p and q,

The constraint is p!=q.

Step 2:

Calculate n; where n=p*q

Step 3:

Find Euler's totient function, $\Phi(n)= (p - 1) * (q - 1)$

Step 4:

Find public key 'e', gcd(e, $\Phi(n)$)=1;

$$1<e< \Phi(n)$$

Step 5:

Find private key 'd',

$de=1 \bmod \Phi(n)$

$d=e^{-1} \bmod \Phi(n)$

$d=(1+k \Phi(n))\%e$

Step 6:Public key size e,n; here d is secret p,q and $\Phi(n)$ are known to r value.

For example p = 3 and q = 11

Compute n = p * q (i.e.) n = 3 * 11 = 33

Compute $\Phi(n) = (p - 1) * (q - 1)$

$\Phi(n) = (3 - 1) * (11 - 1)$

$\Phi(n) = 2 * 10$

$\Phi(n) = 20$

Choose e such that $1 < e < \Phi(n)$ and e and n are co-prime.

Let e = 7

$1< 7 <20$

Compute a value for d such that $(d * e) \% \Phi(n) = 1$

Check the value $d[(d*e)\% \Phi(n)=1]$

One solution is d = 3 [(3 * 7) % 20 = 1]

Public key is (e, n) => (7, 33)

Private Key is (d, n) => (3, 33)

The encryption for message $m = 2$ is $c = 2^7 \% 33 = 29$

The decryption of $c = 29$ is $m = 29^3 \% 33 = 2$

Module 3: Encryption and Decryption using **AES**

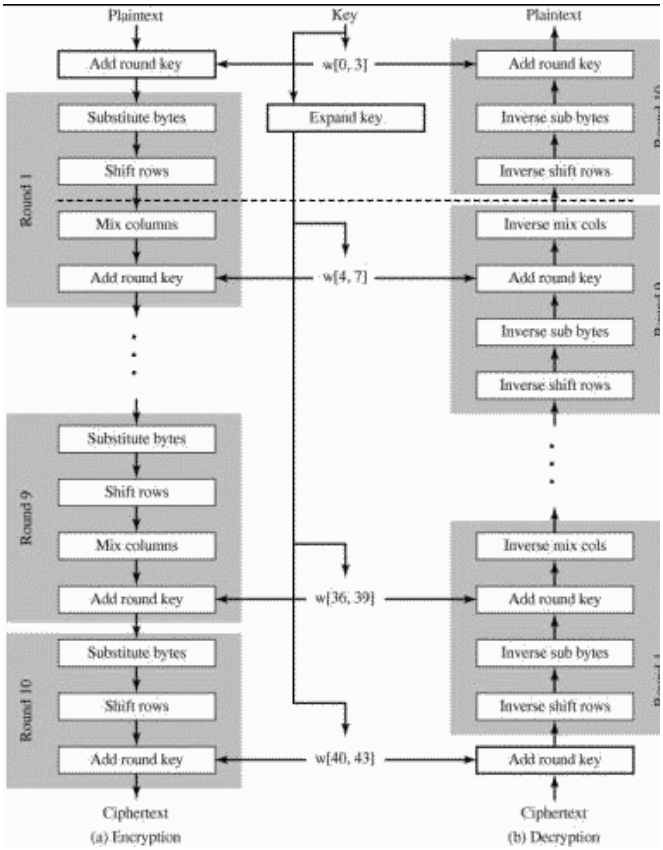



Fig 4. Data encryption and decryption in AES

The AES Encryption process works as follows:
Processor writes the key data to AES Key register. Now the key setup operation has to be performed. The key setup operation is requested by writing the KEY_SETUP command to the AES Control register. The completion of the key setup operation is flagged by the interrupt signal and the busy bit in the AES Status register.

Processor writes plain data to AES Data register. The 128 bit input data block is delivered by four subsequent write commands to the AES Data register. Finally the encoding operation is started by writing the command AES_ENCRYPT to the AES Control register. The completion of the encoding operation is flagged by the interrupt signal and the busy bit in the AES status register. Now the encrypted data may be read from the AES Data register by four subsequent read commands. The key remains in the module until a new key is written or a reset command is performed.

An approach for encryption and decryption In this it consist of various process.

- Substitute bytes
- Shift rows
- Mix column
- Add round key

**Substitute bytes:**
It uses s-boxes to perform a byte by byte substitution of the block
**Shift rows:**
A simple permutation are performed here
**Mix columns:**
A substitution that makes use of arithmetic over GF(28)
**Add round key:**
A simple bitwise XOR of the current block with a portion of the expanded key. Cipher begins with an add round key followed by 9 rounds that each includes all 4 stages followed by a 10th round of stages

### Time Complexity of RSA

Computing $A \cdot B$ in Z has time complexity $O(a \cdot b)$ or $O(\log(A) \cdot \log(B)$ When $b=O(a) b=O(a)$, that time complexity is $O(a^2)$ or $O(\log(A)^2)$. Euclidian division of $A \in Z$ by $N \in N^*$ that is computing $R \in [0 \ldots N-1]$ and $Q \in Z$ such that $A=N \cdot Q+R$, has time complexity identical to that for computing $N \cdot Q$,that is $O(n \cdot q)$; or $O(\log(N) \cdot \log(Q))$.
Modular exponentiation $A^B$ mod N with B>0requires at least $\log_2$ (B) modular multiplications (baring shortcuts possible using the factorization of N), and less than $32 \cdot \log_2$ (B) using square-and-multiply. It thus has time complexity $O(n^2 \cdot b)$ or $O(\log(N)^2 \cdot \log(B))$.
RSA assumption, computing $C^d$modN has time complexity $O(\log(N)^3)$ that's because dd has size in bits or digits proportional to that of NN (and further, close to that), thus the size of dd is O(log(N)). This remains unchanged when the CRT technique is used and there is a fixed number of prime factors in NN (which reduces the computational cost "only" proportionally to the square of the number of factors).
In RSA, if the public exponent ee is fixed, or of bounded size (which is common, and mandated by FIPS 186), then computing $M^e$modN has time complexity $O(\log(N)^2)$ that's because it is done with a fixed or bounded number of modular multiplications (e.g. 17 for the common $e=2^{16}+1$). On the other hand, ee is sometime allowed to be of the same order of magnitude as N (e.g. in PKCS#1), in which (rare) case $M^e$modN has time complexity $O(\log(N)^3)$
$O(\log(N)^3)$ is, of course, not the same as (and more than) $O(\log(N)^2)$ In practice, the RSA private key operation (using d) is often much slower than the public key operation (using e).

## V. FUZZY EVALUATION FOR SECURITY

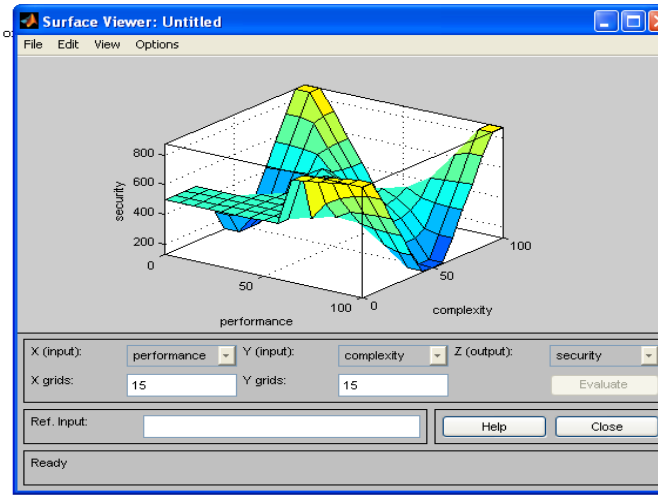



FIG 5: SECURITY EVALUATION

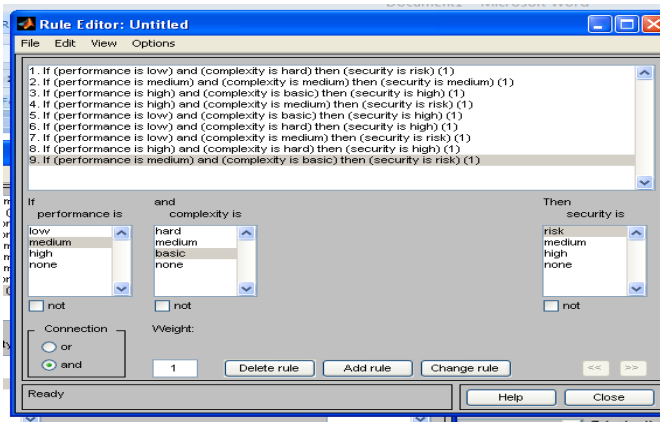



Fig 6: Here we are taking two dimensions for evaluating the security ie; Complexity of the algorithm and Performance. So in fussy we have two inputs which is performance and the second one is Complexity of algorithm. In performance we are taking the time to encrypt and decrypt the data. Then for complexity of algorithm is really based on the combination of alphabets with special characters in text data. [fig 6] shows the selection of possible combination of the inputs using if else and then rules. So in the graph the ups and downs are based on the inputs and rules which we are specified.

## VI. PERFORMANCE EVALUATION

One of the methods used for encryption and decryption is RSA algorithm.

Representation of RSA algorithm

Two prime numbers randomly

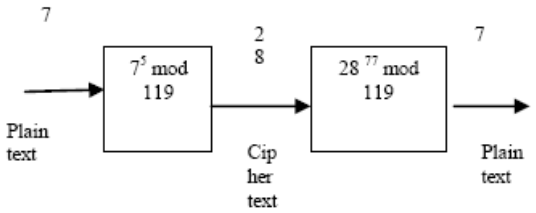
- P=7, q=17
- Φ (n)=(p-1)(q-1)
- Φ (n)=(6)(16)
- Φ (n)=96
- n=p*q 7*17
- Assume e=5(i.e.) e value should not be factor of Φ (n)
- D=d*e Mod Φ (n)=1
- D=d*5 Mod 96=1,so D=77

Now encrypting and decrypting the message HELLO

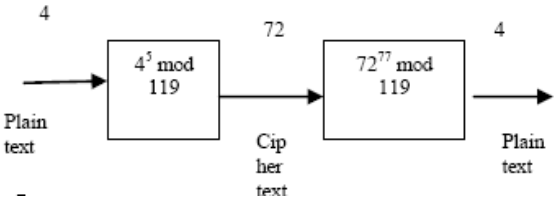
H E L L O

7 4 11 11 12

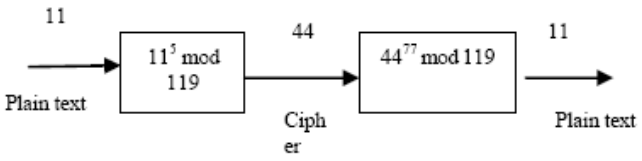
For H



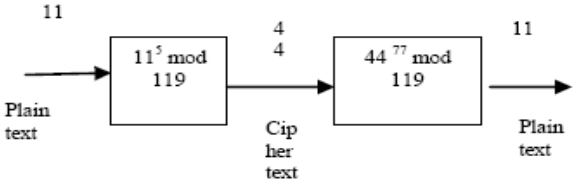

For E



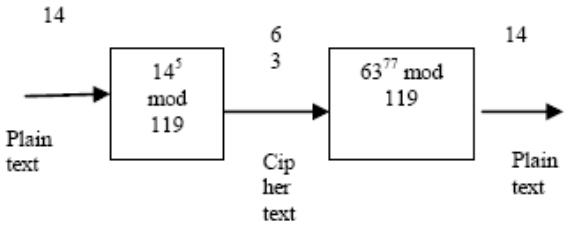

For L




For L



For O




## RELATED WORKS

Cloud computing concept is relatively new concept but it is based on not so many new technologies. Many of the features that makes cloud computing attractive, however has to meet certain basic security criteria. In our paper, we have briefed on various measure ion cloud computing security challenges from

single to multi clouds. While making a cloud secure, the following objectives are to be met: Understanding the cloud computing environment provided by the cloud service provider. The cloud computing solution should meet the basic security and privacy requirements of any firm deploying it. Maintain an account of the privacy of the cloud and data security and applications that are deployed in cloud computing environment. Data Integrity, Service Availability. The user runs customer applications using the service provider's resources

## CONCLUSION

Cloud computing is a new condition that is familiarizing in business surroundings where users can communicate directly with the virtualized resources. Data security has become the most important issue for cloud computing security.These algorithms(RSA & AES) are used for protecting user data include encryption prior to storage user authentication procedures prior to storage or retrieval and building secure channel for data transmission. This method maintains the availability reliability integrity to ensure coded data and at the same time identifies mischievous servers.

## FUTURE ENHANCEMENT

The future work of proposed system can concentrate on the security to multidimensional information which is like to locate the geographic areas and preserve the information about the place of data residence.

## REFERENCE

[1] Amazon Web Services. 2009. *Amazon Web Services: Overview of Security Processes.* [Online] Available from: http://aws.amazon.com/ec2/ [Accessed 26th April 2010]

[2] Bowman s.s, S. Roschke, et aI., "Intrusion Detection in the Cloud," presented at the Eighth IEEE International Conference on Dependable, AutonomIc and Secure Computing, Chengdu, China, 2009.

[3] Birendra Goswani, Dr.S.N.Singh, "Enhancing Security in Cloud computing using Public Key Cryptography with Matrices", International Journal of Engineering Research and Applications, Vol 2, Issue 4, 339-344, July-Aug 2012.

[4] Deyan Chen and Hong Zhao "Data Security and Privacy Protection Issues in Cloud Computing" 2012 IEEE International Conference on Computer Science and Electronics Engineering.

[5] G. Jai Arul Jose, C. Sajeev, Dr. C. Suyambulingom "Implementation of Data Security in Cloud Computing" International Journal of P2P Network Trends and Technology- Volume1 Issue1- 2011.

[6] J Brodkin. (2008). Gartner Seven cloud-computinsecurityAvailable:http://www.networkworld.com/news/

[7] Lizhe WANG, Gregor VON LASZEWSKI, Cloud Computing: a Perspective Study.

[8] Hassan Takabi, James B.D., Joshi, Gail-Joon, Ahn, "Security and Privacy Challenges in Cloud Computing Environments", University of Pittsberg, October 2010.

[9] T. Joachims, "Optimizing Search Engines Using Click through Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

[10] Neil Robinson, Lorenzo Valeri, Jonathan Cave and Tony Starkey (RAND Europe), Hans Graux (time.lex), Sadie Creese and Paul Hopkins (University of Warwick), "The Cloud: Understanding the Security, Privacy and TrustChallenges", TR-933-EC, 30 November 2010, Prepared for Unit F.5, Directorate –General, Information Society and Media, European Commission.

[11] Fuzzy Logic Toolbox™ User's Guide, The MathWorks, Inc.3 Apple Hill Drive Natick, MA 01760-2098

# A Review on Game Playing in AI

**Sagar Sanjay Khomane**
Student: VPASC College, Baramati.
Email:sagarskhomane14@gmail.com

**Abstract:**
**Games hold an inexplicable fascination for many people and the notion that computers might place games has existed at least as long as computers. The ideal way to use a search procedure to find a solution to a problem is to generate moves through the problem space until a goal state is reached. In the context of game-playing programs, a goal state is one in which we win. In the New AI Challenge and Competition, the general game playing has quickly evolved into an established research area. The aim of the paper is to address the aspect of general game playing which provides a unique approach to teaching a number of different topics such as problem solving by search, logic programming and planning.**

**Keywords:**
Artificial Intelligence, game playing, Game Description Language, Knowledge representation.

**Introduction:**
General game playing is the attempt to create a new generation of AI systems that are able to understand the rules of arbitrary games and then learn to play these games without human intervention. Unlike specialized systems like the chess program, Deep Blue, a general game player cannot rely on algorithms signed in advance for specific games. Such a system rather requires a form of general intelligence that enables it to autonomously adapt to new and possibly radically different environments. General game-playing systems are a quite essential example of a new generation of software that end users can customize for their own specific tasks. This makes general game playing an interesting and challenging problem for AI, involving many fundamental issues such as reasoning, learning, planning and decision making. Consequently, general game playing can, and in fact should, be of interest to researchers in a variety of AI disciplines beyond conventional computer game playing. At the same time and for the same reasons, general game playing provides a new anchor for AI education as a unique and attractive for student's framework for teaching multiple basic AI topics, such as problem solving by search, propositional and order logic, logic programming and planning. AI researcher does not lie in general game-playing systems themselves, they can be used as a non-trivial application for a broad range of more theoretically motivated AI research. A further advantage of general game playing is to naturally lead to practical assignments, where students can experiment with various techniques to address an interesting and challenging problem.

**Game Playing:**
Game playing has entered AI education, either in the form of an advanced AI graduate course, with lectures and tutorials but where the special focus lies on practical work; or as part of a general introductory course to AI. Yet another characteristic of general game playing research is often concern the combination and integration of two or more theories and methods, which naturally leads to collaborations involving different AI sub-disciplines. In the relatively short time span since the AI competition in 2005, at least four traditional AI disciplines have proved to be core aspects of research in general game playing:

1. Knowledge Representation
2. Search
3. Planning
4. Learning

**1. Knowledge Representation and Reasoning:**
General game playing requires a formal, symbolic language in which the rules of arbitrary games can be described to a system. The general Game Description Language (GDL) has been developed for the purpose that it can be seen as a specific high-level action description language in the tradition of AI Planning. Languages like STRIPS or PDDL: game worlds are described with the help of individual features (e.g., the position of a piece on a game board), and moves are specified by their preconditions.

2. **Search:**
Once a general game-playing system is capable of computing legal moves and position updates from the game rules, it can search through the space of possible ways in which the game can proceed. Only very simple games allow a complete brute-force search in practice. Hence, one of the fundamental challenges for general game playing is to develop intelligent search techniques without knowing the particulars of the game to which they will be applied. The early general game-playing systems applied standard iterative deepening search in combination with automatically constructed heuristic evaluation function. The state-of-the-art search methods in existing

general game-playing systems gives rise to a number of interesting challenge problems.

**3. Planning:**

Planning is closely related to general game playing, as both are instances of general problem solving, where the specifics of a problem are unknown until runtime the game description language GDL can in fact be seen a special-purpose action description language that follows the tradition of planning languages. The latter, however, describe a problem from the perspective of a single agent, even in case of adversarial planning, and GDL generalize this to the presence of other agents that have their own actions and goals. Reasoning about the intentions of the other players is the basis for opponent modeling one of the crucial aspects in which general game playing goes beyond planning. A further difference is that planning is mainly concerned with solving a problem upfront, whereas much of general game playing is concerned with ending a good course of actions without being able to see all the way to the end of a game.

**4. Learning:**

The very idea of general game playing is to build systems that automatically learn to master arbitrary new games. It is therefore justified to consider learning to be the core research area for this AI challenge. Thus it may be somehow surprising that general AI learning methods are not (yet) dominating the state of the art in general game playing. The main reason seems to be the sheer difficulty of the problem for general-purpose learning algorithms, which is why existing, successful applications of learning techniques in general game playing, focus on improving individual aspects of specific approaches: In [1], transfer learning has been applied to generalize knowledge from one game for other games.

**Further Relevant Research Areas:**

In addition to the four AI disciplines that today build the core of general game playing research. Other areas are likely to eventually become more relevant as the scope of general game playing is broadened decision making. Current state-of-the-art research in general game playing focuses on games that are deterministic and in which players have complete knowledge of the state. As we progress towards the general case of randomized, imperfect-information games, aspects of general decision making will become increasingly important, including utility functions, Markov decision processes, and game-theoretic solutions. In Natural Language Processing most end users of general game-playing systems will expect a better way than using GDL to input the rules of their own invented games. This could be achieved with the help of some controlled form of natural language for describing games. In most cases, this will include a process of disambiguation through a dialogue between the player and its user. This generalization seems particularly attractive because it could be met in a mostly modular fashion, where an existing general game player is coupled with a pre-processing natural language interface game playing robots. General game-playing software provides a great opportunity to make a relatively simple robotic system to act smart.

**Conclusion:**

General game playing is an exciting, still young but on the verge of maturing topic, which touches upon a broad range of aspects of artificial intelligence. In this paper we surveyed the research landscape of general game playing in an attempt to show its many facts and that it provides a rich source of interesting and challenging problems for many AI researcher we also shown that general game playing provides a unique approach to teaching a number of different topics in AI. Students who got exposed to the idea of a general game-playing AI system have repeatedly described it as cool, and the author is inclined to agree.

**References:**

Research Papers:

1. Bikramijt Banerjee and Peter Stone [2007] ."General game learning using knowledge transfer" International Joint Conference on Artificial Intelligence (IJCAI), p 672-677.
2. Michael Thielscher and Sebastian Voigt [2010] "A temporal proof system for general game playing" AAAI Conference on Artificial Intelligence, pages 1000-1005,

Books:

7. Artificial Intelligence Third Edition Mc Graw Hill Education Elaine Rich,Kevin Knight, Shivshankar B. Nair
8. Artificial Intelligence Vision Publication Third Edition

# GIGASPEED TECHNOLOGY : FUTURE APPLICATION OF LIFI

MinalChiranjitDutta
Department Of Computer Application
MAEER's MIT Arts,Commerce And
Science College Alandi(D),Pune
Maharashtra, India.

**Co-Author: Mrs. KavitaMahajan**
Department Of Computer Application
MAEER's MIT Arts,Commerce And
Science College Alandi(D),Pune
Maharashtra, India.
**Email Id:** kdmahajan@mitacsc.ac.in

**Abstract**:

**Technology has become integral part of twenty first century, we can't imagine this world without technology, Today technology has pierced in every aspect of our lives : from entertainment to communications. In today's high-tech world connectivity to networks is extremely important for organizations to survive. . At this point, thereneeds to define wireless technologies to be used.such as bluetooth,wifi and lifitechonologythe LIFI providesfastest wireless data transfer technology GIGASPEED WIRELESS TECHONOLOGY. Our current methods offer effective transmission rates of up to 10 Gbps, allowing a 2 hour HDTV film to be transferred in less than 30 seconds. This can be extended to several 100 Gbps in future versions.In this paper I have discussed about five types of Gigaspeed Techonology,1.GigaDock 2. GigaBeam 3.GigaShower 4.GigaSpot 5.GigaMIMO.**

Keywords: LIFI,Gigaspeed,GigaDock,GigaShower,GigaSpot, GigaBeam,GigaMIMO.

**Introduction**:

Li-Fi is a new paradigm for optical wireless technology to provide unprecedented connectivity within a localized data-centric environment. The increasing demand for higher bandwidths, faster and more secure data transmission as well as environmental and undoubtedly human friendly technology heralds the start of a major shift in wireless technology, a shift from RF to optical technologies.As the light waves are used for the transmission of data then of course it is faster data transmission than wifi and Bluetooth technology.

Alexander Graham Bell is most famous for inventing the telephone, but he also demonstrated the first VLC system in 1880. In fact only 4 years after inventing the telephone, Bell demonstrated the world's first wireless telephone call. He did this with an apparatus called a Photophone that used light, not radio. Bell had to use daylight for transmission, which severely limited its practical use.

Professor Harald Haas, the Chair of Mobile Communications at the University of Edinburgh, is recognized as the founder of Li-Fi. He coined the term Li-Fi and is the co-founder of pureLiFi.in his first demo lecture he used table lamp with LED bulb to transmit a video of blooming flower he then periodically put his hand in between and as per his conc;usionthe table lamp is the source of the video data even in interference through hands.He explained general definition behind that mechanism in simple words is that"Switched off the light means binary 0 and switched on the light means binay 1." There are following characteristics of LIFI[1]

1. Capacity: The visible spectrum of LIFI is more wider than wifi so it can able to cover larger distance with high speed ie 10 gigabits/sec.[1]

2. Availability : lights are available everywhere from the roads to our houses,offices.so there is no issue of the availability and installation.[1]

3. Efficiency: Light consumes less energy and they are highly efficient.[1,2,3]

4. Security : As the light cannot penetrate though walls so light medium is more secure to transmit the data.

We can find LIFI applications in variety of fields.Some of these are given below.

1 Indoor Communication
2 Vehicle to Vehicle communication
3 InDefence for secure data transmission.
4 Underwater Exploration
5 Medical and Healthcare Equipments.
6 Aviation Industry.
7 GigaspeedTechnology

Related Work :Li-Fi technology is based on LEDs for the transfer of data. The transfer of the data can be with the help of all kinds of light, no matter the part of the spectrum that they belong. That is, the light can belong to the invisible, ultraviolet or the visible part of the spectrum. Also, the speed of the internet is incredibly high and we can download movies, games, music etc in just a few minutes with the help of this technology. Also, the technology removes limitations that have been put on the user by the Wi-Fi. We do not need to be in a region that is Wi-Fi enabled to have access to the internet.[2,3,4]
The University of Edinburgh, pureVLC's partner and home to Prof. Harald Haas – "the father of Li-Fi" have already demonstrated 3Gbps on a single color. On a single LED with full color (R,G,B) this could communicate at speeds up to 9Gbps.

In October 2011, companies and industry groups formed the Li-Fi Consortium to promote high-speed optical wireless systems and to overcome the limited amount of radio-based wireless spectrum available by exploiting a completely different part of the electromagnetic spectrum.

A number of companies offer uni-directional VLC products, which is not the same as Li-Fi.

VLC technology was exhibited in 2012 using Li-Fi. By August 2013, data rates of over 1.6 Gbit/s were demonstrated over a single color LED.In September 2013, a press release said that Li-Fi, or VLC systems in general, do not require line-of-sight conditions.In October 2013, it was reported Chinese manufacturers were working on Li-Fi development kits.

In April 2014, the Russian company StinsComan announced the development of a Li-Fi wireless local network called BeamCaster. Their current module transfers data at 1.25 gigabytes per second but they foresee boosting speeds up to 5 GB/second in the near future. In 2014 a new record was established by Sisoft (a Mexican company) that was able to transfer data at speeds of up to 10Gbps across a light spectrum emitted by LED lamps.[11]

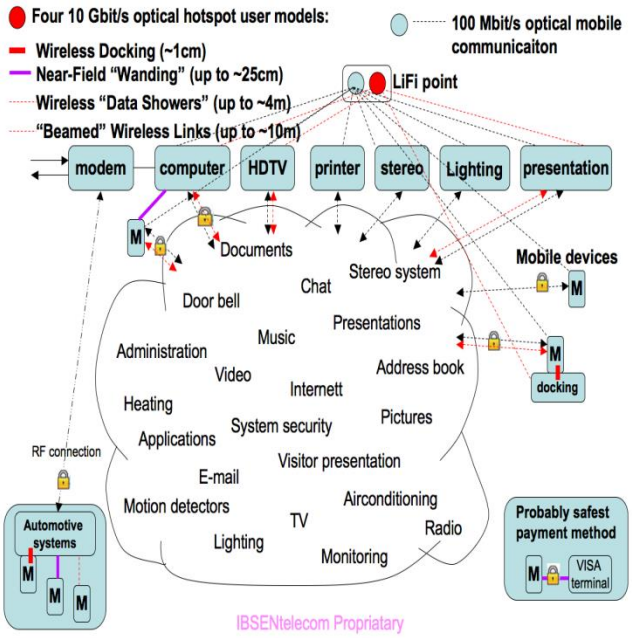



Fig 1 : LIFI Cloud [9]

Within a local Li-Fi cloud several data based services are supported through a heterogeneous communication system. In an initial approach, the Li-Fi Consortium defined different types of technologies to provide secure, reliable and ultra-high-speed wireless communication interfaces. These technologies included giga-speed technologies,[4,5,6]optical mobility technologies and navigation, precision location and gesture recognition technologies.[10] For giga-speed technologies, the Li-Fi Consortium defined Giga Dock, Giga Beam, Giga Shower, Giga Spot and Giga MIMO models to address different user scenarios for wireless indoor and indoor-like data transfers.

1.GigaDock Technology: Giga Dock is a wireless docking solution including wireless charging for smart phones tablets or notebooks, with speeds up to 10 Gbps. It is an embedded device which plugs into the mobile phone in which Docking station fits to charging pad & Mobile device sits on charging pad.TheGigaDock wireless technology docking to the wired devices, such peripheral device allow to charge the mobiles up to 10gigabits/sec along with two way flow of the data simultaneously ie Full Duplex Mode of Transmission. It provides high speed data transmission rate with higher compatibility for wireless charging.[8,9,10]

2.GigaBeamTechonology : the Giga Beam model is a point-to-point data link for kiosk applications or portable-to-portable data exchanges ,[9] As the name suggests it is a beam techonology. It is consist of number of Infra-Red ports for fast bulk transfers which Builts

into kiosks and public stations. IR ports available in mobile device or module.The data rate of this technology is upto 10Gigabit/sec wireless file transfer,It is important to hold the mobile device during transfer.Itis helpful for Wireless kiosk file transfers it provides Portable wireless screen & keyboard.[9,7]

3.GigaShower Technology : It is a shower like Ceiling unit sends up to 10Gbit/s stream it receivsinfra-red signals plug into mobile device.it also has the data rate up to 10 Gigabits/secs.to many users simultaneously [9].Each user can have a separate portion of the transmission path to share information it is called as channel. It is useful to broadcast large volume of the data transmission including voice and non-voice data, audio and video data etc.it can be implemented in public area, in meeting rooms etc.[6,7,8,9] Giga Shower provides unidirectional data services via several channels to multiple users with gigabit-class communication speed over several meters. This is like watching TV channels.[9]

4.GigaSPot Technology : It is the other model for in-house communication. There a transmitter or receiver is mounted into the ceiling connected to, for example, a media server. On the other side are portable or fixed devices on a desk in an office, in an operating room, in a production hall or at an airport.It can provide wireless connection to the peripherals.it can eliminate the use of docking system cables.[8,9,10]

5.GigaMIMO Technology : Giga MIMO is the optical wireless single- and multi-channel Hot Spot solutions offering bidirectional gigabit-class communication in a room, hall or shopping mall.[9] it has Many bidirectional high speed links delivers a data at the rate of 10 gibit/sec.multiple beams attached to it can track user movements.many simultaneous users can be used this technology for Mobile indoor communication.[6,9,10]

Conclusion: There are lots of possibilities and various numerous and can explored further. If this technology can be used into practical life we can used every bulb something similar like a Wi-Fihotspot to transmit wireless signal and data and we will proceed further toward the greener, safer, cleaner and brightest future. The concept of Li-Fi is new technology which currently attracting a great deal of interest, not least because it may offer a genuine and very efficient alternative to radio based wireless. As we know that the number of user increasing day by day and growing number of people and many communicating devices access via wireless internet, the airwaves are becomingincreasingly time to time and clogged, So, it making more and more difficult to get a reliable, high speed signal. This may solve and overcome various issues like such as the shortage of radio frequency bandwidth, poweralso allowing internet in that place where traditional radio based wireless is not allowed like aircraft or hospitals. One of the shortcomings is that it is only applicable in direct line of sight.

**References**:

[1] IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308,WIRELESS DATA TRANSFER USING VISIBLE LIGHT COMMUNICATION

[2] A Hybrid Radio Frequency and BroadcastVisible Light Communication System2nd IEEE Globecom, 2011 Workshop on Optical Wireless Communications, December 5 − 9, 2011, Houston, Texas, USA

[3]IEEE 802.15.7 visible light communication: modulation schemes and dimming support,| Abstract

[4] ISSN: 2321-7782 (Online)/ International Journal of Advance Research in ,Computer Science and Management Studies/*Li-Fi Technology: Data Transmission through Visible Light*

[5] Smart Lighting Annual Industry-Academia Days, February 13-15, 2012/ EXPLODING INTEREST IN VISIBLE LIGHT COMMUNICATIONS: Thomas Little /Boston University

[6] Plant Robot Motor Controller/Serial Communication Protocol Interface.

[7] MCL Technical Report No. 09-07-2011,Lights Off Visible Light Communication,T.Borogovac, M.B. Rahaim, M. Tuganbayeva,and T.D.C. Little, Boston UniversityBoston, Massachusetts 02215

[8] G.B. Stringfellow and M.G. Craford, "High Brightness Light Emitting Diodes",

Semiconductors and Semimetals Vol. 48,*Academic Press*, 1997.

[9]Li-Fi Consortiumnext generation optical wireless communication solution

[10] LI-FI (LIGHT FIDELITY)-THE FUTURE TECHNOLOGY IN WIRELESS COMMUNICATION

**[**11] Li-Fi internet solution from Russian company attracting foreign clients, Russia and India Report, Russia Beyond the Headlines, 1 July 2014

# BIG DATA

**Prof. Pradnya Deshpande**
Arihant College of Arts, Commerce and Science
pradnya.deshpande@arihant.org.in
9730853392

**Abstract:**

**Many studies have combined cloud computing and mobile computing to facilitate better storage, delivery, retrieval, and management of medical files for telecardiology.**
**With information technology, real-time tele-consultation and tele-diagnosis of ECG and images can be practiced via an e-platform for clinical, research, and educational purposes. While being devoted to promote the application of information technology onto telecardiology, we need to resolve several issues: (1) data confidentiality in the cloud, (2) data interoperability, and (3) network latency and accessibility. This paper presents an overview of the Mobile Data Challenge (MDC), a large-scale research initiative aimed at generating innovations around smartphone-based research, as well as community-based evaluation of related mobile data analysis methodologies.Longitudinal smartphone data set for the basis of the MDC. Then, we introduce the Open and Dedicated Tracks of the specific data sets used in each of them and discuss some of the key aspects in order to generate privacy-respecting, challenging, and scientifically relevant**
**mobile data resources for wider use of the research community.**

**Keywords:** mobile computing; cloud computing; telecardiology; electrocardiograph; big data.

**Introduction**

The ongoing growth of mobile communication networks, along with the launch of new mobile devices (e.g. smart phones and tablets) and the increased demand for additional data network traffic, pave the way to the need of a fully interconnected Information Society, supporting the ubiquitous provision of multiple services (e.g. multimedia and data services) to mobile users, located in urban and rural areas. In addition, the new trends of mobile networks that enable the interconnection of new types of devices along with their services (i.e. household appliances) rise the need for adopting new solutions and network architectures to support the provision of novel services and applications. Furthermore, the huge increase of the number of mobile devices with growing requirements for multiple services provision, open the way towards new communication standards that will be adopted by the year 2020 for the deployment of the future mobile networking systems. In this context, the fifth generation (5G) wireless communication systems are envisaged to lead to higher-level mobile users' experiences, significantly supporting increased connection rates (i.e. peak data rates of 10 Gb/s), while it will also contribute to host novel services and applications. With the emergence of ultra-fast 5G mobile networks and Highlyfeatured smartphones, tablets, and wearable computing devices (e.g., with always growing sensing capabilities), the prerequisites will be met for bringing cloud computing to the mobile domain. While first commercial products are restricted to the sharing of files, contacts, etc. among different devices, more sophisticated applications still have to be developed. In addition, it is undoubtedly true that Big Data is as a result of the 2.5 quintillion bytes of data generated each day with pervasive, wearable as well as handheld devices composed by configurations such as machine-to-machine connectivity and the related applications hosting online/m bile social networks. The multi-source collection of data brings into the researchers attention issues that should be investigated like novel access mechanisms and multi-source Big Data Collecting techniques, as well as the distributed Big Data Storing methodologies and finally Intra/Inter Big Data processing. Heterogeneous Big Data Processing in Mobile Cloud Computing Mobile Cloud Computing emerges as a new computing paradigm where mobile devices exploit the available cloud computing platform for performing specific tasks and/or accessing data on demand. With the widespread explotion of information, an increasing number of academic researches and industrial applications result in the appearance of big data from multiple heterogeneous sources in mobile clouds. Storage, transmission, analysis, and processing for such heterogeneous big data are crucially required. The topic in this special session ranges from case studies of particular problems with multisource heterogeneous data to novel learning theories and approaches in mobile clouds. This special session aims to bring together practitioners and researchers with a particular focus on the interesting topics associated with big data from multiple heterogeneous sources for mobile cloud computing.

Big data can expose people's hidden behavioral patterns and even shed light on their intentions. More precisely,it

can bridge the gap between what people want to do and what they actually do as well as how they interact with others and their environment. This information is useful to government agencies as well as private companies to support decision making in areas ranging from law enforcement to social services to homeland security. It's particularly of interest to applied areas of situational awareness and the anticipatory approaches required for near-real-time discovery. In the scientific domain, secondary uses of patient data could lead to the discovery of cures for a wide range of devastating diseases and the prevention of others. By revealing the genetic origin of illnesses, such as mutations related to cancer, the Human Genome Project, completed in 2003, is one project that's a testament to the promises of big data. To extract meaningful business value using analytics, the big data stack needs to be expanded with a number of components and technologies. In short, these are: > the management of data assets, with real-time process control for analytics > scaling up analytics to gain insights from big data, as well as the ability to solve increasingly complex problems using more data > an analytics-as-a-service approach and the flexibility to choose between local and hosted deployment options > securing consumer privacy, while still extracting relevant and timely information from data > and most importantly, domain-specific analytics platforms and SDKs, and access to experts to ensure that true insights are being extracted.

**NEW CHALLENGES**

While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be maintained. For example, both companies and law enforcement agencies increasingly rely on video data for surveillance and criminal investigation. Closed-circuit television (CCTV) is ubiquitous in many commercial buildings and public spaces. Police cars have cameras to record pursuits andtraffic stops, as well as dash-cams for complaint handling.Many agencies are now experimenting with body-worn video cameras to record incidents and gather direct evidence from a crime scene for use in court, obviating the need for eyewitness versions of events. Taser gunsalso now come equipped with tiny cameras. Because all of these devices can quickly generate a large amount of data, which can be expensive to store and time-consuming to process, operators must decide whether it is more costeffective to let them run continuously or only capture selective images or scenes.

Big data also presents new ethical challenges.Corporations are using big data to learn more about

their workforce, increase productivity, and introduce revolutionary business processes. However, these improvements come at a cost: tracking employees' every move and continuously measuring their performance against industry benchmarks introduces a level of oversight that can quash the human spirit. Such monitoring might be in the best interest of a corporation but is not always in the best interest of the people who make up that corporation. In addition, as big multimedia datasets become commonplace, the boundaries between public and private space will blur. Emerging online apps will not only enable users to upload video via mobile social networking but

will soon incorporate wearable devices in the form of a digital watch or glasses to allow for continuous audiovisual capture. People will essentially *become* a camera. This

publicly available data will dwarf that generated by today's CCTV cameras.

However, unlike surveillance cameras, smartphones and wearable devices afford no privacy protection to innocent bystanders who are captured in a video at the right place at the wrong time. For example, in the wake of the recent Boston bombings, images of several people photographed at the scene were mistakenly identified as suspects on social media sites.In fact, one of the major challenges of big data is preserving individual privacy. As we go about our everyday

lives, we leave behind digital footprints that, when combined, could denote unique aspects about ourselves that would otherwise go unnoticed, akin to digital DNA. Examples include our use of language and punctuation in blog and forum posts, the clothes we wear in different contexts, and the places we frequent—do we spend our Sunday mornings outdoors playing sports, indoors online, visiting friends, attending religious services, or cruising a bad part of town? Something as innocuous as when and how we use energy in our homes reveals many details about us. Outside our homes, drones could well be used for ad hoc monitoring, spotting unusual changes in land

use patterns and feeding data back to operation centers about emergencies.

Big data analytics will draw on aspects of our home,work, and social lives to make assumptions beyond typical "market segmentations" and delve deep into ontological questions such as, "Who are you?" This has metaphysical implications. For example, people will consciously alter

their online activity, and will modify their behavior in surveilled spaces, to protect their privacy. Big data will change how we live in both small and large ways. Are we on a trajectory toward an uberveillance society? Big data will change how we live in both small and large ways.

This has widespread social implications. The phones themselves have become instruments for fast communication and collective participation. Further, different user groups, like teenagers, have started to use them in creative ways. At the same time, the number of

sensors embedded in phones and the applications built around them have exploded. In the past few years smartphones remarkably started to carry sensors like GPS, accelerometer, gyroscope, microphone, camera and Bluetooth. Related application and service offering covers e.g. information search, entertainment or healthcare. The ubiquity of mobile phones and the increasing wealth of the data generated from sensors and applications are giving rise to a new research domain across computing and social science. Mobile Cloud Computing emerges as a new computing paradigm where mobile devices exploit the available cloud computing platform for performing specific tasks and/or accessing data on demand. With the widespread explotion of information, an increasing number of acadamic researches and industrial applications result in the appearance of big data from multiple heterogeneous sources in mobile clouds. Storage, transmission, analysis, and processing for such heterogeneous big data are crucially required. The topic in this special session ranges from case studies of particular problems with multisource heterogeneous data to novel learning theories and approaches in mobile clouds.

**Overview of the relationship between big data and mobile:**

- **First, big data needs to be defined:** Big data is most commonly defined as data sets that meet three attributes, known as the three "Vs": volume, variety, and velocity. But there is something more to it. "I like to say there's a fourth V: value," says Kipp Jones, vice president of product at Skyhook. In order for data to be meaningful at all, it needs to be captured and stored efficiently. Then someone has to manage the data, analyze it, and extract value from it. Data, big or not, doesn't add up to anything worthwhile if it doesn't have value to someone.
- **Mobile is particularly well-suited to a big data lens:** Mobile big data isn't only a function of smartphone penetration and consumer usage patterns. The data is also created by apps or other services working in the background. Technically speaking, its not that different from data created using the traditional Web. The difference is that consumers are just producing more of it as we shift our behavior to digital channels, leaving a trail of data documenting our movements and actions. Even when we are ostensibly not using our phones, we are still creating reams of data.
- **This data can be used to optimize and personalize mobile experiences:** Mobile big data can be used for a dizzying variety of purposes, but it is often used for the optimization and personalization of mobile services. App developers, for example, might use Flurry's analytics to improve their apps. Retention is a key metric for developers. Developers can compare their user retention numbers with all other apps and apps within their own categories, to gain insight into how they stack up, and what they might have to change to improve their numbers.
- **Or to help drive an explosion in mobile advertising:** Location data is an essential component of mobile big data - perhaps the primary data type that differentiates mobile from Web-based big data. Location data is expected to help transform the mobile advertising industry. The ability to deliver real-time hyper-local, targeted advertising represents a potentially momentous evolution of the ad market.

**Opportunities**

**SMARTER NETWORKS**

Smarter networks optimize the use of network resources and management of network traffic to improve the consumer experience. This enables the delivery of more compelling service offerings. The mix and behavior of services are not static. Every day, new services become available in app stores, and network traffic patterns for individual services change over time. It is therefore important to have solutions that capture relevant characteristics and excel in a changing environment. Big data analytics provides the means of such large-scale statistical analysis for observing and predicting traffic patterns, as well as quickly detecting trend shifts and anomalies. An example application of big data analytics for smarter networks is the prediction of optimal network parameter settings based on the characteristics of the traffic, to reduce battery consumption on the user device while not increasing network delay. Self-organizing networks (SON) efficiently use network-near analytics in planning, building and optimizing network resources. All SON automation – such as provisioning, configuration and commissioning – can be adapted to changes in the environment and traffic demand based on the insights gained from big data analytics. Finally, next-generation networks will provide ubiquitous mobile communication not only for people but also for connected things. This will add new challenges, such as the ability for the network to handle a massive number of connected devices at low cost and the need for increased energy performance on both the client and network side – challenges where big data analytics will play an important role.

**Conclusion :**

Organizations in every industry are trying to make sense of the massive influx of big data, as well as to develop analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information. When properly captured and analyzed, big data can provide unique insights into market trends, equipment failures, buying patterns, maintenance cycles and many other business issues, lowering costs, and enabling more targeted business decisions. To obtain value from big data, you need a cohesive set of solutions for capturing, processing, and analyzing the data, from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems.

Oracle Advanced Analytics is ideal for uncovering hidden relationships in big data sources. Whether you need to predict customer behavior, anticipate cross/up-sell opportunities, improve marketing campaign response rates, prevent churn, analyze "market baskets" to discover associations, patterns and relationships, leverage influencers in social networks, reduce fraud, or anticipate future demand, Oracle Advanced Analytics can help. For many years, companies developed data warehouses as the focal point for data to support decision making. This is changing as new data sources, platforms, and cloud-based services emerge. As a result, data is becoming more *federated*; that is, data is stored and accessed from multiple places. Adding to this trend are business units such as finance and marketing that have the business needs, resources, and political clout to acquire their own platforms, services, and tools. In many organizations, IT is losing some control over data management. This is not bad if it leads to more agility and better organizational performance. The downside, however, includes data silos that don't share data, data inconsistencies, inefficiencies in storing data, and duplication of resources. Organizations are accepting that data federation is going to exist, at least in the short to medium term, and are instituting greater controls over their data management practices. Some are putting more emphasis on data governance (e.g., data stewards, metadata management, and master data management). They are also creating BI or analytics centers of excellence to provide strategic direction for the use of data and analytics, prioritize projects, provide shareable resources, establish guidelines and standards, participate in tool selection, troubleshoot problems, and more.Organizations are gaining unprecedented insights into customers and operations because of the ability to analyze new data sources and large volumes of highly detailed data. This data is bringing more context and insight to organizational decision making. Success with big data is not guaranteed, however, as there are specific requirements that must be met. Organizations should start with specific, narrowly defined objectives, often related to better understanding and connecting with customers and improving operations. There must be strong, committed sponsorship. Depending on the project(s), the sponsorship can be departmental or at the senior executive level. The CIO is typically responsible for developing and maintaining the big data infrastructure. For some companies (e.g., Google), alignment between the business and IT strategies is second nature because big data is what the business is all about. For others, careful consideration needs to be given to organization structure issues; governance; the skills, experiences, and perspectives of organizational personnel; how business needs are turned into successful projects; and more. There should be a fact-based decision-making culture where the business is run by the numbers and there is constant experimentation to see what works best. The creation and maintenance of this culture depends on senior management. Big data has spawned a variety of new data management technologies, platforms, and approaches. When used in conjunction with open source tools such as Hadoop and MapReduce, this powerful analytic solution delivers everything you need to acquire, organize, analyze and maximize the value of big data within the enterprise while fulfilling fundamental requirements for minimizing data movement, leveraging existing skill sets, and ensuring high levels of security.

# Network Security: A modified Playfair cipher based on Sparse Matrix

**Kiran Tikar**[1]
Assistant Professor,
Department of Computer Science,
Christ College,Pune,India
tikar.kiran@gmail.com

**Dnyaneshwar Tikar**[2]
Sr. Software Engineer,Cybage
Pune,India
dstikar@gmail.com

**Sandeep Patil**[3]
Software Engineer,
Bharat Forge
Pune,India
skpatil.987@gmail.com

**Abstract**

***Playfair cipher is the form of block cipher which operates on block of characters encrypting and decrypting two characters at a time. The plain text is converted to cipher text and vice versa using key and matrix rules of traditional playfair cipher. The traditional Playfair cipher supports 5 X 5 matrix accommodating twenty five uppercase alphabets. To overcome this drawback, various authors have proposed extended Playfair cipher. This paper deals with designing a new technique in playfair cipher.This paper offers a technique to send information over the network consisting of 3 parts(i.e. 1. ASCII value of 1st character in the key matrix, 2. Key details 3. cipher text ). Here, Key matrix can be created in 64! Ways for enhancing security and increasing the difficulties of a crypt-analyst. This techniques makes use of Sparse matrix and matrix transpose concept for hiding the key. The traditional Playfair cipher rules are used for encryption and decryption of text. Here the text can contain spaces also.***

**Keywords:** ***cryptography; encryption; decryption; playfair cipher; security; sparse matrix; transpose of matrix, ASCII values.***

## I. Introduction

In today's internet world, the information/data has become essential to both individuals and organizations. The data must be protected from unauthorized user while it is being transmitted over the network.

Cryptology is the combination of cryptography and cryptanalysis where crypto has come from the Greek word "kryptos" means something hidden not revealed. Cryptography is the science of writing in secret code and is an ancient art.[18]

Cryptography is categorized in substitution technique and transposition technique. It is used for secure communication in the large network. Substitution is the way of substituting any alphabet, number; special characters at the place of plaintext. Encryption is the idea of converting the real message into scramble message while decryption is just the opposite of it. [17]

The Playfair cipher or Playfair square or Wheatstone-Playfair cipher or Wheatstone cipher is a manual symmetric encryption technique and was the first literal digram substitution cipher. The scheme was invented in 1854 by Charles Wheatstone, but bears the name of Lord Playfair who promoted the use of the cipher. [17]

It was used for tactical purposes by British forces in the Second Boer War and in World War I and for the same purpose by the British and Australians during World War II. This was because Playfair is reasonably fast to use and requires no special equipment - just a pencil and some paper. A typical scenario for Playfair use would be to protect important but non-critical secrets during actual combat. By the time that the enemy cryptanalysts could break the message, the information would be useless to them. During the Second World War, the Government of New Zealand used it for communication among New Zealand, the Chatham Islands, and the coast watchers in the Pacific Islands [17]

## II. Literature Review

The traditional **Playfair Cipher** is the polygraphic subsititution. The technique encrypts pairs of letters (bigrams or digrams), instead of single letters as in the simple substitution cipher and rather more complex Vigenère cipher systems then in use. The Playfair is thus significantly harder to break since the frequency analysis used for simple substitution ciphers does not work with it. The frequency analysis of bigrams is possible, with 600 possible bigrams rather than the 26 possible monograms (single symbols, usually letters in this context), a considerably larger cipher text is required in order to be useful. The traditional Playfair consist of 5 X 5 matrix incorporating I & J are single position. For instance, if we select puzzle as the confidential keyword the matrix is shown in Table 2.1 [17]

**Table 2.1:** Original Playfair 5 x 5 Matrix

| | | | | |
|---|---|---|---|---|
| P | U | Z | L | E |
| A | B | C | D | F |
| G | H | J | K | M |
| N | O | Q | R | S |
| T | V | W | X | Y |

Here, each distinct letter is inserted in the matrix and repeated letters are ignored. The remaining places are filled with the rest of the alphabets. The original message is divided into the group of two letters each. The substitution occurs depending on the following three principles.

**Encryption Principle**

1. Just in case both the letters are in the same row, replace them with the letter on the right of the letter. If the letter is at the start, go to the next letter. If the letter is at the end of the row then move circular(i.e. move at the beginning of the row).

2. Just in case both the letters are in the same column, replace them with the letter below them. If the letter is at the bottom, go to the top of the column and use the letter to replace with bottom letter.

3. If neither of the alphabets lies in the same column nor same row, imagine creating a rectangle form and write the corners alphabets. [17]

**Encryption Process**

Ex: If **ballon** is the plaintext and **puzzle** is the keyword, the corresponding ciphertext will be created as follows.

1. Initially the plaintext string is converted into upper case format i.e. **BALLON**
2. If there are repeating letters then insert X in between the two repeated letters i.e. **BALXLON**
3. Create pairs of 2 letters each. If total no. of letters are odd then insert X at the end of the string. In the second step the string contains 7 letters (odd) hence insert X at the end of the string i.e. **BALXLONX** Now create pairs of 2. i.e. **BA LX LO NX**
4. Now by using the table 2.1 key matrix and applying playfair cipher principals for converting plain text into cipher text.
5. BA – they lie within the same row, hence by rule one the cipher text is **CB**.
6. LX – they lie within the same column, hence by rule two the cipher text is **DL**.
7. LO – they are neither within the same row or column. hence by rule three creat an imaginary rectangle considering L and O as corners and use other two corners alphabets as cipher text ($1^{st}$ row alphabet and $2^{nd}$ as column alphabet). Hence cipher text is **UR**.
8. NX – they are neither within the same row or column. hence by rule three creat an imaginary rectangle considering N and X as corners and use other two corners alphabets as cipher text ($1^{st}$ row alphabet and $2^{nd}$ as column alphabet). Hence cipher text is **RT**.
9. So the cipher text is **CB DL UR RT**.

**Decryption Principle**

1. Just in case both the letters are in the same row, replace them with the letter on the left of the letter. If the letter is at the start, go back to the end of the same row and take the letter to replace with start letter.

2. Just in case both the letters are in the same column, replace them with the letter above them. If the letter is at the top, go back to the bottom of the column and use the letter to replace with top letter.

3. If neither of the alphabets lies in the same column nor same row, imagine creating a rectangle form and write the corners alphabets [16]

**Table 2.2**: Playfair Encryption and Decryption

| | | | | |
|---|---|---|---|---|
| plaintext | BA | LX | LO | NX |
| method | 1 | 2 | 3 | 3 |
| Encryption | CB | DL | UR | RT |
| cipher text | CB | DL | UR | RT |
| method | 4 | 5 | 6 | 6 |
| Decryption | BA | LX | LO | NX |

**Drawbacks of traditional Playfair Cipher**

The original Playfair comprises of 5*5 grid in which 25 letters can be placed that to of uppercase so it cannot encrypt lowercase letters, whitespaces, different printable characters. Moreover one letter will be discarded due to 25 squares. This is the main downside so several new proposals have been discussed.

**III. Proposed algorithm for encryption**

Here, the matrix size is increased to 8X8 for enhancing security. The matrix incorporates special symbols, numbers and alphabets hence it's an alphanumeric matrix. This matrix is used for hiding the key. Matrix elements are stored in the increasing order of their ASCII values. Matrix can be created in 64! Ways. This algorithm can handle white spaces and numbers in the string. Here, white spaces are replaced by '!' character.

For transposition of plain text into cipher text the traditional playfair cipher principals are used.

**Table 3.1:** Matrix for hiding the key

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ! | “ | # | $ | % | & | ( | ) |
| * | + | , | - | . | / | 0 | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| : | ; | < | = | > | ? | @ | A |
| B | C | D | E | F | G | H | J |
| K | L | M | N | O | P | Q | R |
| S | T | U | V | W | X | Y | Z |
| [ | \ | ] | ^ | _ | ‘ | { | } |

**Encryption Process**

**STEP 1:**
**Ex:**The key is **PLAYFAIREXAMPLE**
The key can be represented in 64! ways (matrix size – 8X8) by applying the permutation rule.

$^{n}P_{r}$ = n!/(n-r)! [19]
$^{64}P_{1}$ = 64!/ (64-1)!
= 64!
= 1.26886932E89 (in scientific notation)

**STEP 2:** Create sparse matrix representation for the key using table 3.1
In numerical analysis, a sparse matrix is a matrix in which most of the elements are zero. By contrast, if most of the elements are nonzero, then the matrix is considered dense. The fraction of non-zero elements over the total number of elements in a matrix is called the sparsity (density).
Conceptually, sparsity corresponds to systems which are loosely coupled. The concept of sparsity is useful in combinatorics and application areas such as network theory, which have a low density of significant data or connections. Large sparse matrices often appear in scientific or engineering applications when solving partial differential equations.
When storing and manipulating sparse matrices on a computer, it is beneficial and often necessary to use specialized algorithms and data structures that take advantage of the sparse structure of the matrix. Sparse data is by nature more easily compressed and thus require significantly less storage. [17]

**Table 3.2:** Sparse matrix representation for key “PLAYFAIREXAMPLE”

| | | |
|---|---|---|
| 8 | 8 | 15 |
| 5 | 5 | 12 |
| 5 | 1 | 13 |
| 3 | 7 | 5 |
| 6 | 6 | 0 |
| 4 | 4 | 0 |
| 4 | 7 | 0 |
| 5 | 7 | 0 |
| 4 | 4 | 14 |
| 6 | 5 | 0 |
| 5 | 2 | 0 |

Matrix description:-
1. 1st row consist of no. of rows , no. of columns of the actual matrix and no. of characters in the key.
2. From 2nd row onwards each row gives location details about distinct characters in the key.
3. 2nd and subsequent rows consist of location of the character in (row , col )format. And the last value tells whether the character is repeated in the key or not. If repeated then at which location. Here, location starts from 0. If not repeated then the value 0 will be inserted in the last column.

**STEP 3:** Perform transpose of the matrix by interchanging row with columns

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 5 | 5 | 3 | 6 | 4 | 4 | 5 | 4 | 6 | 5 |
| 8 | 5 | 1 | 7 | 6 | 4 | 7 | 7 | 4 | 5 | 2 |
| 15 | 12 | 13 | 5 | 0 | 0 | 0 | 0 | 14 | 0 | 0 |

**Table 3.3:** Matrix after performing transpose

**STEP 4:**Create Key matrix by inserting distinct characters from the key first and then insert remaining characters, digits and finally all special symbols (in the increasing order of their ASCII value)

**Table 3.4:** Matrix used for generating cipher text

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P | L | A | Y | F | I | R | E |
| X | M | B | C | D | G | H | K |
| N | O | Q | S | T | U | V | W |
| Z | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | ! | “ | # | $ | % |
| & | ( | ) | * | + | , | - | . |
| / | : | ; | < | = | > | ? | @ |
| [ | \ | ] | ^ | _ | ‘ | { | } |

**STEP 5:** Ex: Plain text is **“Gold is inside the trunk”**
1. Convert plain text into upper case format i.e. **“GOLD IS INSIDE THE TRUNK”**
2. If there are repeating letters then insert ‘X’ in between the two repeated letters

3. Create pairs of 2 letters each. If total no. of characters including white space is odd then insert X at the end of the string. Insert '!' in place of white space i.e.

   GO LD !I S! IN SI DE !T HE !T RU NK

4. Now by using the table 3.1 key matrix and applying traditional playfair cipher principals for converting plain text into cipher text.
5. GO – they are neither within the same row or column. Hence by rule three create an imaginary rectangle considering G and O as corners and use other two corners alphabets as cipher text ($1^{st}$ row alphabet and $2^{nd}$ as column alphabet). Hence cipher text is **MU**.
6. LD – they are neither within the same row or column. Hence by rule three create an imaginary rectangle considering L and D as corners and use other two corners alphabets as cipher text ($1^{st}$ row alphabet and $2^{nd}$ as column alphabet). Hence cipher text is **FM**.
7. S! - they lie within the same column, hence by rule one the cipher text is **2***. And so on.
8. So the cipher text is
   **MU FM #Y 2* PU UY KF "S KR "S IV WX**

**STEP 6:** The encrypted message hence created consist of 3 parts

1. ASCII value of the first character in the key matrix. It will be required for decryption.
2. Key in the form of transpose of Sparse matrix (row wise)
3. Cipher text

Hence encrypted message is as follows:

**33** 8 8 15 5 5 12 5 1 13 3 7 5 6 6 0 4 4 0 4 7 0 5 7 0 4 4 14 6 5 0 5 2 0 **MU FM #Y 2* PU UY KF "S KR "S IV WX**

## IV Proposed algorithm for Decryption

**STEP 1:** Regenerate the key matrix from the first number in the encrypted text. First number represents the ASCII value of the $1^{st}$ character in the matrix. Matching ASCII value matrix will be retrieved from the pool of available combinations.

| | | |
|---|---|---|
| 8 | 8 | 15 |
| 5 | 5 | 12 |
| 5 | 1 | 13 |
| 3 | 7 | 5 |
| 6 | 6 | 0 |
| 4 | 4 | 0 |
| 4 | 7 | 0 |
| 5 | 7 | 0 |
| 4 | 4 | 14 |
| 6 | 5 | 0 |
| 5 | 2 | 0 |

**Table 3.5:** Matrix for hiding the key

**STEP 2:** Beginning with the second number of the received message, first the number series in the message is converted into 3 by 'n' matrix (since sparse matrix has only 3 row's and encrypted message consist of transpose of matrix hence 3 columns.)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 5 | 5 | 3 | 6 | 4 | 4 | 5 | 4 | 6 | 5 |
| 8 | 5 | 1 | 7 | 6 | 4 | 7 | 7 | 4 | 5 | 2 |
| 15 | 12 | 13 | 5 | 0 | 0 | 0 | 0 | 14 | 0 | 0 |

**Table 3.6:** Matrix regenerated from the message

**STEP 3:** Perform transpose of the retrieved matrix.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ! | " | # | $ | % | & | ( | ) |
| * | + | , | - | . | / | 0 | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| : | ; | < | = | > | ? | @ | A |
| B | C | D | E | F | G | H | J |
| K | L | M | N | O | P | Q | R |
| S | T | U | V | W | X | Y | Z |
| [ | \ | ] | ^ | _ | ' | { | } |

**Table 3.7:** Sparse matrix regenerated

**STEP 4:** Using above sparse matrix description regenerate key and create decryption matrix from it.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| P | L | A | Y | F | I | R | E |
| X | M | B | C | D | G | H | K |
| N | O | Q | S | T | U | V | W |
| Z | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | ! | " | # | $ | % |
| & | ( | ) | * | + | , | - | . |
| / | : | ; | < | = | > | ? | @ |
| [ | \ | ] | ^ | _ | ' | { | } |

**Table 3.8:**Decryption matrix

**STEP 5:** By applying the traditional playfair cipher principals convert cipher text into plain text.

## V Conclusion

This paper focuses on the modification of the original Playfair cipher and hiding the key in the sparse matrix format. The traditional Playfair Cipher method due to its drawback (i.e. if the key is retrieved by an unauthorized person then the message can be easily decrypted) is not secure. Through this paper the encryption process becomes more secure and flexible for long strings of data with numbers and white spaces.

**VI Future Enhancement**

The future work of proposed algorithm can concentrate on the encryption and decryption process in the playfair cipher for enhancing security of valuable data sent over the network.

**VII Reference**

[1] S. Basu and U. K. Ray, "Modified Playfair Cipher using Rectangular Matrix", International Journal of Computer Applications (0975 – 8887) vol. 46, no. 9, (2012) May.

[2] P. Murali and G. Senthilkumar, "Modified Version of Playfair Cipher using Linear Feedback Shift Register", IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 12, (2008) December.

[3] R. S. Bhadoria, D. Sahu and M. Dixit, "Proficient Routing in Wireless Sensor Networks through Grid Based Protocol", International Journal of Communication Systems and Networks, vol. 1, no. 2, (2012), pp. 104-109.

[4] K. Ravindra Babu, S. Uday Kumar, A. Vinay Babu, I. V. N. S. Aditya and P. Komuraiah, "An Extension to Traditional Playfair Cryptographic Method", International Journal of Computer Applications (0975 – 8887) vol. 17, no. 5, (2011) March.

[5] S. S. Srivastava and N. Gupta, "A Novel Approach to Security using Extended Playfair Cipher", International Journal of Computer Applications (0975 – 8887) vol. 20, no. 6, (2011) April.

[6] G. Agrawal, S. Singh and M. Agarwal, "An Enhanced and Secure Playfair Cipher by Introducing the Frequency of Letters in any Plain text", Journal of Current Computer Science and Technology, vol. 1, no. 3, (2011), pp. 10-16

[7] A. Alam, S. Ullah, I. Wahid and S. Khalid, "Universal Playfair Cipher Using MXN Matrix", International Journal of Advanced Computer Science, vol. 1, no. 3, (2011).

[8] S. S. Srivastava and N. Gupta, "Security aspects of the Extended Playfair cipher. In Communication Systems and Network Technologies (CSNT)", 2011 International Conference on IEEE, (2011) June, pp. 144- 147.

[9] S. S. Srivastava and N. Gupta, "Modified Version of Playfair Cipher by using 8x8 Matrix and Random Number Generation", Rajaram jaiswal, in Proceedings of IEEE 3rd International Conference on Computer Modeling and Simulation (ICCMS 2011).

[10] R. S. Bhadoria, "Security Architecture for Cloud Computing", Handbook of Research on Securing Cloud-Based Databases with Biometric Applications, IGI Global Publisher, USA, (2014), pp. 47-71.

[11] V. Verma, D. Kaur, R. K. Singh and A. Kaur, "3D-Playfair cipher with additional bitwise operation", In Control Computing Communication & Materials (ICCCCM), 2013 International Conference on IEEE, (2013), August, pp. 1-6.

[12] R. S. Bhadoria, M. Dixit and V. M. K. S. Jadon, "Enhancing Web Technology through Wiki-shell Architecture", In Proceedings of IEEE World Congress on Information and Communication Technologies, (2011) December, Mumbai, India.

[13] N. Chand and S. Bhattacharyya, "A Novel Approach for Encryption of Text Messages Using PLAY-FAIR Cipher 6 by 6 Matrix with Four Iteration Steps", International Journal of Engineering Science and Innovative Technology (IJESIT) vol. 3, no. 1, (2014) January, pp. 478-484.

[14] S. S. Dhenakaran and M. Ilayaraja, "Extension of Playfair Cipher using 16X16 Matrix", International Journal of Computer Applications (0975 – 888) vol. 48, no. 7, (2012) June.

[15] Swati Hans, Rahul Johari, Vishakha Gautam, "An Extended PlayFair Cipher using Rotation and Random Swap patterns," 5th IEEE International Conference on Computer and Communication Technology, (2014).

[16] R. S. Bhadoria and R. Jaiswal, "Competent Search in Blog Ranking Algorithm", In Proceedings of Springer International Conference on Computation Intelligence & Information Technology, (2011) June, Pune, India.

[17] https://en.wikipedia.org/wiki/Playfair_cipher

[18] http://www.garykessler.net/library/crypto.htm

[19]http://tutors4you.com/permutationcombinationtutorial.htm

# Predicting Patients Curability using Data mining Signatures

**S. Sreeji ME.;MBA**
Asst. Prof Dept of Computer Application
MGR University, Chennai
sreeji.cse@gmail.com

***Abstract-***

**Providing a Graphical representation of Individuals predicted health history by matching the history of the patient with related information's in the system is the core idea of the project. In this project, temporal string pattern matching is one of the core algorithmic approach used for matching the patient's history with individual's symptoms to predict the foredooming health issues. Restrictions/Filtering of data will be done to slash out the appropriate disease related matching through temporal string pattern algorithm. Pattern marching will be done with back ground data fetching with the help B Tree searching concept and regaining back from the filtered criteria to the original dataset using Back Tracking algorithm. Focusing towards the process of checking constraints such as Data constraints, Drawing constraints and Interface constraint, enables the possibility of obtaining the exact filter criteria thereby matching the exact issues from patients history.**

## I.INTRODUCTION

Data mining research has successfully produced numerous methods, tools, and algorithms for handling large amounts of data to solve real-world problems. Traditional data mining has become an integral part of many application domains including bioinformatics, data warehousing, business intelligence, predictive analytics, and decision support systems. Primary objectives of the data mining process are to electively handle large-scale data, extract actionable patterns, and gain insightful knowledge. Because social media is widely used for various purposes, vast amounts of user-generated data exist and can be made available

For data mining. Data mining of social media can expand researchers' capability of under- standing new phenomena due to the use of social media and improve business intelligence to provide better services and develop innovative opportunities. For example, data mining techniques can help identify the inuential people in the vast blogosphere, detect implicit or hidden groups in a social networking site, sense user sentiments for proactive planning, develop recommendation systems for tasks ranging from buying septic products to mak- ing new friends, understand network evolution and changing entity relationships, protect user privacy and security, or build and strengthen trust among users or between users and entities. Mining social media is a burgeoning multidisciplinary area where researchers of different backgrounds can make important contributions that matter for social media research and development.

The objective of this tutorial is to introduce social media, data mining, and their con- uence|mining social media. We attempt to achieve the goal by presenting representative and interesting research issues and important social media tasks based on our experience and research. This tutorial _rst reviews data mining, social media and its types, and the importance of social media mining. In x2, we brief introduce representative issues in social media mining. In x3, we highlight the impact of social media mining using three examples based on our current research. How social media mining is applied in some real-world applications two projects on humanitarian assistance and disaster relief (HADR) carried out in the Data Mining and Machine Learning Laboratory (DMML) at Arizona State University (ASU).

The patient history needs to be fed into the system. Appropriate information about the relevant data will be matched from the system. As usual, our system is provided with the Factor based temporal matching algorithm in matching the relevant contents from the historical databases in order to predict the foredooming issues of an individual. Data's are tuned up to a very lower level of temporal information to be maintaining high accuracy of information.

## II.RELATED WORK

The main goal of the paper[ ] is to show how valuable is to perform log query mining, by presenting several different applications of this idea combined with

standard usage mining. In this we have a tendency to present two applications of the computer program log. First, victimisation the query distribution, we have a tendency to present an inverted file organization that has three levels: pre computed answers, main and secondary memory indexes. Second, we have a tendency to present an algorithmic rule that uses queries and clicks to enhance ranking that captures linguistics relations of queries and web content.

In this paper the data mining supported neural networks is researched very well. The key technology and ways that to attain the data mining supported neural networks also are mentioned during this paper. Data mininging process is composed by three main phases like knowledge preparation, data processing, expression and interpretation of the results.

This paper introduces a link-analysis procedure for discovering relationships in a very relational database or a graph, generalizing each straightforward and multiple correspondence analyses. it's supported a random-walk model through the database defining a Markov chain having as many countries as elements within the database.

This paper proposes to use stochastic complementation for extracting a sub graph containing the elements of interest from the original graph and also introduces a kernel-based extension of the basic diffusion map for displaying and analyzing the reduced sub graph.

## III.PROPOSED WORK

In our proposed system, the patient history needs to be fed into the system. Appropriate information about the relevant data will be matched from the system. As usual, our system is provided with the Factor based temporal matching algorithm in matching the relevant contents from the historical databases in order to predict the foredooming issues of an individual. Data's are tuned up to a very lower level of temporal information to be maintaining high accuracy of information. The existing system doesn't have a high level filter for criteria's such as age and gender. Whereas proposed system considers these criteria for predicting the possibility of disease for a patient.

Accuracy of data is maintained up to nanoseconds that are highly important for the patients. Age and gender based criterion match predicts more accurate disease than any other factor considered.

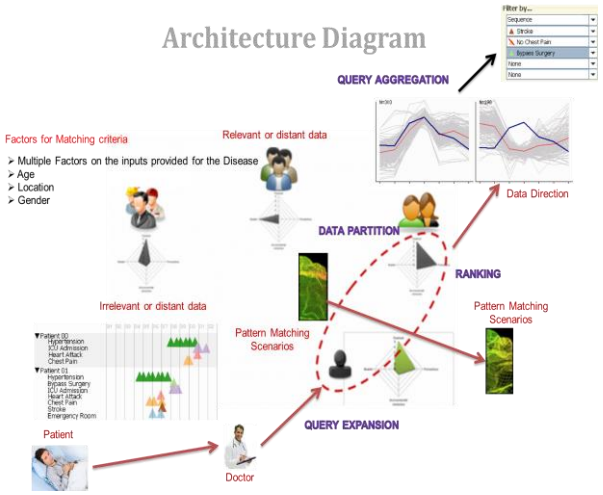



fig 3.1 Architecture diagram of the proposed model

**Use case Diagram:**

A use case is a set of scenarios that describing an interaction between a user and a system. A use case diagram displays the relationship among actors and use cases. The two main components a user or another system that will interact with the system modelled. A use case is an external view of the system that represents some action the user might perform in order to complete a task.

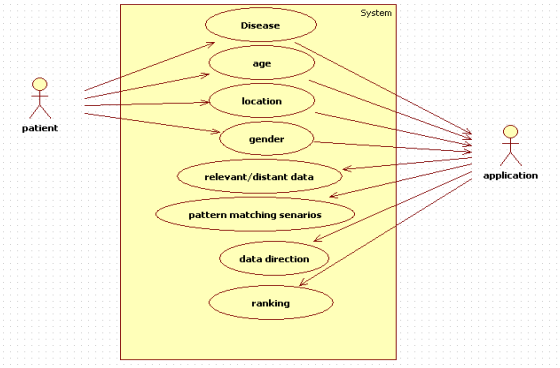



fig 3.2 Usecase diagram of the proposed model

**Temporal Pattern Search Algorithm**

Set of arrays, where each array contains all events of the same type, sorted by time stamps. Particular item in the pattern using a binary search over the appropriate arrays. Considerably more expensive per item, it allows TPS to skip many unnecessary events in personal histories.

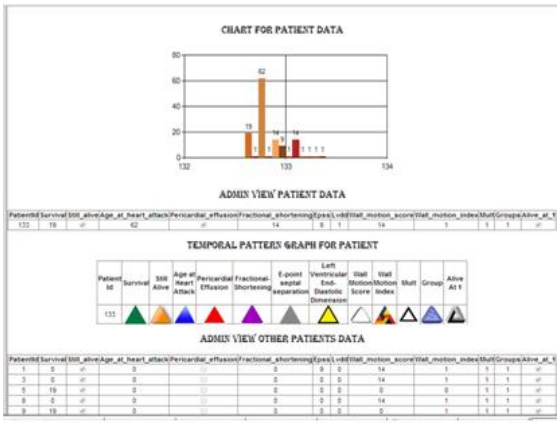



fig 3.3 **Temporal Pattern Graph for Patients**

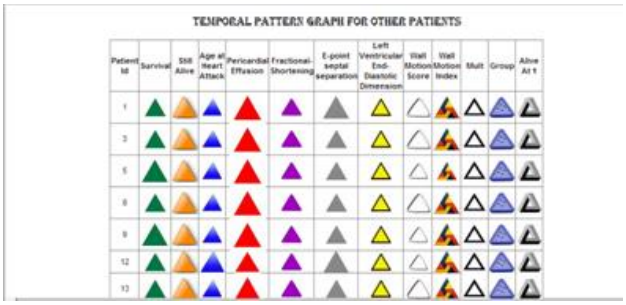



fig 3.4 **B-Tree Algorithm**

Tree data structure that keeps data sorted and allows searches, sequential access, insertions, and deletions in logarithmic time. It was difficult to access a large amount of data from the secondary memory. Many of the algorithms were introduced to make our search very fast, to access the required data from the secondary memory B-trees are more effective and faster B-trees. The payoff of the B-tree insert and delete rules are that B-trees are always "balanced". Searching an unbalanced tree may require traversing an arbitrary and unpredictable number of nodes and pointers.

Binary search is an algorithmic approach for locating an object in the large dataset. The pattern gets matched based on the key that maps with object in the datasets. The pattern is the regular and intelligible form or sequence discernible in the way in which something happens or is done. In this module, a pattern for the patient is formulated and matched with the previous patient record.

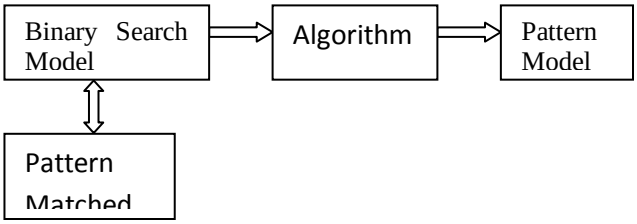



fig 3.5

**Backtracking Algorithm**

Partial candidate solution" and a relatively quick test of whether it can possibly be completed to a valid solution. Traverses this search tree recursively, from the root down, in depth-first order. The algorithm checks whether c can be completed to a valid solution.

Problem searching for a set of solutions or which require optimal solution can be solved by using Backtracking method. To apply backtracking method the problem must be expressible as an n-tuple($_{x1,x2\cdots Xn}$),where the xi are chosen from some finite set $_{si,}$ Backtracking is a fascinating programming technique which can be used for optimizations, solving puzzles (jigsaw and other) and creating computer opponents in thinking games like chess. Backtracking is a general algorithmic technique that considers searching every possible combination in order to solve the problem recursively. In this module, the individual patient's details are grouped recursively with their factors and the matching symptoms.

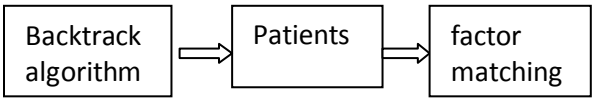



fig 3.6

Performance Analysis

**Performance means the accomplishment of a given task measured against preset known standards of accuracy, completeness, cost, and speed. Performance Analysis module is the final module that checks the correctness and accuracy of the system factor's consideration of the patient and their prediction of survival. This Analysis of performance is visualised graphically**

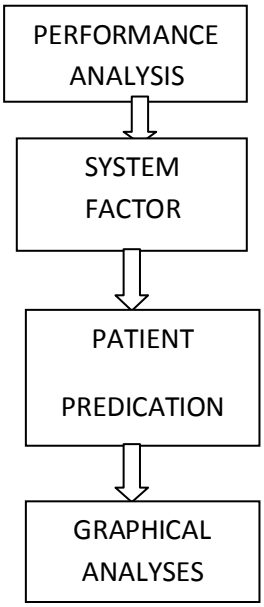



fig 3.7

IV. CONCLUSION

In this paper, the patient history is fed into the system. Appropriate information about the relevant data is matched from the system. Our system is provided with the Factor based temporal matching algorithm in that matched the relevant contents from the historical databases and predicted the foredooming issues of an individual. Data's are tuned up to a very lower level of temporal information that maintains high accuracy of information.

**FUTURE ENHANCEMENT:**

The patient history can be viewed in terms of images that may help in mining the data more efficiently. The process can be extended in terms of cloud so that the entire records of all the patients can be accessed from anywhere.

V. REFERENCES

[1] B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen, "Detect
And Track Latent Factors with Online Nonnegative Matrix
Factorization," Proc. 20th Int'l Joint Conf. Artificial Intelligence,
pp. 2689-2694, 2007.

[2] F.R.K. Chung, Spectral Graph Theory. Am. Math. Soc., 1997.

[3] C. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative
Matrix Factorizations," IEEE Trans. Pattern Analysis and Machine
Intelligence, vol. 32, no. 1, pp. 45-55, Jan. 2010.

[4] M. Dong, "A Tutorial on Nonlinear Time-Series Data Mining in
Engineering Asset Health and Reliability Prediction: Concepts,
Models, and Algorithms," Math. Problems in Eng., vol. 2010, pp. 1-
23, 2010.

[5] J. Eggert and E. Korner, "Sparse Coding and NMF," Proc. IEEE
Int'l Joint Conf. Neural Networks, vol. 2, pp. 2529-2533, 2004.

[6] W. Fei, L. Ping, and K. Christian, "Online Nonnegative Matrix
Factorization for Document Clustering," Proc. 11th SIAM Int'l
Conf. Data Mining, 2011.

[7] C. Fe´votte and J. Idier, Algorithms for Nonnegative Matrix
Factorization with the Beta-Divergence, arXiv:1010.1763, 2010.

[8] P.O. Hoyer, "Non-Negative Matrix Factorization with Sparseness
Constraints," J. Machine Learning Research, vol. 5, pp. 1457-1469,
2004.

[9] P.O. Hoyer, "Non-Negative Sparse Coding," Proc. 12th IEEE
Workshop Neural Networks for Signal Processing, 2002.

[10] Y.R. Ramesh Kumar and P.A. Govardhan, "Stock Market
Predictions—Integrating User Perception for Extracting Better
Prediction a Framework," Int'l J. Eng. Science, vol. 2, no. 7, pp. 3305-3310, 2010.

# Review on Artificial Intellegence

**Amit.R.Atole**
Student, vpasc college, Baramati.
atole.amit@gmail.com

**ABSTRACT**

**Artificial intellegence is study of how to make computer do things which at the moment, people do better. AI methods can be divided into two broad categories.**
**1) Symbolic AI: This focuses on development of knowledge-based system.**
**2) Computational intelligence: This includes such methods as neural networks, fuzzy systems and evolutionary computing.**

**AI helps to improve our way of life, machines will able to do jobs that require detailed instructions. AI works with pattern matching methods which attempt to describe objects, events or processes in terms of their qualitative features and logical and computational relationship.**
**AI related to similar task of using computer to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.**
**Keywords:** Artificial Intelligence

## INTRODUCTION

Artificial intellegence is the branch of computer science which deals with intelligence of machine where an intellegent agent is a system that takes action which maximizes its chances of success.
It is the study of idea which enables computers to do the things that make people seem intellegent.
The central principal of AI includes Reasoning, Knowledge, Learning, Communication, Perception and the ability to move and manipulate the objects. It is the science and engineering of making intelligent machines, especially intelligent computer programs.

## LITERATURE REVIEW

Researches have worked in terms of:
1) The theoretical properties of the algorithms used for qualitative temporal reasoning.
2) Development of axioms that give sound and complete axiomazation for the logic of reasoning.
The literature on complexity of reasoning relates to spatial congruence and expressive description logics.
This review of AI is prepared with the help of:

a) Reference Book.
b) Internet.
c) Group Discussion.
d) Newspaper.

## PROBLEMS IN AI

Now AI research became more advanced and techniques for handling larger amount of world knowledge were developed. Some progress was made on the task just described and new task could reasonably be attempted. These include perception (vision and speech), natural language understanding and problem solving in specialised domains such as medical diagnosis and chemical analysis.

The problem areas where AI is now flourishing most as a practical discipline as primarily the domains that require only speacialised expertise without the assistance of commonsense knowledge. AI problems span a very broad spectrum. They appear to have very little in common except that they are hard. A well defined problem can be described by:-
1) Initial state.
2) State space: All states reachable from initial by any sequence of actions.
3) Path: sequence through state space.
4) Goal test: Test to determine if at goal state.

## CONCLUSION

In its short existence, AI has increased understanding of the nature of intellegence and provided an impressive array of application in a wide range of areas.
It has sharpened understanding of human reasoning and of the nature of intelligence in general. At the same time, it has revealed the complexity of modeling human reasoning providing new areas and rich challenges in feature.

## REFERENCES

1)http://en.wikibooks.org/wiki/computer_science:artificialintellegence
2)http://www.google.co.in
3)http://machinelikeus.com/news/whatever-happened-artificial-intellegence
4)http://www.library.thinkquest.org

# Implementation Of Protus2.0 By Ontology For Advanced E Learning

**Japala Saikrishna Rao***
Software Engineer
Paxterra Software, Banglore

**ABSTRACT**

***Protus2.0 E_ learning environment is used in learning process for decision making, communication and problem-solving. The semantic web used in real-time annotations of these ontology for debriefing of the students, student self study and better guidance of the learning methodologies of mentors. A ever-present methodology environment will provides an interoperable, pervasive, and flawless learning structural design to connect, fix together, and share three major scope of learning resources such as learning collaborators, learning services, and learning contents. ever-present learning is characterized for identify right learning collaborators, right learning contents and right learning services in the right place at the right time. it promotes the achievement of practical knowledge as well as decision production, communication, and problem solving with help of ontological Skills-based learning environments. These environments are a important to give feedback about the students from the practically conducted sessions and comments of learners actions can notify to the assessment of their quality process. And also, those learning environments are helping to the researchers for better understand the learning process. The proposed system examined study environments. Also, achieve the better student improving the knowledge as well as feedback of the understanding of learning environments; we proposed a new mechanism such as semantic based approach with the ontology is used. Index Terms: semantic web, ontology mapping methods, fuzzy logic, fuzzy grouping, k-means algorithms, tutorial systems.***

## I .INTRODUCTION

To provide an effective learning environments to the students (for self study) using learning materials such eBooks. To offer an efficient learning materials to the students and researchers with help of semantic based annotations of the recording of simulated environments and with effective mapping method. Ontology is used to promote the gaining of practical skills as well as team working, decision making, communication and problem solving. They can be incorporated into opinion of student performance, which brings a condition that the approaches for estimation and feedback need to be , valid, reliable, feasible, educational, and of course acceptable to practitioners. Through the ontology, the student experiences are designed to be exactly as they would experience in the workplace in real time. By this ontology method student can analyze himself whether he is good in a particular domain and it is very useful for the persons who are not able to pay the amount for technical educations. in this approach the questionnaires all are related to the frequently asked questions with the correct approach.

## II. RELATED ARTICLES

Ballan et al. (2011)has proposed approaches for robust detection and representation of spatio-temporal interest points and motion features, modelling of events and approaches to represent domain knowledge and contextual information of activities and actions. The methods are used to several different domains. The real time based event detection is very complicated. This research does not discuss about any observer's audio. Also, the location identification of objects is very difficult. Tiropanis et al. (2009) developed a paper for establishes the availability of semantic tools and services , searching and matching, repository and collaborative annotations &semantic infrastructure development. The relate things for semantic tools and services are knowledge, teaching and Institutional challenge for higher education. The tool does not provide an efficient semantic link between the other videos. McDonald et al. (2008) developed a paper for how Semantic Grid technologies can be used to generate enhanced tools for data collection that provide enabling technologies for interdisciplinary work, thereby enhancing the capacity to address substantive social science research. The research work focus on only simulation environments. The research work is not suitable for visual and textual videos. Lindgren et al. (2008) developed a paper for describe DIVER, a platform designed to solve a set of core challenges we have identified in supporting video collaboratories. Characterize five Collaboration Design Patterns (CDPs) for video-based practices. CDPs used for characterizing interaction patterns. The DIVER video platform embodies a new kind of communication infrastructure for video conversations by providing persistent and searchable records of video pointing activities by participants to specific time and space moments. The learning technologies community to refine and advance our conceptualizations of collaboration design patterns for video platform uses in research and education and the creation of systems that supports the important needs for video conversation in the work practices of educator and researchers. J.A Muras (2006)Proposed a novel taxonomy of pervasive healthcare systems. The taxonomy is using hierarchy structure of the properties of pervasive healthcare

systems and can be used as a framework for system categorization. The taxonomy identify a set of fundamental properties, that properties are used in a system to be describe according to its user's kind, its purpose and environment of use, as well as the technologies implementation.
The properties of invasive healthcare system are straight in hierarchical manner initial from the root of the taxonomy. The root describes the relationships between all seven main feature categories. Novel taxonomy is based on the ICF.ICF provide standard language and a structure for the explanation of health and disability. Novel taxonomy was considering the attributes of the users in a number of different ways. Taxonomy was extended for novel system properties without reorganisation of existing structure. K.R. Page et al. (2005) developed a paper for hypertext and knowledge based tools .this Paper supporting e-Science, also explores the similarities and application of CoAKTinG (Collaborative Advanced Knowledge Technologies in the Grid) technologies as part of a human-centred design approach to e-Learning. Introduced the tools that have been developed by the COAKTING project and identified how they are typically used in meetings, and in support of collaborative science in the Semantic Grid. Includes the application of Semantic Web technologies to help with the initial identification of the members of the virtual organisation.

## III SUMMARY OF EXISTING SYSTEM

This is used to establish a complete attribution trail through to learned output, enabling researchers to follow back to the original data. The Task Computing project involves Semantic Web technologies (Resource Description Framework, Ontology Web Language) and Web Services (Simple Object Access Protocol, Web Service Description Language) to pervasive computing, aiming to "fill the gaps between tasks and services." Learners will see the tasks are possible in their current context and are assist in designing composite tasks from easy tasks, they can be reused. Disadvantages of learning environments are cannot offer effective way of learning. Time usage is high. Issues in accessing the learning related material to by the research team. The collection, storage, and dissemination of staff and student data could not be anonymized.

## IV PROPOSED SYSTEM

A proposed system constructed an ontology that contains all the entities describing the, sessions, and participants. This ontology allowed the basic online ontological framework to be independent of the specific structure of ontology. The system takes a domain ontology that includes the following conditions of tutorial libraries such as a series of workshops, observational sessions, and discussion groups. Identified the types of annotation (manual textual, location based are shows how to create ontology's with semantic based), the constructed ontology representing the range of annotations applicable in the scenarios. The ontology consists the basis for the sharing of an knowledge is to be developed. For an ontology for successful, it is important to design prompts that are easily identifiable and familiar to the users (students). In this protus system by using the fuzzy logic algorithms the user has to qualify the particular level of the learning system then only he will enter into the second level after cleared in second level third level like that he will achieve a total knowledge about the particular tutorial in this paper we are going to implement for java language. The below figure shows the advanced e-learning in proposed system. This diagram describes the total flow of the project and the approach how is the approach for the proposed system. And by this architecture we can assume the proposed system very clear manner and this is mostly used in ontology system this ontology will be very used in the intelligent tutorial systems. And by this ontology system all the systems are merging and it is used for using the multi database at a time.

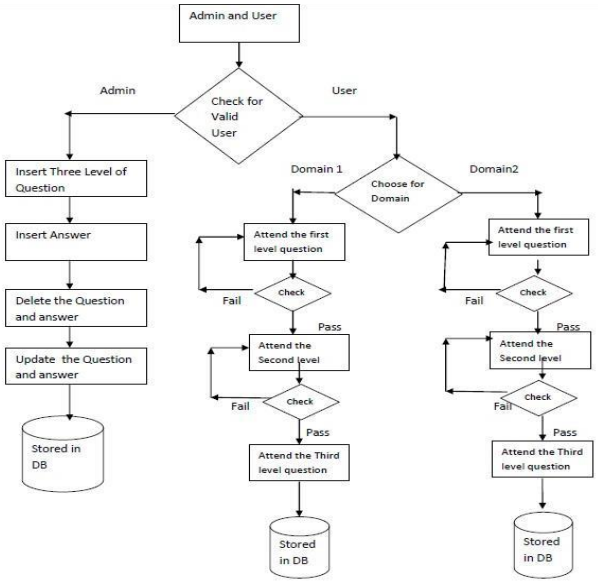



Figure 1: functional architecture

In above diagram it is clearly understood that first learner should enter into the admin module after that only he will enter into the learning the particular courses if he clears the starting level he is eligible to enter into the next level like that till the last module this process will continued. V IMPLEMENTATION In this implementation of protus 2.0 ontological system, the detailed information regarding how the tutorial system exactly in these by using four modules this system was implemented by using the ontology and fuzzy logics .

### *A. administration :-*

in the implementation part admin is the authorized person to maintain on line exam, admin has the responsibilities of preparing the questions and depends upon the level from easy level to hard level preparation of questions will takes

place here by using intelligence tool the questions are generated depends upon the user automatically these questions will be generated ontology rules are applied in this phase
Admin will stores all the questions and answers in database. Depends upon the particular question the answer will be evaluated these above all things are to be in the admin phase. Here by intelligence system the person has failed in particular level he will not be enter into the next module.

***B. user system***

The user system has the certain rules first he has to learn the particular tutorial he has to register in administration module he has to give all the details like name ,security questions like that all things need to give in this phase the learner or user must and should be aware about the system generation of the questions and navigation systems he has to know all the tutorial system for that by using tutorial system and mapping methods we are assigning the users to the administrator.

***C. online question paper***

The online question paper phase is responsible for generating the marks for the correct option and also in this phase some particular time has to be given if the user will edit the correct answer are not it doesn't bother after the finishing of the time slot the question paper will disappears these all things are done in this Phase.

***D. result evaluation***

In this phase the evaluation of result takes place here in this the advanced result manipulations like graph based evaluation takes place and depends upon the levels this results will be displayed by this user will understood in which level he is lagging.

## VI. PERFORMANCE EVALUTION

In the existed system the ontology rules are not that much implemented and here in this tutorial we are adding the online compiler by this user can easily achieve the programming concepts actually in this tutorial particularly implemented for java tutorial system. By this proposed system we can achieve the following things. A. Coherent navigation B. flexible entry points C. concept. Matching D. reasoning And the result evaluation will be done in a advanced manner .user will achieve his goal. The above fig 1.2 shows how the proposed system and existing system performed. The proposed system contained the symbol annotation so the system having more performance than existing system. From the graph we are taking the y-axis as number of Exams conducted. By writing the we can easily analyse the online exam with symbol annotation which was proposed in the proposed system.

## VII CONCLUSION

The protus 2.0 main aim is to provide a flexible tutorial system by this user can evaluate his status in particular field by this we are using advanced technologies like ontology and fuzzy rules.

## VIII REFERENCES

[1] M.J. Weal, D. Michaelides, K.R. Page, D.C. De Roure, E. Monger, and M. Gobbi, "Semantic Annotation of Ubiquitous Learning Environments," IEEE Trans. Learning Technologies, vol. 5, no. 2, pp. 143-156, Apr.-June 2012.

[2] T. Tiropanis, H. Davis, D. Millard, and M. Weal, "Semantic Technologies for Learning and Teaching in the Web 2.0 Era—A Survey," Proc. Society On-Line (WebSci '09), http://eprints.ecs. soton.ac.uk/17106, 2009.

[3] J.W. McDonald, M.O. Gobbi, D. Michaelides, E. Monger, M.J.Weal, and D. De Roure, "Grid-Enabled Data Collection and Analysis: Semantic Annotation in Skills-Based Learning," Proc.Fourth Int'l Conf. E-Social Science, June 2008.

[4] M.J. Weal, D.T. Michaelides, D.C. De Roure, M. Gobbi, E. Monger, and J.W. McDonald, "Semantic Annotation in Ubiquitous Healthcare Skills-Based Learning Environments," Proc. Workshop Semantic Web in Ubiquitous Healthcare (ISWC), Nov. 2007.

[5] K.R. Page, D.T. Michaelides, S.J.B. Shum, Y.-H. Chen-Burger, J.Dalton, D.C. De Roure, M. Eisenstadt, S. Potter, N.R. Shadbolt, A.Tate, M. Bachler, and J. Komzak, "Collaboration in the Semantic Grid: A Basis for E-learning," Applied Artificial Intelligence, vol. 19, nos. 9/10, pp. 881-904, Nov. 2005.

[6] O. Lassila and M. Adler, "Semantic Gadgets: Ubiquitous Computing Meets the Semantic Web," Spinning the Semantic Web, D. Fensel, ed., MIT, pp. 363-376, 2003.[7] T. D. of Health, "Delivery 21st Century IT Support for the NHS: National Strategic Programme," technical Stationary report Office, June 2002.

# Secure Information Using Cryptography Based Authentication methods

**Prof.Jankar Jayshree Prakash Lecturer**
Prof. Ramkrishna More Arts, Commerce and Science College, Akurdi ,Pune-44
Mo. No 9730651115
Email address – jay19jankar@gmail.com

**ABSTRACT**

**Cryptography is the science of coding and decoding message so as to keep these message secure. In this study we discussed main cryptographic services as well as specific authentication services actually based on Cryptographic services. Authentication is an absolutely essential element of a typical security model. It is the process of confirming the identification of a user that is trying to access resources. Now a days information security and confidentiality play vital role in many sectors. This study talks about the simplest authentication method as well as the available biometric authentication methods**

**Keywords**- Information Security , Cryptography , Authentication methods, Biometric Authentication

## 1.Introduction

Information security means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction. Information security is concerned with the confidentiality, integrity and availability of data regardless of the form the data may take: electronic, print, or other forms. In a simplest form encryption is to convert the dada in some unreadable form. This help in protecting the privacy while sending the data from sender to receiver .Encryption ensures that only authorized users (confidentiality) have access to accurate and complete information (integrity) when required. Cryptography provides the information security for other useful applications such as in encryption, message digests, and digital signatures. The length and strength of the Cryptography keys are considered an important mechanism. The keys used for encryption and decryption must be strong enough to produce strong encryption. They must be protected from unauthorized users and must be available when they are needed. Cryptography also contributes to computer science, particularly, in the techniques used in computer and network security for access control and information confidentiality.

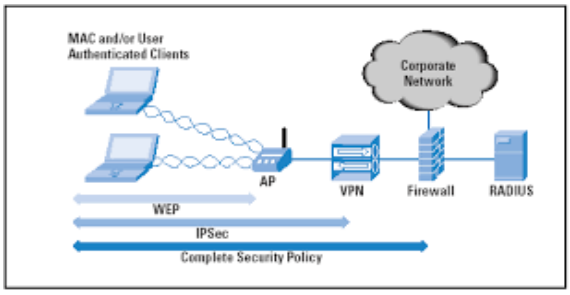



Figure1 :Information Security

## 2. Cryptography

Cryptography includes techniques such as microdots, merging words with images, and other ways to hide information in storage or transit. However, in today's computer-centric world, cryptography is most often associated with scrambling plaintext (ordinary text, sometimes referred to as clear text) into cipher text (a process called encryption), then back again (known as decryption). Individuals who practice this field are known as cryptographers.

Modern cryptography concerns itself with the following four objectives:

1) Confidentiality -the information cannot be understood by anyone for whom it was unintended.

2) Integrity -the information cannot be altered in storage or transit between sender and intended receiver without the alteration being detected.

3) Non-repudiation -the creator/sender of the information cannot deny at a later stage his or her intentions in the creation or transmission of the information.

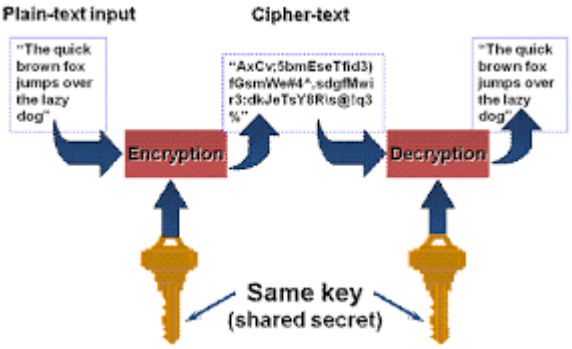



Figure 2: Cryptography scheme

## I.Types Of Cryptographic Algorithms

There are several ways of classifying cryptographic algorithms. For purposes of this paper, they will be categorized based on the number of keys that are employed for encryption and decryption, and further defined by their application and use. The three types of algorithms that will be discussed are (Figure 1):

- Secret Key Cryptography (SKC): Uses a single key for both encryption and decryption
- Public Key Cryptography (PKC): Uses one key for encryption and another for decryption
- Hash Functions: Uses a mathematical transformation to irreversibly "encrypt" information

### 1.Symmetric Key Cryptography (Secret Key Cryptography)

**a)**.Same Key is used by both parties

**b)**. Simpler and Faster

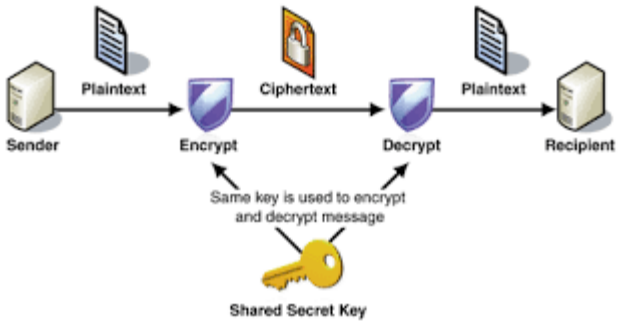



Figure3: Symmetric Key Cryptography

### 2 .Asymmetric Key Cryptography (Public Key Cryptography)

**a)**.Two different keys are used Users get the Key from an Certificate Authority.

**b)**. Authentication in asymmetric cryptography is more secured but the process is relatively more complex as the certificate has to be obtained from certification authority.

Figure 4: Asymmetric Key Cryptography

### 3.HashFunction

**a)**.Uses mathematical transformation to irreversibly encrypt information.

**b)**.It is a one-way encryption

**c)**.uses no key for encryption and decryption

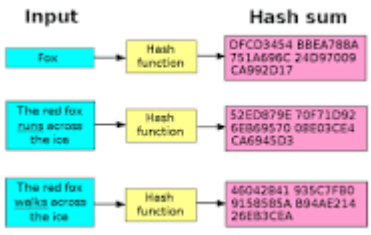



Figure 5: HashFunction

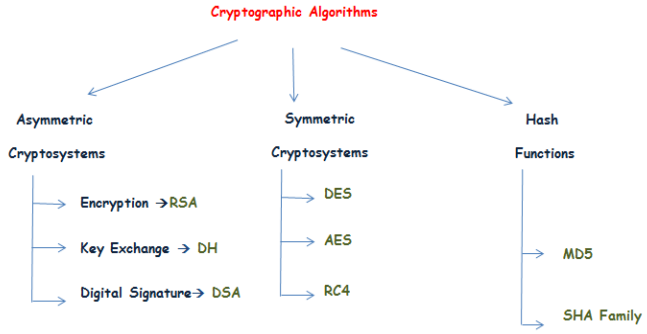



Figure 6: Main branches of Cryptography

### 3. Authentication methods

#### I .Password authentication

Most of us are familiar with password authentication. To log onto a computer or network, you enter a user account name and the password assigned to that account. This password is checked against a database that contains all authorized users and their passwords. In a Windows 2000 network, for example, this information is contained in Active Directory.

To preserve the security of the network, passwords must be "strong," that is, they should contain a combination of alpha and numeric characters and symbols, they should not be words that are found in a dictionary, and they should be relatively long (eight characters or more).Password authentication is vulnerable to a password "cracker" who uses a brute force attack (trying every possible combination until hitting upon the right one) or who uses a protocol "sniffer" to capture packets if passwords are not encrypted when they are sent over the network.

Figure 7: Password Authentication

#### II. Smart card authentication

Smart cards are credit card-sized devices that hold a small computer chip, which is used to store public and private keys and other personal information used to identify a person and authenticate him or her to the system. Logging onto the network with a smart card requires that you physically insert the card into (or slide it through) a reader and then enter a Personal Identification Number (PIN) in much the same way that you use an ATM card to access an automatic teller machine.

Smart cards use cryptography-based authentication and provide stronger security than a password because in order to gain access, the user must be in physical possession of the card and must know the PIN.

#### III. Public key authentication: Diffie-Hellman Authentication

The key exchange is an important method in public-key Cryptography for providing authentication cryptographic service. It was the first public-key cryptographic scenario as

developed by Whitfield Diffie and Martin Hellman ,were the first who developed the key exchange algorithm that is called DH. In DH, keys are exchanged between the users according to Cryptography protocols which are based on the key exchange problem. They highlighted the most important method of exchanging the keys by using the discrete logarithm hard problem(refer to Figure8)

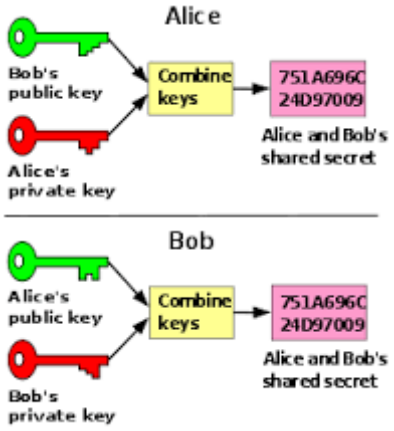



Figure 8: Diffie-Hellman Authentication

### 4. Biometrics Authentication

The word "biometrics" came from Greek and we can divide it into two roots: "bio" means life and "metrics" –to measure

Biometrics refers to metrics related to human characteristics. Biometrics authentication (or realistic authentication)is used in computer science as a form of identification and access control. It is also used to identify individuals in groups that are under surveillance.

Biometric identifiers are the distinctive, measurable characteristics used to label and describe individuals. Biometric identifiers are often categorized as physiological

versus behavioral characteristics.[Physiological characteristics are related to the shape of the body. Examples include, but are not limited to fingerprint, palm veins, face recognition, DNA, palm print, hand geometry, iris recognition, retina and odour /scent. Behavioral characteristics are related to the pattern of behavior of a person, including but not limited to typing rhythm, gait, and voice.

Biometric authentication systems compare the current biometric data capture to stored, confirmed authentic data in a database. If both samples of the biometric data match, authentication is confirmed and access is granted. The process is sometimes part of a multifactor authentication system. For example, a smart phone user might log on with his personal identification number (PIN) and then provide an iris scan to complete the authentication process.

Biometric characteristics of a person are unique. Most of such keys are impossible to copy and exactly produce. Theoretically these are ideal keys. But by using biometric identification a lot of specific problems appear. All biometric identifiers can be divided into two big groups:

**1)Physiological**

**2)Behavior**

Though behavior biometrics is less expensive and less dangerous for the user, physio-logical characteristics offer highly exact identification of a person. Nevertheless, all two

types provide high level of identification than passwords and cards. Spheres of use:

•Criminalistics (biometric identifiers are used to recognize victims, unidentified body and protection of children against kidnapping.)

•Marketing (methods of biometrics are used to identify owners of loyal cards)

•Time accounting systems at work, schools, etc

•Security systems ( are use to control the access to the rooms and control access to internet resources)

•Voting system ( during the functionality of voting system identification/authentication of people, that take part in voting is demanded)

•According to actual international demands ( for example, according to the standard of ICAO there should be biometric part in passport.)

•Biometric identifiers are used for registration if immigrants and foreign workers. It allows identifying people even without documents.

•For organization of distribution of social help

.Methods of biometric authentication differ according their degree of safeness:

•Fingerprint

•DNA

•Iris recognition

•Face recognition

•Voice

### I. Fingerprint recognition

Fingerprint recognition refers to the automated method of identifying or confirming the identity of an individual based on the comparison of two fingerprints. **Fingerprint recognition** is one of the most well known biometrics, and it is by far the most used biometric solution for authentication on computerized systems. The reasons for fingerprint recognition being so popular are the ease of acquisition, established use and acceptance when compared to other biometrics, and the fact that there are numerous (ten) sources of this biometric on each individual.

Figure 9: Fingerprint recognition

II.DNA Matching

*Chemical Biometric* The identification of an individual using the analysis of segments from DNA.

Figure 10: **DNA Matching**

### III. Eyes - Iris Recognition

Iris recognition is the best of breed authentication process available today. While many mistake it for retinal scanning, iris recognition simply involves taking a picture of the iris; this picture is used solely for authentication..

Figure 11: Eyes - Iris Recognition

### IV. Face Recognition

*Visual Biometric* The analysis of facial features or patterns for the authentication or recognition of an individuals identity. Most face recognition systems either use faces or local feature analysis.

**Figure 12:** Face Recognition

.

### V. Voice biometric authentication

Is the use of the voice pattern to recognize the identity of the person. Meanwhile,

voice authentication is now considered as a fast wide deployed form of biometric authentication. However, it is one of the best methods of determining the biometric method efficacy. However, voice recognition is categorized into five types: speaker dependent system, speaker independent system, discrete speech recognition, continuous speech recognition, and natural language.

## 5.CONCLUSION

This paper summarizes the difference between cryptography based authentication methods. It shows the possibility of using public key cryptosystem to provide high level of authentication security. The surveyed methods are password, public key sharing, symmetric key, and biometric authentication methods. The performance of these methods are actually different according to their tools cost, execution time and the security. While biometrics technology provides a strong user authentication solution, there are other variables to be considered in the authentication protocol. When a high level of security is needed, it is recommended that you combine other authentication factors with biometrics, but sometime they need extra tools which are required extra cost. Therefore, the surveyed public key authentication is good alternative to the traditional biometric methods.

## REFERENCES

1. Biometrics, Aidan Dysart

http://www.zdnet.com/pcmag/features/biometrics/bench.html

2. W. Stallings, “Cryptography and Network Security Principles and

Practices”, Pearson Education, 5rd edition, 2011.

3. Government Chief Information Officer. "e-Authentication".

http://www.e-authentication.gov.hk/en/professional/skey.htm, 2014.

4.http://searchsecurity.techtarget.com/tip/Biometric-authentication-know-how-Devices-systems-and-implementation

5Proceedings of the World Congress on Engineering and Computer Science 2014 Vol I

6. Iris Recognition: Issues

http://www.iris-scan.com/iris_cautionary.htm

# Segmenting and Hiding Data Randomly using Image Steganography

**Ankita Rameshchandra Mithaiwala**
M.Tech Student, Computer Engineering, CGPIT, Surat
ankitamithaiwala10@gmail.com

**Prof. Asha Hemaush Nagoriya**
Christ College,Pune
ahnagoriya@gmail.com

## ABSTRACT

**Security is major issue in today internet world. Information that passes on the internet is not safe or secure from the outsider or any third party. Data hiding provide the way to send the information in secure way. There are three way to provide the data hiding steganography, watermarking and fingerprint reader. Steganography is the one of the method to hide the data using innocent looking cover object. Image Steganography achieve using LSB, masking and filtering and transform techniques. Least significant bit (LSB) is the technique to hide the secret information into image but using simple LSB any one can get the information from image. In this paper provide review on SHDRIC technique using LSB that gives better result compared to simple LSB. In SHDRIC technique take the input of text data and segment the data into odd and even segment. After segmenting data one index channel is selected out of RGB channel and other two channels used to send the information. By switching index channel create more confusion to any attacker to get the message from image and it will also achieved two level of security. This technique implemented in MATLAB and result compared with existing simple LSB technique.**

**KEY WORDS: Data Hiding, Steganography, Index Channel, LSB, Data Segmentation**

## 1. INTRODUCTION

In today's world of digital information, one of the major issues is to decrease loss of information during the transmission. Data hiding plays a significant role to serve the said purpose. Information that sends in this virtual world is not secure for outsider to provide the security of the information data hiding used. Data hiding is the technique to send the secret information into some cover object. There is various ways to achieve the data hiding like Watermarking, steganography, cryptography. Watermarking is used to send the secret information as one watermark but this technique easily detectable because watermark place at some fixed position using used for digital signature. Steganography is the other technique for data hiding.Stganography is the technique for hiding the secret information in some cover object like image, audio, video and text.

Steganography" is a Greek origin word which means "hidden writing". Steganography word is classified into two parts: Steganos which means "secret or covered" (where you want to hide the secret messages) and the graphic which means "writing" (text). However, in the hiding information the meaning of steganography is hiding text or secret messages into another media file such as image, text, sound, and video.

Steganography and watermarking differs in a number of ways, which includes detection/extraction methods, aim, specification etc. The ultimate difference is the object of communication. Watermarking takes the host signal as object of communication. In which the copyright information are embedded. In steganography the object of communication is used to send the secret message.

A good technique of image steganography aims at three aspects. First one is capacity (the maximum data that can be stored inside cover image). Second one is the imperceptibility (the visual quality of stego-image after data hiding) and the robustness [5].

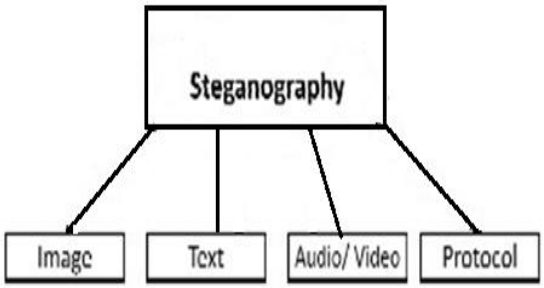



Figure 1.1 Types of Steganography

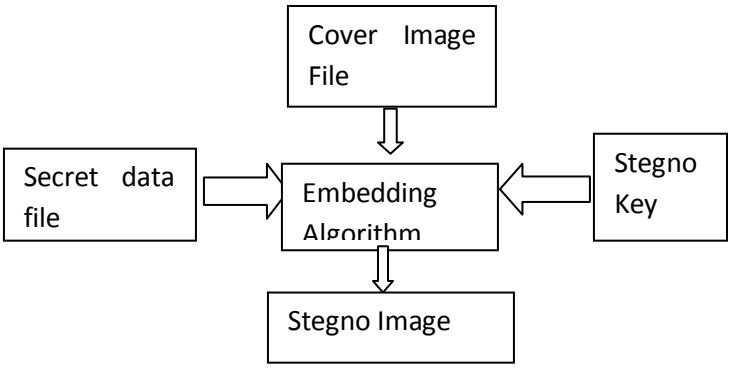



Figure 1.2 Basic approaches to produced stegno image

A general model for the steganography in the following steps:

Step- 1 Insertion of secret data files into cover file

Step- 2 Distribution of stego image

Step- 3 Data extraction/detection

Three basic types of stego systems are available:

• Pure stego systems - no key is used.

• Secret-key stego systems - secret key is used.

• Public-key stego systems - public key is used.

In this paper implement pure stego system.no key used in transmission.

The remainder of this paper is organized as follows. Section 2 Contains steganography techniques. In Section 3 Segmenting And Hiding Data Randomly Using Based On Index Channel. Section 4 Describe Performance Measure. Section 5 Describe the Result and Discussion. Then in Section 6 Conclusions and Future scope of SHDRIC.

## 2. STEGANOGRAPHY TECHNIQUES

Three are three ways to doing Image steganography:

1. Least Significant bit
2. Masking and filtering
3. Transform techniques

### 2.1 Least Significant Bit

LSB is the simple and widely used technique for image steganography. In computer or over internet image pass as the RGB format. An image is a collection of pixels which in binary format. In image least significant bit not used to pass any important information. In LSB technique use this least significant bit pass the secret information.

Image that we send on the internet is many formats like JPEG, BMP, and PNG etc. Using the simple LSB any one can get the information from least significant bit.

Advantages:

1. Easy to implement
2. High payload can pass
3. Imperceptibility is low

Techniques for LSB: There are various technique used to perform the LSB

1. Simple LSB
2. Segmenting and hiding data randomly using based on index channel.
3. Standard Deviation Converges
4. Edge Adaptive Image
5. Message Segmentation
6. Huffman Coding

### 2.2 Masking and Filtering

These techniques used to hide information by marking an image example watermarks image. These techniques embed the information in the more significant areas. Watermarking techniques can be applied. These techniques easily detectable because watermark image go in visible portion or anyone can guess the location of secrete image.

### 2.3 Transform Techniques

These techniques are used to send the information in transform or frequency domain. These techniques are complex and send less amount of data compare to LSB and Masking. There are mainly two techniques used for transform domain:

1. Discrete cosine transformation technique (DCT)
2 Discrete Wavelet transformation techniques (DWT)

## 3. SEGMENTING AND HIDING DATA RANDOMLY USING BASED ON INDEX CHANNEL [7]

In this image steganography which is the combination of the simple LSB and message segmentation. In this method require the input file which contain secret information that send and innocent looking

image which used to send the data inside the image which is called as cover image.

Image is a collection of pixels. Image that pass over internet is in RGB format. So each image contain three channel that is RED, GREEN, BLUE .In this method out of the three channel one channel used as the index channel for two other channels. When using 24 bit color image, two bits of two colors can be used to store the total 2 bits.

Suppose Message is:
"hello world"

Table 3.1: Original data and index position

| h | e | l | l | o | | w | o | r | L | d |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Table 3.2: Even Segment

| h | l | o | w | r | d |
|---|---|---|---|---|---|

Table 3.3: Odd Segment

| e | l | | o | l |
|---|---|---|---|---|

In RGB three channels used in that one channel is used as index channel. If red channel used as index channel other two channel green and blue used to send the data.

If the index channel contain even number of"1"s than even segment data store in least significant bit of channel 1 and least significant bit of channel 2. If the number of index channel has odd number of "1"s than odd segment data stored in two channel least significant bits.
If the RGB channel has following bits:

Table 3.4: Index channel and embedding operation

| | Index | Channel 1 | Channel 2 | |
|---|---|---|---|---|
| Pixel(1): | **100101**1 | 00011**01** | 11101**01** | Even |
| Pixel(2): | **111**0000 | 00011**10** | 11110**11** | Odd |

In that example if the number of "1"s is even in index channel than even segment 4 bit out of 2 bit store in 1 channel and 2 bit in 2 channel least significant bit. Same steps perform for the odd number of bits in odd segment.

Table 3.5 Meaning of index channel values [7]

| No of "1" in index channel | Channel1 | Channel 2 |
|---|---|---|
| Even | 2 bit of even segment | 2 bit of Even segment |
| Odd | 2 bit of odd segment | 2 bit of odd segment |

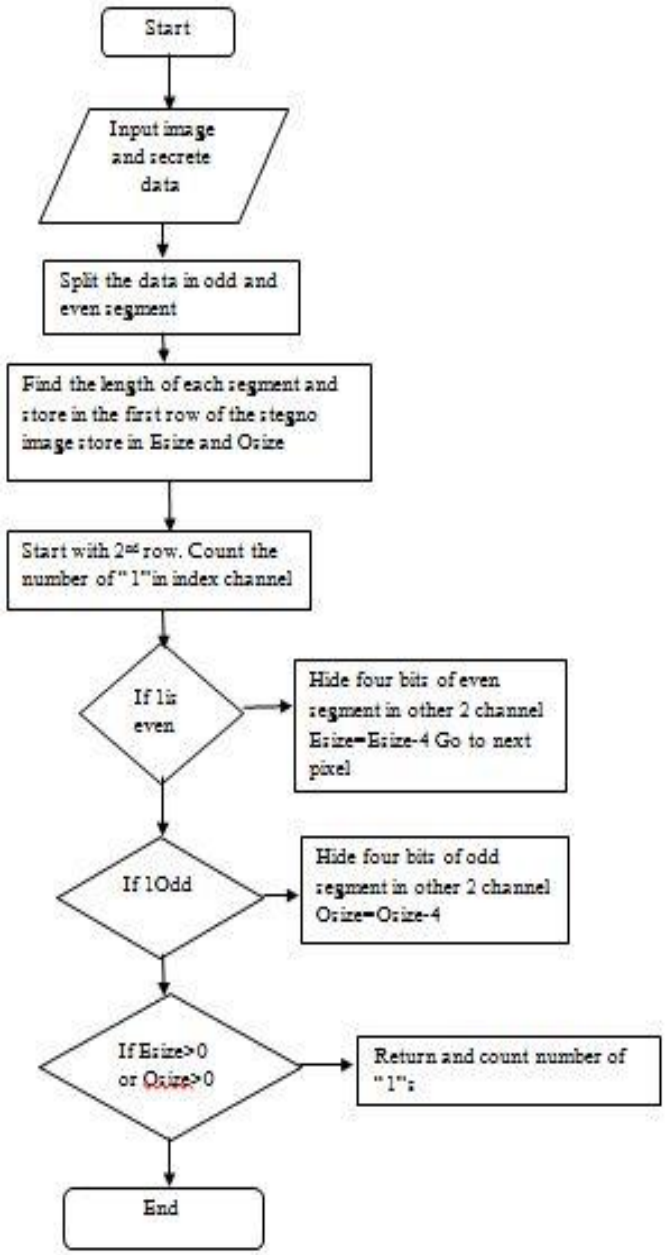



Figure 3.1 Encoding Steps

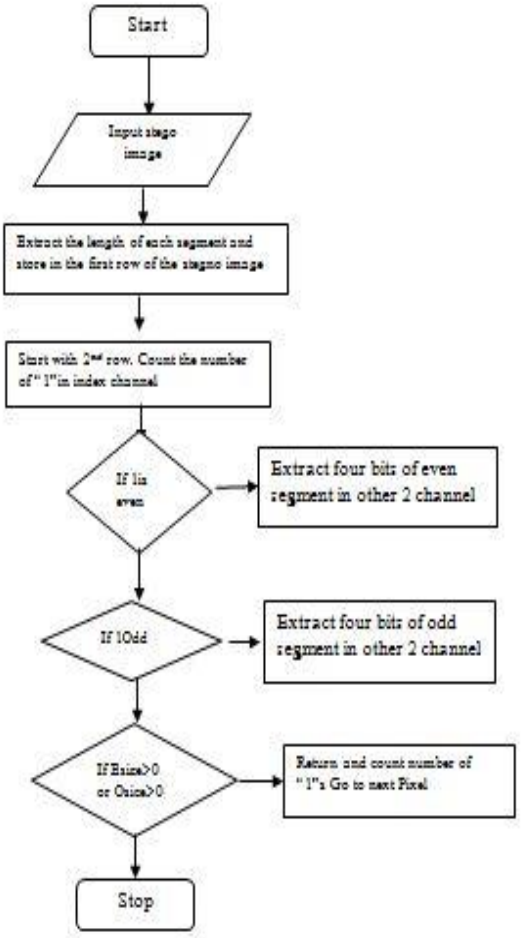



Figure 3.2 Decoding Steps

**4. PERFORMANCE MEASURE**

As a performance measure for image distortion due to embedding, the well-known peak-signal-to noise ratio (PSNR), which is categorized under difference distortion metrics, can be applied to stego images [1].

1. **MSE:**

Mean Squared Error (MSE) is computed by performing byte by byte comparisons of the CVR and stego-image. The computation can be expressed as follows:

$$\mathbf{MSE=1/MN\ {}^{M}\sum_{i=1}\ {}^{N}\sum_{j=1}(f_{ij}-g_{ij})^2 \quad [3]}$$

Where;

M, N is the number of rows and columns in the CVR matrix,

fij is the pixel value from CVR, and

gij is the pixel value from the stego-image.

Higher value of MSE indicates dissimilarity between compared images.

**2. PSNR:**

Peak signal-to-noise ratio measures in decibels the quality of the stego-image compared with the CVR.

The higher value of PSNR gives the better the quality of the image. PSNR is computed using the following equation:

$$\mathbf{PSNR=10log_{10}\ L^2/MSE \quad [4]}$$

Where; L is the peak signal level of cover image for gray scale image L=255.

**5. Result and Discussion**

Implementation of SHDRIC algorithm is carried out using **MATLAB 7.8.0 (R2009a)**. As we require two different files for data hiding, so the secret message is plaintext file (.txt). Embedded message is a plaintext. Along with that cover image is used for embedding secret message is **bird.jpeg**. Cover image is color image of size 512x512 pixels

(a) (b)

Figure 5.1 (a) Cover image (b) Stego image

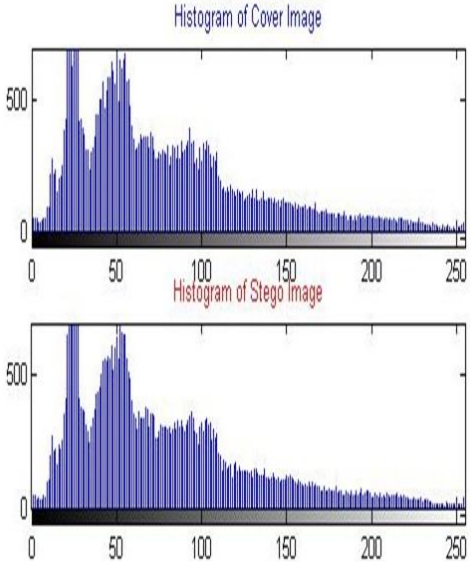



Figure 5.2 Histogram comparisons of cover image and stego image

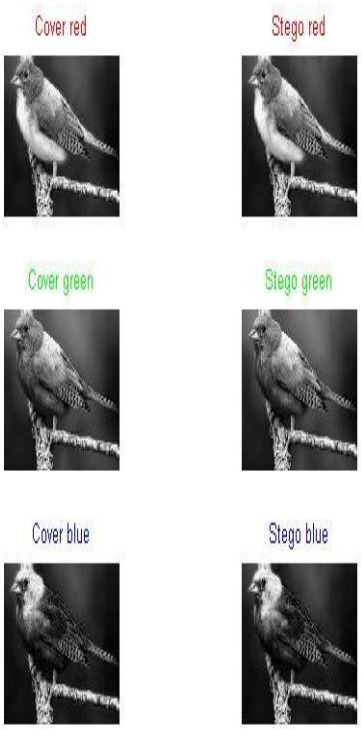



Figure 5.3 R, G & B Monochromes of cover image and stego image

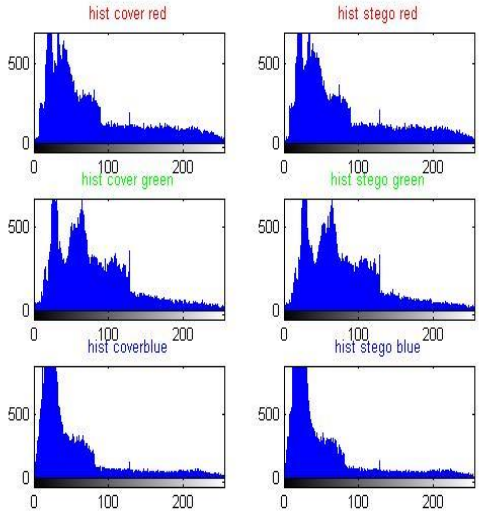



Figure 5.4 Histograms of all monochromes of cover image and stego image

In this comparison between SHDRIC and LSB is done based on Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). For five different cover images of different file formats and secret message files are used to reckon MSE and their results are shown in Table 5.1.

Table 5.1 Calculated MSE of Images

| Images | SHDRIC MSE | LSB MSE |
|---|---|---|
| Bird.jpeg | 6.31 | 42.12 |
| Love.jpeg | 7.06 | 83.58 |
| Lenna.tiff | 31.06 | 2.38 |
| Baboon.jpeg | 11.75 | 55.31 |
| Gilr.tiff | 51.33 | 78.134 |

Table 5.2 Calculated PSNR of images

| Images | SHDRIC PSNR | LSB PSNR |
|---|---|---|
| Bird.jpeg | 40.13 | 1.82 |
| Love.jpeg | 39.64 | 14.00 |
| Lenna.tiff | 33.21 | 81.41 |
| Baboon.jpeg | 37.43 | 7.91 |
| Girl.tiff | 37.03 | 0.80 |

Peak signal-to-noise ratio measures in decibels the quality of the stego-image compared with the CVR. The value PSNR get higher it shows the better the quality of the image.

From Table 5.1, Table 5.2 it is easy to analyze that MSE of approach (SHDRIC) is better than LSB approach. Likewise PSNR results of approach (SHDRIC) for all five cover images & secret message files are also better than LSB.

## 5. CONCLUSION AND FUTURE SCOPE

In today's digital world of digital information, information security has become a primary matter of concern for covert communication across several fields. These techniques get the better security than Simple LSB and achieve the two level of security. SHDRIC technique achieves more payload compare to other techniques. The result shows using this technique achieve the 30 to 50 % PSNR ratio.

In future SHDRIS extends to CDICS provides three layers of security. Due to implicit end marker the CDICS does require size of data to be embedded. So the crucial benefits include random embedding, implicit end marker, high data embedding capacity and higher visual quality and also add the cryptography and achieve one more level of security.

## 6. REFERENCES

[1] Alaa A. Jabbar Altaay, Shahrin bin Sahib and Mazdak Zamani, "An Introduction to Image Steganography Techniques", 2012 International Conference on Advanced Computer Science Applications and Technologies IEEE, 2013

[2] A. Shaddad, J. Condell, K. Curran, and P. Mckevtt. "Biometric inspired digital image steganography." In Proc. of the 15th Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems, 2008. Pp. 159-168.

[3] B. Jamil Mohd, S. Abed ,T. Al-Hayajneh, and S. Alouneh, "FPGA Hardware of the LSB Steganography Method", IEEE Computers, 2012.

[4] Babloo Saha, Shuchi Sharma (2012), "Steganographic Techniques of Data Hiding Using Digital Image", vol. 62, pp. 11-18

[5] C. Kessler. (2001). Steganography: Hiding Data within Data. An edited version of this paper with tbe title "Hiding Data in Data".Windows&.NETMagazine.http://www.garykessler. netllibrary/steganog aphy.btml

[6] Dr. Mohammed Abbas Fadhil Al-Husainy, "Message Segmentation to Enhance the Security of LSB Image Steganography", International Journal of Advanced Computer Science and Applications (IJACSA), 2012

[7] Emad T. Khalaf and Norrozila Sulaiman, "Segmenting and Hiding Data Randomly Based on Index Channel", IJCSI International Journal of Computer Science Issues, vol. 8, issue 3, No. 1, May 2011

[8] G. Karthigai Selvi, Leon Mariadhasan, K.L. Sunmugnanthan, "Steganography Using Edge Adaptive Image", International Conference on Computing, Electronics and Electrical Technologies [ICCEET] IEEE, 2012

[9] Rengarajan Amritharajan, P. Archana, V. Rajesh, G. Devipriya, J.B.B. Rayappan, "Standard Deviation Converges for Random Image Steganography", IEEE Conference on Information and Communication Technologies (ICT), 2013

[10] Rig Das and Themrichon Tuithung, "A Novel Steganography Method for Image Based on Huffman Encoding", IEEE, 2012

[11] Katzenbeisser, S. and Petitcolas, F.A.P., (2000) Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Inc., Boston, London.

[12] Z.Z Kermani and M. Jamzad. "A robust steganography algorithm based on texture similarity using Gabor filter." in Proc. of IEEE 5th International Symposium on Signal Processing and Information Technology ISSPIT, 2005. pp. 578-582.

[13] MAB. Younes, A. Jantan, "Image encryption Using BlockBased Transformation Algorithm,"AENG International Journal of Computer Science, Vol. 35, Issue. 1, 2008, pp.1523.

[14] Saurabh Singh, Gaurav Agarwal Use of image to secure text message with the help of LSB replacement INTERNATIONAL JOURNAL OF APPLIED ENGINEERING RESEARCH, DINDIGUL Volume 1, No1, 2010

[15] Petitcolas, F.A.P., Anderson, R. J. and Kuhn, M.G. (1999) "Information Hiding -A Survey", Proceedings of the IEEE, Special issue on Protection of Multimedia Content, vol. 87, no. 7, pp.1062- 1078.

[16] Johnson, N.F. and Jajodia, S. (1998) "Exploring Steganography: Seeing the Unseen", *IEEE, Computer*, vol. 31, no. 2, pp. 26-34.

[17] Kahate, A. (2008), "*Cryptography and Network Security*", 2nd Edition, Tata McGraw-Hill.

[18] Gonzalez, R. C. and Woods, R. E. (2002), "*Digital Image Processing*", 2nd edition, Prentice Hall, Inc.

# The study of Shimin Wei Digital Signature Schemes based on Integer Factorization and discrete logarithm

**A.B.Nimbalkar, Gadekar M.J.**
Annasaheb Magar College, Hadapsar, Pune28 India.
nimbalkar_ab@yahoo.com
tomanisha15@gmail.com

***Abstract***
**A digital signature is an electronic signature that can be used to authenticate the identity of the sender of a message or the signer of a document, and possibly to ensure that the original content of the message or document that has been sent is unchanged. Now a days document security over internet is must. One can use various ways to form the digital signature, like Factoring and Discrete Logarithm. Shimin Wei (2007) scheme is based on both hard problems FAC and DL. But there are some draw backs for implementation of this scheme. Also some attacks can be possible on this scheme.**

**Key words**: Cryptography, discrete logarithm(DL), digital signature, Factoring(FAC)

## I. INTRODUCTION

Cryptography is an ancient art, historically used for securely communicating. Today, Cryptography has many uses beyond military ones. The millions of users daily to purchase goods over the Internet. Cryptography even has applications beyond secure communication, such as electronic cash..

There are two main categories of cryptographic communication protocols: symmetric (i.e. private key) and asymmetric (i.e. public key)[3]

Digital signature : A digital signature scheme allows one to sign an electronic message and later the produced signature that can be validated by the owner of the message or by any verifier. Most of the existing digital signature schemes were developed based on a single hard problem like factoring, discrete logarithm, residuosity or elliptic curve discrete logarithm problems.

Cryptography: The art of protecting information by transforming it (*encrypting* it) into an unreadable format, called cipher text. Only those who possess a secret *key* can decipher (or *decrypt*) the message into plain text. Encrypted messages can sometimes be broken by cryptanalysis, also called *codebreaking*, although modern cryptography techniques are virtually unbreakable.

Cryptography systems can be broadly classified into symmetric-key systems that use a single key that both the

sender and recipient have, and *public-key* systems that use two keys, a public key known to everyone and a private key that only the recipient of messages uses.

## II Integer factorization:[13]

The security of many cryptographic techniques depends upon the intractability of the integer factorization problem. A partial list of such schemes includes the RSA public-key encryption scheme and the RSA signature scheme.

This problem is believed to be hard for general n when n is large. Some ingenious methods have been devised in an attempt to factorize large composite numbers n.

The three methods that are most effective on very large numbers are the quadratic sieve, the elliptic curve method and the number field sieve.

## III Discrete Logarithmic method:

The discrete logarithm problem applies to mathematical structures called groups. The discrete logarithmic problem is as follows: given an element g in finite group G and another element h in G, find inter x such that $g^x=h$.

The discrete logarithm problem has received much attention in recent years. The best discrete logarithm algorithms have expected running times similar to those of the best factoring algorithms.

The security of many cryptographic techniques depends on the intractability of the discrete logarithm problem. A partial list of these includes Diffie-Hellman key agreement [1] and its derivatives, ElGamal [2] encryption.

**Definition of discrete logarithm:** Let G be a finite cyclic group of order n. Let $\alpha$ be a generator of G, and let $\beta \in G$. The *discrete logarithm of* $\beta$ *to the base* $\alpha$, denoted $\log_\alpha \beta$, is the unique integer x, $0 \le x \le n-1$, such that $\beta = \alpha^x$.

**Definition** The *discrete logarithm problem* (DLP) is the following: given a prime p, a generator α of $Z^*_P$ , and an element β ∈ $Z^*_P$ , find the integer x, $0 \le x \le p-2$, such that $\alpha^x \equiv \beta \pmod p$.

- Baby-step giant-step algorithm.
- Pollard's rho algorithm for logarithms.
- Pohlig-Hellman algorithm.
- Index-calculus algorithm.
-

**IV. Digital Signature Schemes:**

A signature consists of three probabilistic polynomial algorithms:

**Key-Generator**: it takes as input a security parameter k, and outputs a pair (sk,pk), where sk is secrete key of the user, and pk is the matching public key.

**Signature:** this algorithm takes as input a message m and the secrete key sk and produces a signature Θ.

**Verification**: Finally, the verification algorithm takes as input a message m, a signature Θ and the public key pk, and returns "true" if Θ is a valid signature of m, and "false" otherwise.

**Shimin wei's Digital Signature Scheme 2007 [16]**

1) Algorithm :

Let p be big prime $p=4p_1q_1+1$ ,

Where $p_1 = 2p_2 +1$ , $q_1 = 2q_2 +1$ and $p_1, q_1$, $p_2 , q_2$ are all large primes. These parameters must be kept secret from every user.

Let g be an element with order $p_1q_1$ of the finite field GF(p).

Secrete key x is : $1<x< p_1q_1/2$

Public key y is : $y=g^{x^2-x^{-2}} \pmod p$( $n= p_1q_1$)

To sign the message m user does the following:

1) Randomly choose an integer t as $1 < t <n/2$
   Randomly choose an odd integer k as $1 < k <n/2$
   And calculates u and v
   $u=g^{t^2-t^{-2}} \pmod p$ and $v=u^{k^2} \pmod p$
2) Compute s and r such that
   $xs+x^{-1}r=um^2t+vmkt^{-1} \mod n$ -------- (1)

   $x^{-1}s+xr=um^2 t^{-1}+vmkt \mod n$ -------- (2)

To find r multiply eq1 by $x^{-1}$ and eq2 by x and solve simultaneous eq. we get

$s+x^{-2}r = x^{-1} um^2t + vmkt^{-1}x^{-1} \mod n$ ----- (3)

$s+x^2r = xum^2 t^{-1} + vmktx \mod n$ ----- (4)

$(x^{-2}- x^2)r = um^2(x^{-1}t – x t^{-1}) + vmk(t^{-1} x^{-1} – tx)$

$$\mathbf{r = (x^{-2}- x^2)^{-1} .[ um^2(x^{-1}t – x t^{-1}) + vmk(t^{-1} x^{-1} – tx)]}$$ ---- (5)

To find s multiply eq1 by x and eq2 by $x^{-1}$ and solve simultaneous eq.
we get

$x^2s+ r = xum^2t + xvmkt^{-1} \mod n$ --- --- (6)

$x^{-2}s+ r = x^{-1}um^2 t^{-1}+ x^{-1}vmkt \mod n$ -------(7)

$s(x^2- x^{-2}) = um^2(xt - x^{-1} t^{-1}) + vmk (xt^{-1}-x^{-1}t)$

$$\mathbf{s = (x^2- x^{-2})^{-1} .[ um^2(xt - x^{-1} t^{-1}) + vmk(xt^{-1}-x^{-1}t)]}$$

---- (8)

3) Send signature(u,v,r,s) as signature of m.
   The verification equation of sign (u,v,r,s) is:

$u^{u^2m^4} = v^{v^2m^2}\ y^{s^2-r^2} \pmod p$.

**V. output: from 100 number only two combinations gives value of p, p1 ,p2,q1,q2 primes**

**If all numbers r prime then message is displayed**

**Number with p indicates Prime and N indicate Not prime number.**

During evaluation of In[376]:= 281p 199p 563p 399N 898549p

During evaluation of In[376]:= all r prime s

During evaluation of In[376]:= 5p 251p 11p 503p 22133p

During evaluation of In[376]:= all r prime s

During evaluation of In[376]:= 359p 23p 719p 47p 135173p

During evaluation of In[376]:= 89p 307p 179p 615N 440341N

It is very time consuming to generate prime p with all p1,p2,q1,q2 are prime.

In this Implementation I have checked p prime, if not changed the values of p2,q2

**3) Implementation :** Shimin Wei's 2007 new Scheme is implemented in Mathematica9.0. the result is as follows :

Time start :39.75
Number of Iteration required to chk Prime num5
P is :300627332549
P1 is :272759
q1 is :275543
P2 is :136379
q2 is :137771
n is :75156833137
g is :16
half37578416568
Inverse of x2 - x-2 is :35073188797
Inverse of x-2 - x2 is :40083644340
Number of Iteration required to chk X is Invertible num0
Number of Iteration required to chk X Square is Invertible num0
Number of Iteration required to chk X Subtraction is Invertible num0
x 37578416568
x square 56367624853
x Sq Inver 4
Time start before Y :39.75
Time after Y :39.75
Y is : 173878468153
t 32626738225
t Square 53946693196
t sq Inv 30171388021
k odd : 30326043119
Time start before u :39.75
t2 munus t Sq Inv =23775305175
Time after U :39.75
U is :109635503242
k sq 2216886357
Time after v :39.75
V is :248135796731
x Inv 75156833135
t Inv 47575767273
r is 34093181405
s is 36101290727
uSq is :5988957479
umForth is :20736
Lhs = 16326062450
mul v271111472980
v rhs30747005722
y rhs 126054144244
rhs = 16326062450

The Following attack is possible :

Public key y is : $y=g^{x^2-x^{-2}} \pmod p$

$u=g^{t^2-t^{-2}} \pmod p$ and $v=u^{k^2} \pmod p$

The verification equation of sign (u,v,r,s) is:

$$u^{u^2m^4} = v^{v^2m^2}\, y^{s^2-r^2} \pmod p. \quad \text{----------(1)}$$

For message m attacker substitute $u=y^2$ and $v=y^3$ on verification identity Eq(1).

## VI. CONCLUSION

The electronic transactions are tremendously increasing over an internet, which have led to a critical need for ensuing the authenticity and also nonrepudiation. The Digital Signature serve the purpose of authentication and validation e-documents. We encounter the problems in shiminWei2007 schemes. The problems

The Major problem with Shimin Wei signature

i) The attack is shown by Lin, Gun and Chen, so it is no more secure as claimed by Shimin Wei.

ii) There is no sufficient information for generating the prime p, because as per scheme p should be prime. But Shimin Wei scheme not generating P as prime.

## REFERENCES

1. G.J. Simmons. "Contemporary Cryptography: The Science of Information Integrity". IEEE Press, 1992.

2.John W. Rittinghouse, William M. Hancock, " Cyber Security Operation Hand Book", Elsevier Digital Press 2003

3.Whitfield Diffie and Martin E. Hellman. New directions in cryptography. IEEE Transactions on Information Theory, 22(6):644-654, 1976.

4. Rivest, R., A. Shamir and L. Adleman, 1978. A method for obtaining digital signature and publickey cryptosystem. Commun. ACM., 21: 120-126. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.5588.

5. ElGamal, T., 1985. A public key cryptosystem and a signature scheme based on discrete logarithm. IEEE Trans. Inform. Theory IT, 25: 469-472. http://dsns.csie.nctu.edu.tw/research/crypto/HTML/PDF/C84/10.PDF.

6. V. Miller, "Uses of elliptic curves in cryptography", Proc of CRYPTO'85, 1986, pp.417–426.

7.N. Koblitz, "Elliptic Curve Cryptosystems", Mathematics of Computation, 1987, pp.203–209.

8. RABIN, M.O.: 'Digitalized signatures and public key functions as intractable as factorization'. Tech. Rep., MITILCSITR-212, MIT Laboratory for Computer Science, MIT, Cambridge, MA, 1979

9. Nyang, D , and Song, J.: 'Fast digital signature scheme based on the quadratic residue problem', Electron. Lett., 1997, 27, (3), pp. 205- 206

10 Wei-Hua He and Tzong-Chen Wu, "Improvement to Nyang-Song fast digital signature scheme" © IEE 1997 Electronics Letters Online No: 19971243

11.Hard jono T. and Zheng Y. (1992). A practical digital multi-signature scheme based on discrete logarithm, Advance in Cryptology – Auscrypto – 92 Springer and Verlag, p.p.16-21.

12. Harn L. and Kiesler T. (1989). New Scheme for Digital Multi-signature, Electronic Letters – 25 (15), p.p. 1002-1003.

13. D.R. Stinson. "Cryptography theory and practice". CRC Press, 1995

14. Feige,Fiaza., and Shamir ,A.: 'Zero knowledge proofs of identity', J. Cryptol., 1988, 1, (2), pp. 77-94

15. Hwang T. and Chen C.C. (2001). A new proxy multi-signature signature scheme, International Workshop on Cryptography and Network Security, Taipei, p.p. 22 – 28.

16. Wei S., "A New Digital signature schemes based on factoring and discrete logarithms", *Progress on Cryptography The Springer International Series in Engineering and Computer Science*, Vol 769, 107-111, (2004).

17. Desmedt, Y. and Frankel Y. (1990). Threshold cryptosystems, Advances in Cryptology –Crypto- 89, Springer Verlag, LNCS # 293, p.p. 307-315.

18. Boyar J., Chaum D., Damgard I. and Pederson T. (1990). Convertible undeniable signatures, Advances in Cryptology – Crypto - 90, Springer Verlag, LNCS # 537, p.p.189-205.

19. Chaum D. (1990). Zero knowledge undeniable signatures, Advance in Cryptology-Eurocrypt -90, Springer Verlag, LNCS # 473, p.p. 458-464.

# Transposition Technique with Modified Caesar Cipher

**Dr. Kavita Khobragade[1], Shubhankar G Keskar[2]**
[1] Associate Professor, Computer Science Department, Fergusson College, Pune - 4 kavitanand@rediffmail.com
[2] T.Y. Student, Fergusson College, Pune
shubhankarkeskar@yahoo.in

***Abstract***

**Conversion of original message into its non-readable form is called Cryptography. It is an art and science of converting original message into non-readable form. Transposition and Substitution techniques are used for converting data into non-readable form. Caesar cipher is an example of substitution method. As Modified Caesar cipher has various limitations so this paper will present a perspective on combination of substitution and transposition techniques together. A transposition technique is applied on Modified Caesar cipher in order to overcome all limitations of Modified Caesar cipher and provide much more secure and strong cipher and also difficult to decrypt if keys are not known. In this paper we tried to achieve better security while data transmission by using our technique.**

**Keywords – Cryptography, Encryption, Decryption, Key, Transposition, Modified Caesar Cipher.**

## I. INTRODUCTION

Data security is one of the important aspects in computing. There should be an implementation of a method to ensure secure data transfer. Today internet is a large web of interconnected computers. We can share various kinds of information on the web which can also contain sensitive data. This data remains unprotected. So there should be proper implementations of methods involving data security. So to do this the term cryptography comes in picture [1].

*Plain text - > Encryption - > Cipher Text - > Decryption - > Plain Text*

Cryptography contains terms like encryption, decryption, plain text & cipher text. The method of encoding plain text in such a way that the original contents are hidden is called encryption. The process of converting cipher text to its original plain text is called decryption [2, 5, 6, 7]. Data that can be read and understood easily is called plain text i.e. original data. Cipher text is the encrypted text of original data. Key is a set of numbers that the cipher operates on [1]. It is also an input to the encryption algorithm.

## II. MODIFIED CAESAR CIPHER

Modified Caesar cipher is one of the types of substitution method. In this method each letter of plain text is replaced by a letter of some fixed number of positions down the alphabet or each letter in set of alphabets is assigned other random alphabet which should not be repeated for the other alphabets [3, 4]. There are some weak points of modified Caesar cipher which are given below.

**i.** The algorithm for encryption and decryption is already known.

**ii.** Only 25 keys can be considered for the first part of this method.

**iii.** The language of plain text is known and can be recognized easily.

## III. TRANSPOSITION TECHNIQUE

Transposition technique is one in which the letters of a message are rearranged or permuted which includes methods like Rail Fence Method, Columnar Method etc. Transposition ciphers differ from substitution ciphers. Transposition ciphers do not simply replace one alphabet with another. They also perform some permutation over the plain text alphabet [2]. In this method we write the key & give numbers to the alphabets. We write the plain text horizontally in rows padded to fill the matrix. Then cipher text can be represented as columns, starting with the column whose key letter is lowest. Weak points of Transposition Technique are as follows:

**i.** Here letters are not repeated in the key.

**ii.** Key can be known if it is present in dictionary or a predefined word.

**iii.** Algorithm is known.

**iv.** Numbering of alphabets in key depends on order in which the alphabets are in its set.

## IV. LITERATURE REVIEW AND OBJECTIVE

Imran et. Al, used the Caesar cipher method. They used three methods in which each letter in the plaintext is replaced by some fixed number of position down the alphabet. The result is a data which is encrypted and be

decrypted to its readable form [1]. P.Subhasri et.al used reverse Caesar cipher algorithm for cloud computing [8].

The objectives of this research are to generate the best cipher text so that the security is preserved. To achieve this we have initiated this study with following objectives:

**i.** Implement two methods Modified Caesar Cipher & Transposition Technique combined together.

**ii.** To compare the proposed encryption method with both methods mentioned above.

**iii.** To create encryption & decryption method for the proposed research.

## V. PROPOSED ALGORITHM

This is an algorithm which is used to encrypt & decrypt data which provides more data security than Modified Caesar Cipher & Transposition Technique. In our work, we proposed the combination of both methods. The proposed algorithm is as follows:

### A. Encryption

Step 1 – Accept a message from the user which is nothing but the plain text which we have to encrypt.

Step 2 – Accept key 1 for first part of algorithm i.e. for transposition technique having max length 10 as we using 0-9 numbers for numbering the key. We can take any combination of 26 alphabets with repetitions as key. For example, if length is 5, it means we have 5 places. We have 26 alphabets all together. So there will be

$^{26}C_5 + {}^{26}C_4 + {}^{26}C_3 + {}^{26}C_2 + {}^{26}C_1 = 83681$ combinations for 5 places.

Step 3 – Start numbering the key letters. Here we use 0-9 numbers for numbering. Numbering of key letters can be any combination of 0-9 numbers but without repetitions. And it does not depend on the order of letters in the key in which they appear in the set of alphabets. For example, consider above example and we have 10 numbers so there will be $^{10}C_5 + {}^{9}C_4 + {}^{8}C_3 + {}^{7}C_2 + {}^{6}C_1 = 461$ combinations without repetitions for 5 places.

Step 4 – Then follow the process of transposition technique (columnar method).

Step 5 – Result of this will be our first cipher text which will be given further as input to Modified Caesar Cipher.

Step 6 – Now in first part of Modified Caesar Cipher initialize the alphabets with other alphabets in the set such that the initialized alphabets should not be repeated which will be key 2.

Example, A=Z, B=X … and so on till Z. We can assign the alphabets in any order every time.

Step 7 - Apply these initializations to the first cipher text. Result of this will give us cipher text 2.

Step 8 – Accept i from user where i=0, 1… 26. Now key 3 will be character + i where each letter in plain text (cipher text 2) is replaced by character + i letter.

Step 9 – Result of this will be the cipher text 3 which will be the final strongly encrypted cipher text.

Thus the original data is encrypted 3 times using these combined methods. We get a very strong encrypted message which cannot be decrypted without knowing the key and its order.

### B. Decryption

In decryption we are accepting a message which consists of all keys used in encryption in a sequential order. So for example, the new message will contain "key1key2key3". As we can see, the new message has unique pattern where key 1 is succeeded by key 2 and then key 3. So we can differentiate between these keys in that message.

Step 1 – Separate every key in that message i.e. tokenize the message.

Step 2 – Start applying the algorithms (Modified Caesar Cipher & Transposition Technique) backward by considering the keys.

Step 3 – Result will be the original message which is now delivered to the receiver.

### C. EXAMPLES USING PROPOSED METHOD

**Example 1** –

**Encryption** : Consider a message "Viking" that we have to encrypt and send it to the receiver.

Take length of key as 5 i.e. n=5.

Plain Text – Viking

Key 1 – a p q r z which is random combination of 26 alphabets with repetitions allowed. Start numbering the key letters which is also a random combination of 0-9 numbers with no repetitions allowed.

Columnar method -

Table 1.1

| a | p | q | r | z |
|---|---|---|---|---|
| 2 | 5 | 6 | 8 | 1 |
| v | i | k | i | n |
| g | | | | |

Now start writing the letters by considering the numbers given to the key letters sequentially.

Thus the Cipher Text 1 – nvgiki

Now we will apply Modified Caesar Cipher to the encrypted message.

Apply Cipher Text 1 to the first part of this method where all 26 alphabets are assigned to other alphabets.

Table 1.2

| A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z | X | Y | W | V | U | S | T | R | O | P | Q | K |

| N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| J | L | I | H | G | F | E | D | C | A | B | M | N |

Thus by applying above assignments to Cipher Text 1 we get,
Cipher Text 2 – jcsrpr and Key 2 - ZXYWVUSTROPQKJLIHGFEDCABMN
Now in part 2 we apply Key 3 – Character + i where i=0, 1… 26.
Take i=3.
For j, j + 3 = m.
For c, c + 3 = f.
For s, s + 3 = v.
For r, r + 3 = u.
For p, p + 3 = s.
Thus by applying this we get,
Cipher Text 3 – mfvusu
With the proposed method we got the encrypted message which is quite difficult to decrypt. Another message is sent to the receiver having all keys or the sender would send this message in header part of the actual message to be sent. So the message which will be sent with the key is given as follows:

"apqrz25681ZXYWVUSTROPQKJLIHGFEDCABMN3mfvusu".

**Decryption** : Once receiver got the encrypted message, the message should be decrypted. For decryption, tokenize the message and separate out the keys and apply the methods in reverse order.

Take last key i.e. 3
Apply Character - 3 to Cipher Text 3
Result will be Cipher Text 2 – jcsrpr
Then take the 2nd key and represent again in a tabular format as –

Table 1.3

| A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z | X | Y | W | V | U | S | T | R | O | P | Q | K | J |

| O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | I | H | G | F | E | D | C | A | B | M | N |

Now substitute back the values to the Cipher Text 2 by considering this table.
The result will be Cipher Text 1 – nvgiki.
Then take the 1st key and Cipher Text 1 to decrypt back to get original message using Transposition Technique. Here plot the letters of Cipher Text 1 sequentially in the table according to the numbering i.e. n is 1st letter so n will go in the column having number 1, v is 2nd so v will go in the column having number 2 and so on.

Table 1.4

| a | p | q | r | z |
|---|---|---|---|---|
| 2 | 5 | 6 | 8 | 1 |
| v | i | k | i | n |
| g | | | | |

Write it again without considering the numbering i.e. start writing it row-column wise just like we write the elements in any matrix sequentially. Therefore, the result of this is the original message i.e. "Viking".
**Example 2** : Just to verify our proposed method, we have encrypted and decrypted various messages. Let us see one more example.
**Encryption** : Consider a message "Trojan Horse" that we have to encrypt and send it to the receiver. Use Transposition Technique.
Take length of key as 4 i.e. n=4.
Plain Text – Trojan Horse
Here we will consider a case where all key letters are same.
Key 1 – a a a a which is random combination of 26 alphabets with repetitions allowed
Here the numbering does not matter as well depending on the key letters.
Start numbering the key letters which is also a random combination of 0-9 numbers with no repetitions allowed.
Columnar method - Table 2.1

| A | a | a | a |
|---|---|---|---|
| 3 | 1 | 4 | 2 |
| T | r | o | j |
| A | n | h | o |
| R | s | e | |

Now start writing the letters by considering the numbers given to the key letters sequentially. Thus the Cipher Text 1 is rnsjotarohe.
Apply Modified Caesar Cipher. Apply Cipher Text 1 to the first part of this method where all 26 alphabets are assigned to other alphabets.

Table 2.2

| A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z | X | Y | W | V | U | S | T | R | O | P | Q | K | J |

| O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | I | H | G | F | E | D | C | A | B | M | N |

Thus by applying above assignments to Cipher Text 1 we get,
Cipher Text 2 – gjfolezgltv and Key 2 - ZXYWVUSTROPQKJLIHGFEDCABMN

Now in part 2 we apply Key 3 – Character + i where i=0, 1… 26.
Take i=5.
Thus by applying this we get,
Cipher Text 3 – loktqjelqya
Decryption : The message sent to the receiver consists of encrypted text i.e. Cipher Text 3 and the key. So the message which contains keys is
"aaaa3142ZXYWVUSTROPQKJLIHGFEDCABMN5loktqjelqya".
Receiver first tokenize the message and separate out the keys and apply the methods in reverse order.

Take last key i.e. 5
Now apply Character - 5 to Cipher Text 3
Result will be Cipher Text 2 – gjfolezgltv
Then take the 2nd key and represent again in a tabular format as –

Table 2.3

| A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z | X | Y | W | V | U | S | T | R | O | P | Q | K | J |

| O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L | I | H | G | F | E | D | C | A | B | M | N |

Then substitute back the values to the Cipher Text 2 by considering this table.
The result will be Cipher Text 1 – rnsjotarohe.
Now take the 1st key and Cipher Text 1 to decrypt back to get original message using Transposition Technique. Here plot the letters of Cipher Text 1 sequentially in the table according to the numbering i.e. n is 1st letter so n will go in the column having number 1, v is 2nd so v will go in the column having number 2 and so on.

Table 2.4

| a | a | a | a |
|---|---|---|---|
| 3 | 1 | 4 | 2 |
| t | r | o | j |
| a | n | h | o |
| r | s | e | |

Write the message again without considering the numbering i.e. start writing it row-column wise just like we write the elements in any matrix sequentially. Therefore, the result of this is the original message i.e. "Trojan Horse".

## VI. CONCLUSION

Caesar Cipher method is simple which can be decrypted easily. That is why in this paper we have applied combination of Modified Caesar Cipher method and Transposition Technique. Both these methods can be easily performed on computers and by using any high level programming language. This combination provides more secure & strong cipher. The final cipher is strongly encrypted which is very difficult to break without knowing keys and the pattern of it. In substitution the letters are replaced by other letters and transposition method changes the position of the letters. The method described above uses combination of Modified Caesar Cipher & Transposition Technique to produce a strong & secure cipher.

## REFERENCES

[1]_ Enas Ismael Imran & Farah abdul ameeraabdulkareem, "Enhancement Caesar Cipher for Better Security", *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 01-05.

[2] Malay B. Pramanik, "Implementation of Cryptography Technique using Columnar Transposition", *International Journal of Computer Applications (0975 – 8887) Second National Conference on Recent Trends in Information Security*, GHRCE, Nagpur, India, Jan-2014.

[3]_ Mr. Vinod Saroha, Suman Mor & Anurag Dagar, "Enhancing Security of Caesar Cipher by Double Columnar Transposition Method", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 10, October 2012.

[4] Atul Kahate , "*Cryptography and Network Security"*, 2nd edition, McGraw-Hill, 2009.

[5] William Stalling, "Network Security Essentials (Applications and Standards)", Pearson Education,2004.

[6] Stallings W, "*Cryptography and Network Security"*, 2nd edition, Prentice Hall, 1999.

[7] William Stallings, *Cryptography and Network Security*, 3rd edition, Pearson Education, 2003.

[8] P.Subhasri, Dr. A. Padmapriya, " Implementation of Reverse Caesar Cipher Algorithm for Cloud Computing", *International Journal for Advance Research in Engineering and Technology*, Vol. 1, Issue VI,pp: 43-46, July 2013 [ISSN 2320-6802].

# Secure Data Deduplication in Cloud Computing

**MINI JOSWIN**
DEPT OF COMP.ENGG
G.H.R.I.E.T
PUNE,MAHARASHTRA
minijoswin2000@yahoo.com

***ABSTRACT***

**Deduplication is a popular technique used to save storage spaces and bandwidth in the cloud by removing identical copies of repeating data. This article explains the basic principles and implementation issues of data deduplication and covers some methodolies at work today. To retain the confidentiality of sensitive data while supporting the deduplication, the data is encrypted before outsourcing using convergent encryption technique. Although convergent encryption has been extensively adopted for secure deduplication, a critical issue of making convergent encryption practical is to efficiently and reliably manage a huge number of convergent keys. Further on we analyse a differential authorization duplicate check , in which each user is issued a set of privileges during system initialization and deduplication is performed according to the assigned privileges.**

**Keywords— Deduplication;Convergent key; differential privileges**

## I. INTRODUCTION

Cloud has become inevitable in every enterprise chiefly for storing and ubiquitous retrieval of data ranging from organizational data to personal data. Delivering robust yet high performance storage has been one of the greatest hardware and software challenges in the explosion of cloud computing. The most cited complaint from the users of many leading Infrastructure-as-a-Service (I-a-a- S) clouds is the poor storage performance. Organizations have to pay a huge sum of money to the Cloud Service Provider, for each and every GB of memory utilized in cloud. An important obstacle to cost efficient cloud storage is the presence of multiple copies or duplicates, which emaciates the storage space to a great extent [6]. Data deduplication is an advanced technology that can dramatically reduce the amount of backup data stored by eliminating redundant data. Data deduplication maximizes storage utilization while allowing IT to retain more nearline backup data for a longer time. This tremendously improves the efficiency of disk-based backup, changing the way data is protected [8].In general, data deduplication compares new data with existing data from previous backup or archiving jobs, and eliminates the redundancies. Advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less- expensive and faster offsite replication of backup data. The management of the ever-increasing volume of data is one of the critical challenges of today's cloud storage services.

According to the analysis report of IDC, the volume of data in the wild is expected to reach 40 trillion gigabytes in 2020 [11]. Data deduplication works by comparing blocks of data or objects (files) in order to detect duplicates. Deduplication can take place at two levels file and sub-file(block) level. In some systems, only complete files are compared, which is called Single Instance Storage (SIS). This is not as efficient as sub-file(block) deduplication, as entire files have to be stored again as a result of any minor modification to that file. From a users perspective, data outsourcing raises security and privacy concerns. We must trust third-party cloud providers to properly enforce confidentiality, integrity checking, and access control mechanisms against any insider and outsider attacks. However, deduplication, while improving storage and bandwidth efficiency, is incompatible with traditional encryption. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to different cipher texts , making deduplication impossible.

Convergent encryption provides a feasible option to enforce data confidentiality while realizing deduplication .It encrypts/decrypts a data copy with a convergent key, which is derived by computing the cryptographic hash value of the content of the data copy itself . After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since encryption is deterministic, identical data copies will generate the same convergent key and the same cipher text. This allows the cloud to perform deduplication on the cipher texts. The cipher texts can only be decrypted by the corresponding data owners with their convergent keys.

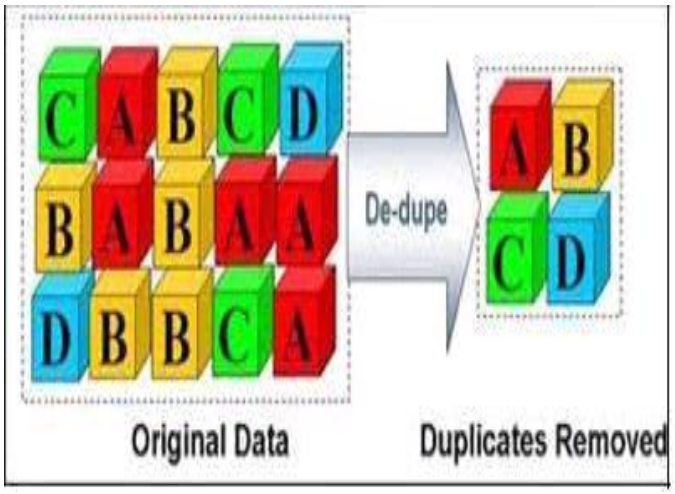



Fig 1: Removing redundant information through Deduplication

## II. RELATED WORK

Data security is another area of increasing importance in modern storage systems and, unfortunately, deduplication and encryption are, to a great extent, diametrically opposed to one another. Deduplication takes advantage of data similarity in order to reduce storage space whereas the goal of cryptography is to make cipher text indistinguishable from theoretically random data. Thus, the goal of a secure deduplication system is to provide data security, against both inside and outside adversaries, without compromising the space efficiency achievable through single-instance-storage techniques.[12] proposes two approaches to secure deduplication: authenticated and anonymous. While the two models are similar, they each offer slightly different security properties. Both can be applied to single server storage as well as distributed storage. In the former, single server storage, clients interact with a single file server that stores both data and metadata. In the later, metadata is stored on an independent metadata server, and data is stored on a series of object-based storage devices (OSDs). Current systems that utilize single instance storage rely upon one of three primary deduplication strategies: whole file, fixed-sized chunks, and variable-sized chunks. The first, **whole file, typically utilizes a file's hash value as** its identifier. Thus, if two or more files hash to the same value, they are assumed to have identical contents and only stored once (not including redundant copies). This form of content addressable storage (CAS) is used in the EMC Centerasystem[12]. The second type of deduplication, per-block deduplication files are broken into fixed sized blocks before deduplication, so files that share some identical contents (but not all), may still yield storage savings. The third, and most flexible form, breaks files into variable-length **"**chunks**"** using a hash value on a sliding window; by using techniques such as Rabin fingerprints, chunking can be done very efficiently.

Today's commercial cloud storage services, such as Dropbox, Mozy, and Memopal, have been applying deduplication to user data to save maintenance cost. Convergent encryption [7] provides data confidentiality in deduplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user derives a tag for the data copy, such that the tag will be used to detect duplicates. It is assumed that the tag correctness property holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side. In [1], a new construction Dekey is proposed to provide efficient and reliable convergent key management through convergent key deduplication and secret sharing. Dekey supports both file-level and block-level deduplication.

Deduplication-based systems require solutions tailored to the type of data they are expected to handle. Data requires different degrees of protection that depend on how popular a datum is.[2] suggests that the files that undergo backup can be divided into those uploaded by many users and those uploaded by one or very few users only. Files falling in the former category will benefit strongly from deduplication because of their popularity and may not be particularly sensitive from a confidentiality standpoint. Files falling in the latter category, may instead contain user-generated content which requires confidentiality, and would by definition not allow reclaiming a lot of space via deduplication. This paper implements a multi-layered cryptosystem. All files are initially declared unpopular and are encrypted with two layers, as illustrated in Figure 2: the inner layer is applied using a convergent cryptosystem, whereas the outer layer is applied using a semantically secure threshold cryptosystem. Uploaders of an unpopular file attach a decryption share to the cipher text. In this way, when sufficient distinct copies of an unpopular file have been uploaded, the threshold layer can be removed. This step has two consequences: (i) the security notion for the now popular file is downgraded from semantic to standard convergent and (ii) the properties of the remaining convergent encryption layer allow deduplication to happen naturally. Security is thus traded for storage efficiency as for every file that transits from unpopular to popular status, storage space can be reclaimed. Once a file reaches the popular status, space is reclaimed for the copies uploaded

so far, and normal deduplication can take place for future copies.

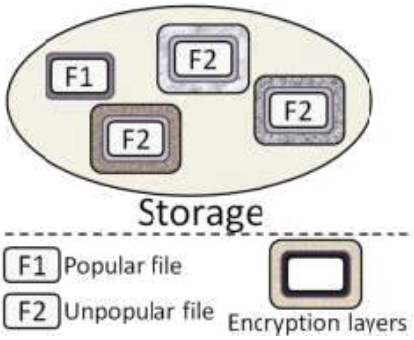



Fig. 2 MultiLayer Cryptosystem [2]

Two attacks possible against convergent encryption: confirmation of a file (COF) and learn-the-remaining-information (LRI). In COF, an attacker who already knows the full plaintext of a file, can check if a copy of that file has already been stored. If the attacker is the cloud provider or an insider, he might also learn which users are the owners of that file. In LRI, the attacker alreadyknows a big part of a file and tries to guess the unknown parts by checking if the result of the encryption matches the observed cipher text. This is the case of those documents that have a predefined template and a small part of variable content. The ClouDedup[3] scheme consists of two basic components: a server that is in charge of access control and that achieves the main protection against COF and LRI attacks; another component, named as metadata manager (MM), is in charge of the actual deduplication and key management operations. A simple solution to prevent the attacks against convergent encryption (CE) consists of encrypting the cipher texts resulting from CE with another encryption algorithm using the same keying material for all input. This solution is compatible with the deduplication requirement since identical cipher texts resulting from CE would yield identical outputs even after the additional encryption operation. The main components of this solution are (a) User

The role of the user is limited to splitting files into blocks, encrypting them with the convergent encryption technique, signing the resulting encrypted blocks and creating the storage request. In addition, the user also encrypts each key derived from the corresponding block with the previous one and his secret key in order to outsource the keying material as well and thus only store the key derived from the first block and the file identifier. For each file, this key will be used to decrypt and re-build the file when it will be retrieved. Instead, the file identifier is necessary to univocally identify a file over the whole system. Finally, the user also signs each block with a special signature scheme.

(b) Server

The server has three main roles: authenticating users during the storage/retrieval request, performing access control by verifying block signatures embedded in the data, encrypting/decrypting data traveling from users to the cloud and vice versa. The server takes care of adding an additional layer of encryption to the data (blocks, keys and signatures)uploaded by users.

Metadata Manager (MM)

MM is the component responsible for storing metadata, which include encrypted keys and block signatures, and handling deduplication. Indeed, MM maintains a linked list and a small database in order to keep track of file ownerships, file composition and avoid the storage of multiple copies of the same data segments. The tables used for this purpose are file, pointer and signature tables.

## III. AUTHORISED DATA DEDUPLICATION

The deduplication systems studied previously failed to consider the access rights of the various users storing and retrieving data from the cloud. The system under study[1] aims to develop authorized data de-duplication for protecting the data security by including differential privileges of users in the duplicate check. In such an authorized deduplication system, each user is issued a set of privileges during system initialization. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for a file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud.

There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig. 3.

- S-CSP. This is an entity that provides a data storage service in public cloud. The S-CSP provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data.

- Data users. A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.
- Private cloud. Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for **facilitating user's secure** usage of cloud service.

Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/ owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

*A. System Description*

Suppose that there are *N* users in the system and the privileges in the universe is defined as $P = \{p_1,\ldots,p_s\}$.For each privilege *p* in *P*, a private key $k_p$ will be selected. For a user *U* with a set of privileges $P_U$ , he will be assigned the set of keys $\{k_{pi}\}$*where* $p_i \in P_U$.

*1) File Uploading :*

Suppose that a data owner *U* with privilege set $P_U$ wants to upload and share a file F with users who have the privilege set $P_F = \{p_j\}$.The user computes and sends S-CSP the file token $\phi'_{F,p} = \text{TagGen}(F,k_p)$ for all $p \in P_F$ .

- If a duplicate is found by the S-CSP, the user proceeds proof of ownership of this file with the S-CSP. If the proof is passed, the user will be assigned a pointer, which allows him to access the file.

Otherwise, if no duplicate is found, the user computes the encrypted file $C_F = \text{Enc}_{CE}(k_F)$, with the convergent key $k_F = \text{KeyGen}_{CE}(F)$ and uploads $(C_F, \{\phi'_{F,p}\})$ to the cloud server. The convergent key $k_F$ is stored by the user locally.

*2) File retrieving:*

Suppose a user wants to download a file F. It first sends a request and the file name to the S-CSP. Upon receiving the request and file name, the S-CSP will check whether the user is eligible to download F. If failed, the S-CSP sends back an signal to the user to indicate the download failure. Otherwise, the S-CSP returns the corresponding cipher-text $C_F$ . Upon receiving the encrypted data from the S-CSP, the user uses the key k F stored locally to recover the original file F.

*B. Drawbacks of the system*

The flaws in the above system can be detailed as follows:

- First, each user will be issued private keys $\{k_{pi}\}$where $p_i \in P_U$ for their corresponding privileges, denoted by $P_U$. These private keys $\{k_{pi}\}$where $p_i \in P_U$ can be applied by the user to generate file token for duplicate check. However, during file uploading, the user needs to compute file tokens for sharing with other users with privileges $P_F$ . To compute these file tokens, the user needs to know the private keys for $P_F$, which means $P_F$ could only be chosen from $P_U$. Such a restriction makes the authorized deduplication system unable to be widely used and limited.

- Second, the above deduplication system cannot pre-vent the privilege private key sharing among users. The users will be issued the same private key for the same privilege in the construction. As a result, the users may collude and generate privilege private keys for a new privilege set *P** that does not belong to any of the colluded user. For example, a user with privilege set $P_{U1}$ may collude with another user with privilege set $P_{U2}$ to get a privilege set $P^* = P_{U1} \cup P_{U2}$.

- The construction is inherently subject to brute-force attacks that can recover files falling into a known set. That is, the deduplication system cannot protect the security of predictable files. One of critical reasons is that the traditional convergent encryption system can only protect the semantic security of unpredictable files.

## IV. CONCLUSION

In this paper we have explored various techniques for efficient data deduplication on the cloud. Primary purpose of deduplication is to reduce theupload bandwidth and enable efficient utilisation of the space available on the cloud servers. A primary analysis was done on the authorised deduplication which takes into account the access rights of the users.

## REFERENCES


1. J.Li, X.Chen,M. Li, J. Li, P.C. Lee, and W.Lou - ”Secure Deduplication with Efficient and Reliable Convergent Key Management”-IEEE Transactions On Parallel And Dis Tributed Systems, Vol. 25, No. 6, June 2014, pp 1615-1625

2. J.Stanek, A.Sorniotti, E.Androulaki, L.Kencl-”A Secure Data Deduplication Scheme for Cloud Storage”- Research Report-Czech Technical University, IBM Research-Zurich , Aug 2013

3. P.Puzio,S.Antipolis,R.Molva, S.Loureiro - ”ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage” - IEEE CloudCom2013, Bristol UK

4. J.Li, X.Chen,M. Li, J. Li, P.C. Lee, and W.Lou - "A Hybrid Cloud Approach for Secure Authorized Deduplication" IEEE Transactions On Parallel And Distributed Systems, Vol. 26, No. 5, May 2015

5. M.Li, C.Qin, and P.P.C. Lee -” CDStore:

Toward Reliable, Secure, and Cost-Efcient Cloud Storage via Convergent Dispersal”

6. L.G.Manzano, A. Orfila -” An efficient

confidentiality-preserving Proof of Ownership for Deduplication ” - Journal of Network and Computer Applications, 2014

7. S. Baskaran and G. Madhubal : ”File

Deduplication using Genetic Programming for Efficient Cloud Storage” - Advances in Natural and Applied Science, pp 651-660

8. M. Bellare, S. Keelveedhi, and T. Ristenpart,”Message-Locked Encryption and Secure Deduplication,” - Proc. IACR Cryptology ePrint Archive, 2012

9. FalconStar Software- How Data Deduplication Works - A white paper N.O.Agrawal, S.S.Kulkarni-”Secure

10. Deduplication And Data Security With Efficient And Reliable CEKM ” - International Journal of Application or Innovation in Engineering Management (IJAIEM)- pg 335-340

11. D.T Meyer,W.J.Bolosky. A study of practical deduplication. ACM Transactions on Storage (TOS), 2012.

12. M.W. Storer, K. Greenan, D.D.E. Long, and E.L. Miller - ”Secure Data Deduplication” in Proc. StorageSS, 2008,pp. 1-10.

# A REVIEW IN VIRTUAL PRIVATE NETWORK

**Asst.Prof. Shailesh Khandare**
Arihant College of Arts Commerce and Science

**Abstract:**

**A Virtual Private Network is Cost Effective solution for connecting to remote Personal Computer and Office over Public Network(internet) making virtual Netwok,tunnel.it is as secure,reliable,fast as just like we connecteing seprate physcial cable lines to connect office, PC to other pc,office over the Globe. we discussing in this paper what VPN,how it works,type Technolgy.**

***Keywords:***Virtual Private Network, ISDN,ISP

## INTRODUCTION

As a business grows, it might expand to multiple shops or offices across the country and around the world. To keep things running efficiently, the people working in those locations need a fast, secure and reliable way to share information across computer networks. In addition, traveling employees like salespeople need an equally secure and reliable way to connect to their business's computer network from remote locations

A virtual private network (VPN) extends a private network across a public network, such as the Internet. It enables users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network, and thus are benefiting from the functionality, security and management policies of the private network (wikipedia)

A VPN is a private network that uses a public infrastructure (usually the Internet) to connect remote sites or users. The VPN as the name suggest uses "virtual "connections routed through the Internet from the business's private network to the remote site or remote employee. It is a new technology which can be applied to LAN as well as to WLAN.A VPN maintains privacy of data through security procedures and tunneling protocols. The Protocols **IPSecVPN** or **SSL VPN** used as per specific requirment. In effect, data is encrypted at sender"s side and forwarded via "tunnel" which is then decrypted at receiver,,s side. An additional layer of security can be added by encrypting not only the data, but also the originating and receiving network addresses.

VPN was not the first technology to make remote connections. Several years ago, the most common way to connect computers between multiple offices was by using a leased line. Leased lines, such as ISDN (integrated services digital network, 128 Kbps), are private network connections that a telecommunications company could lease to its customers. Leased lines provided a company with a way to expand its private network beyond its immediate geographic area. These connections form a single wide-area network (WAN) for the business. Though leased lines are reliable and secure, the leases are expensive, with costs rising as the distance between offices increases.

Today, the Internet is more accessible than ever before, and Internet service providers (ISPs) continue to develop faster and more reliable services at lower costs than leased lines. To take advantage of this, most businesses have replaced leased lines with new technologies that use Internet connections without sacrificing performance and security. Businesses started by establishing intranets, which are private internal networks designed for use only by company employees. Intranets enabled distant colleagues to work together through technologies such as desktop sharing. By adding a VPN, a business can extend all its intranet's resources to employees working from remote offices or their homes.

### How VPNs Work

Imagine that you live on an island in a huge ocean. There are thousands of other islands all around you, some very close and others farther away. The common means of travel between islands is via ferry. Traveling on the ferry means that you have almost no privacy: Other people can see everything you do.

Let's say that each island represents a private local area network (LAN) and the ocean is the Internet. Traveling by ferry is like connecting to a Web server or other device through the Internet. You have no control over the wires and routers that make up the Internet, just like you have no control over the other people on the ferry. This leaves you susceptible to security issues if you're trying to connect two private networks using a public resource.

Continuing with our analogy, your island decides to build a bridge to another island so that people have an easier, more secure and direct way to travel between the two islands. It is expensive to build and maintain the bridge, even if the islands are close together. However, the need for a reliable, secure path is so great that you do it anyway. Your island would like to connect to yet another island that is much farther away, but decides that the costs are simply too much to bear.

This scenario represents having a leased line. The bridges (leased lines) are separate from the ocean (Internet), yet are able to connect the islands (LANs). Companies who choose this option do so because of the need for security and reliability in connecting their remote offices. However, if the offices are very far apart, the cost can be prohibitively high -- just like trying to build a bridge that spans a great distance.

So how does a VPN fit in? Using our analogy, suppose each inhabitant on your island has a small submarine. Let's assume that each submarine has these amazing properties:

- It's fast.
- It's easy to take with you wherever you go.
- It's able to completely hide you from any other boats or submarines.
- It's dependable.
- It costs little to add additional submarines to your fleet once you've purchased the first one.

Although they're traveling in the ocean along with other traffic, the people could travel between islands whenever they wanted to with privacy and security. That's essentially how a VPN works. Each remote member of your network can communicate in a secure and reliable manner using the Internet as the medium to connect to the private LAN. A VPN can grow to accommodate more users and different locations much more easily than a leased line. In fact, scalability is a major advantage that VPNs have over leased lines. Moreover, the distance doesn't matter, because VPNs can easily connect multiple geographic locations worldwide.

**]Keeping VPN Traffic in the Tunnel**

Tunneling is the process of placing an entire packet within another packet before it's transported over the Internet. That outer packet protects the contents from public view and ensures that the packet moves within a virtual tunnel.

This layering of packets is called encapsulation. Computers or other network devices at both ends of the tunnel, called tunnel interfaces, can encapsulate outgoing packets and reopen incoming packets. Users (at one end of the tunnel) and IT personnel (at one or both ends of the tunnel) configure the tunnel interfaces they're responsible for to use a tunneling protocol. Also called an encapsulation protocol, a tunneling protocol is a standardized way to encapsulate packets

The purpose of the tunneling protocol is to add a layer of security that protects each packet on its journey over the Internet. The packet is traveling with the same transport protocol it would have used without the tunnel; this protocol defines how each computer sends and receives data over its ISP. Each inner packet still maintains the passenger protocol, such as Internet protocol (IP) or AppleTalk, which defines how it travels on the LANs at each end of the tunnel. (See the sidebar for more about how computers use common network protocols to communicate.) The tunneling protocol used for encapsulation adds a layer of security to protect the packet on its journey over the Internet.

To better understand the relationships between protocols, think of tunneling as having a computer delivered to you by a shipping company. The vendor who is sending you the computer packs the computer (passenger protocol) in a box (tunneling protocol). Shippers then place that box on a shipping truck (transport protocol) at the vendor's warehouse (one tunnel interface). The truck (transport protocol) travels over the highways (Internet) to your home (the other tunnel interface) and delivers the computer. You open the box (tunneling protocol) and remove the computer (passenger protocol).

**Two VPN technologies** that are being used are:
**A)Site -to-Site VPN**
**B)Remote Access VPN**

**A)Site-to-site VPN**-

Asite-to-site VPN allows multiple offices in fixed locations to establish secure connections with each other over a public network such as the Internet. It also provides extensibility to resources by making them available to employees at other locations.

A remote-access VPN is great for individual employees, but what about entire branch offices with dozens or even hundreds of employees? Next, we'll look at another type of VPN used to keep businesses connected LAN-to-LAN.

A site-to-site VPN allows offices in multiple fixed locations to establish secure connections with each other over a public network such as the Internet. Site-to-site VPN extends the company's network, making computer resources from one location available to employees at other

locations. An example of a company that needs a site-to-site VPN is a growing corporation with dozens of branch offices around the world.

There are two types of site-to-site VPNs:

**Intranet-based** -- If a company has one or more remote locations that they wish to join in a single private network, they can create an intranet VPN to connect each separate LAN to a single WAN.

**Extranet-based** -- When a company has a close relationship with another company (such as a partner, supplier or customer), it can build an extranet VPN that connects those companies' LANs. This extranet VPN allows the companies to work together in a secure, shared network environment while preventing access to their separate intranets.

Even though the purpose of a site-to-site VPN is different from that of a remote-access VPN, it could use some of the same software and equipment. Ideally, though, a site-to-site VPN should eliminate the need for each computer to run VPN client software as if it were on a remote-access VPN. Dedicated VPN client equipment, described later in this article, can accomplish this goal in a site-to-site VPN.

**B)Remote Access VPN**

A remote-access VPN allows individual users to establish secure connections with a remote computer network. These users can access the secure resources on that network as if

they were directly plugged in to the network's servers.

A remote-access VPN connection allows an individual user to connect to a private business network from a remote location using a laptop or desktop computer connected to the Internet.

A remote-access VPN connection allows an individual user to connect to a private business network from a remote location using a laptop or desktop computer connected to the Internet.

A remote-access VPN allows individual users to establish secure connections with a remote computer network. Those users can access the secure resources on that network as if they were directly plugged in to the network's servers. An example of a company that needs a remote-access VPN is a large firm with hundreds of salespeople in the field. Another name for this type of VPN is virtual private dial-up network (VPDN), acknowledging that in its earliest form, a remote-access VPN required dialing in to a server using an analog telephone system.

There are two components required in a remote-access VPN. The first is a network access server (NAS, usually pronounced "nazz" conversationally), also called a media gateway or a remote-access server (RAS). (Note: IT professionals also use NAS to mean network-attached storage.) A NAS might be a dedicated server, or it might be one of multiple software applications running on a shared server. It's a NAS that a user connects to from the Internet in order to use a VPN. The NAS requires that user to provide valid credentials to sign in to the VPN. To authenticate the user's credentials, the NAS uses either its own authentication process or a separate authentication server running on the network.

The other required component of remote-access VPNs is client software. In other words, employees who want to use the VPN from their computers require software on those computers that can establish and maintain a connection to the VPN. Most operating systems today have built-in software that can connect to remote-access VPNs, though some VPNs might require users to install a specific application instead. The client software sets up the tunneled connection to a NAS, which the user indicates by its Internet address. The software also manages the encryption required to keep the connection secure.

Large corporations or businesses with knowledgeable IT staff typically purchase, deploy and maintain their own remote-access VPNs. Businesses can also choose to outsource their remote-access VPN services through an enterprise service provider (ESP). The ESP sets up a NAS for the business and keeps that NAS running smoothly.

**Equipment Used in a VPN**

While a VPN can be configured on generic computer equipment such as standard servers, most businesses opt for dedicated equipment optimized for the VPN and general network security. A small company might have all of its VPN equipment on site or, as mentioned earlier, might outsource its VPN services to an enterprise service provider. A larger company with branch offices might choose to co-locate some of its VPN equipment, meaning that it will set up that equipment in a co-location facility (or colo). A colo is a large data center that rents space to businesses that need to set up servers and other network equipment on a very fast, highly reliable Internet connection.

As mentioned earlier, there is no standard that all VPNs follow in terms of their setup. When planning or extending

a VPN, though, you should consider the following equipment

**Network access server** -- As previously described, a NAS is responsible for setting up and maintaining each tunnel in a remote-access VPN.

**Firewall --** A firewall provides a strong barrier between your private network and the Internet. IT staff can set firewalls to restrict what type of traffic can pass through from the Internet onto a LAN, and on what TCP and UDP ports. Even without a VPN, a LAN should include a firewall to help protect against malicious Internet traffic.

**AAA Server --** The acronym stands for the server's three responsibilities: authentication, authorization and accounting. For each VPN connection, the AAA server confirms who you are (authentication), identifies what you're allowed to access over the connection (authorization) and tracks what you do while you're logged in (accounting).

One widely used standard for AAA servers is Remote Authentication Dial-in User Service (RADIUS). Despite its name, RADIUS isn't just for dial-up users. When a RADIUS server is part of a VPN, it handles authentication for all connections coming through through the VPN's NAS.

VPN components can run alongside other software on a shared server, but this is not typical, and it could put the security and reliability of the VPN at risk. A small business that isn't outsourcing its VPN services might deploy firewall and RADIUS software on generic servers. However, as a business's VPN needs increase, so does its need for equipment that's optimized for the VPN. The following are dedicated VPN devices a business can add to its network. You can purchase these devices from companies that produce network equipment, such as Cisco:

**VPN Concentrator** -- This device replaces an AAA server installed on a generic server. The hardware and software work together to establish VPN tunnels and handle large numbers of simultaneous connections.

**VPN-enabled/VPN-optimized Router** -- This is a typical router that delegates traffic on a network, but with the added feature of routing traffic using protocols specific to VPNs.

**VPN-enabled Firewall** -- This is a conventional firewall protecting traffic between networks, but with the added feature of managing traffic using protocols specific to VPNs.

**VPN Client** -- This is software running on a dedicated device that acts as the tunnel interface for multiple connections. This setup spares each computer from having to run its own VPN client software

**Encryption and Security Protocols in a VPN**

let's take a closer look at the encryption and protocols that VPN components use.

**Encryption** is the process of encoding data so that only a computer with the right decoder will be able to read and use it. You could use encryption to protect files on your computer or e-mails you send to friends or colleagues. An encryption key tells the computer what computations to perform on data in order to encrypt or decrypt it. The most common forms of encryption are symmetric-key encryption or public-key encryption:

In **symmetric-key encryption**, all computers (or users) share the same key used to both encrypt and decrypt a message.

In **public-key encryption,** each computer (or user) has a public-private key pair. One computer uses its private key to encrypt a message, and another computer uses the corresponding public key to decrypt that message.

In a VPN, the computers at each end of the tunnel encrypt the data entering the tunnel and decrypt it at the other end. However, a VPN needs more than just a pair of keys to apply encryption. That's where protocols come in. A site-to-site VPN could use either Internet protocol security protocol (IPSec) or generic routing encapsulation (GRE). GRE provides the framework for how to package the passenger protocol for transport over the Internet protocol (IP). This framework includes information on what type of packet you're encapsulating and the connection between sender and receiver.

**IPSec** is a widely used protocol for securing traffic on IP networks, including the Internet. IPSec can encrypt data between various devices, including router to router, firewall to router, desktop to router, and desktop to server. IPSec consists of two sub-protocols which provide the instructions a VPN needs to secure its packets:

Encapsulated Security Payload (ESP) encrypts the packet's payload (the data it's transporting) with a symmetric key.

Authentication Header (AH) uses a hashing operation on the packet header to help hide certain packet information (like the sender's identity) until it gets to its destination.

Networked devices can use IPSec in one of two encryption modes. In transport mode, devices encrypt the data traveling between them. In tunnel mode, the devices build a virtual tunnel between two networks. As you might guess, VPNs use IPSec in tunnel mode with IPSec ESP and IPSec AH working together [source: Friedl].

In a remote- access VPN, tunneling typically relies on **Point-to-point Protocol (PPP)** which is part of the native protocols used by the Internet. More accurately, though, remote-access VPNs use one of three protocols based on PPP:

L2F (Layer 2 Forwarding) -- Developed by Cisco; uses any authentication scheme supported by PPP

PPTP (Point-to-point Tunneling Protocol) -- Supports 40-bit and 128-bit encryption and any authentication scheme supported by PPP

L2TP (Layer 2 Tunneling Protocol) -- Combines features of PPTP and L2F and fully supports IPSec; also applicable in site-to-site VPNs

IPSec and SSL VPN are most papular protocls

**Features in VPN**

- ·Provide extended connections across multiple geographic locations without using a leased line.
- ·Improved security mechanism for data by using encryption techniques.
- ·Provides flexibility for remote offices and employees to use the business intranetover an existing Internet connection as if they're directly connected to the network
- ·Saves time and expense for employees who commute from virtual workplaces

**CONCLUSION**

This Paper Specifies, How VPN is cost effective,reliable soultion for us over ISDN

without compromising securities to connect remote office and Indiual user over the globe.

**REFERENCES**

1] Wikipedia (www.en.wikipedia.org/wiki/vpn)

2]Cisco. "How Virtual Private Networks Work." Oct. 13, 2008. (April 4, 2011)http://www.cisco.com/application/pdf/paws/14106/how_vpn_works.pdf

3]Friedl, Stephen J. "Steve Friedl's Unixwiz.net Tech Tips: An Illustrated Guide to IPSec." Aug. 24, 2005. (April 4, 2011)http://www.unixwiz.net/techtips/iguide-ipsec.html

4]Microsoft. "TechNect: VPN Tunneling Protocols." 2011. (April 3, 2011)http://technet.microsoft.com/en-us/library/cc771298(WS.10).aspx

5] Pandya, Hiten M. "FreeBSD Handbook: Understanding IPSec." The FreeBSD Documentation Project. (April 4, 2011)http://www.freebsd.org/doc/en_US.ISO8859-1/books/handbook/ipsec.html

# Quality improvisationin Game design techniques throughGame Development life Cycle (GDLC) Model.

**Author: Prof Sarita A Dhawale**
Department of Computer
Ashoka Center for Business And Computer Studies, Nashik
Savitribai Phule Pune University.

**CoAuthor:**
Mr Kamlesh Dubey
Ashoka Center for Business And Computer Studies, Nashik

**ABSTRACT:**

**Game is a kind of software with goal to provide entertainment to the youngster's. It is a form of competitive activity or sportplayed according to rules.Player eager or willing to do something new or challenging with defined set of rules.The application of typical elements of game playing (e.g. point scoring, competition with others, rules of play) to other areas of activity, typically as an onlinemarketing technique to encourage engagement with a product or service: gamification is exciting because it promises to make the hard stuff in life fun .Gamification has become a trend in recent years for youths. More and more companies recognize the prioritization & benefits and have started to apply game elements for their internal and external business processes. Gamification has also been addressed by the different software development phases in order to influence youth's behaviour in a positive manner. Different Gamification tools exist to support requirements gathering, Game development and Game testing allowing improving the overall quality of software engineering.**

**This research paper consists of a literature study on Gamification and how its principles may be successfully applied to the process of software development. However, during the real game development practice, simply adopting the software development life cycle (SDLC) is not enough, as the developers face several challenges during its life cycle. To address the problem, game development companies uses a kind of specific approach called game development life cycle (GDLC) to direct the game development. However, none of the existing GDLCs explicitly address how to successfully deliver a good quality game and a cross-integrated gaming system over the different phases has not been encountered in practice yet. This paper presents a new game development life cycle model and guidelines to successfully deliver a good quality game. Several quality criteria's are explicitly considered at each phase.**

**Keywords**: Gamification, GDLC (Game Development Life Cycle)

## 1. INTRODUCTION:

Stating that games are engaging is becoming increasingly redundant. In a report published in April 2014, the Entertainment Software Association claims that 59% of Americans play video games [1]. From the same report, it can be concluded that most notably casual games are making this form of entertainment more and more accessible. Looking at this trend, it is only logical that it would be desirable to take some elements of what is driving this popularity, and implementing it in other situations. This so-called gamificationhas already been successfully adopted in several large organizations with many positive results, including better long-term collaboration, creativity, productivity, loyalty and learning [2]. On the other side, many software projects nowadays still fail to deliver on time, within cost and with expected scope. One main reason for project failure is unsatisfied or not met requirements, often caused by poor stakeholder collaboration, incomplete understanding of needs, insufficient domain knowledge, fluctuating requirements and others [3].

Therefore, this paper will focus specifically on the software development process and how it can be improved through gamification. The promises of gamificationin this context include a reduction in the amount of time necessary for the project, a decreased amount of errors, redesigns and development dead-ends [2]. Indeed, the main point of this paper will be on the

quality of the process, rather than that of the resulting product. The state-of-the art of gamification will be thoroughly explored, starting by introducing all the major concepts. Using this basis, the paper continues by explaining the term of gamification, followed by a discussion of different types of game elements. Furthermore, the use of gamificationin software engineering is demonstrated, including real-life examples as well as a conceptual model approach for integrating it over the entire software development lifecycle.

## 2. CONTEXT & SIGNIFICANCE OF THE STUDY:

In order to study gamificationin software engineering, the aim and scope of the study needs to be defined. Based on existing literature, it can be concluded that gamification has already been successfully applied in different work situations and has improved the engagement and productivity of employees. In this study, however, the focus is specifically on software development.

The aim is to provide a clear overview of exactly how gamificationcan improve the process quality. Therefore, the main research question of the study is: in what way does gamification enhance the software engineering process? In order to answer this question, the following sub-questions were formulated to get the conceptual answers for the same.

- What is gamification?
- What are the game elements that can be used for software engineering?
- What are today are best practices of gamification in software engineering?
- How can gamification be applied over the entire development lifecycle?

## RESEARCH METHODOLOGY:

Software Development life Cycle is (SDLC) a conceptual study used to apply in the gaming industry to convert it in to GDLC i.e. Game Development Life Cycle to improve the process management of Gaming industry .The study aims at understanding the essence of improvisation of Gaming process .The conceptual analysis is completed through subjective study of the impact and essence of GDLC. The study also focuses on specific booming industry of Gaming rather than general software industry. This study is conceptual and subjective in nature and may only derive further understanding of GDLC model.

## 3. LITERATURE REVIEW

To get a clear picture of the current state of gamificationas well as the process of software engineering, done the literature study. For this, different methods were used for multiple sources. The primary source of information was scientific literature. Using Google Scholar, relevant scientific journals and articles were identified. A side from these, a number of books were found pertaining to either gamificationor the software development process. Among these less formal sources were statements made by so-called GamificationGurus, who are experts in the field. The process of selecting the most relevant source material was done in several steps. References used to gain a more complete picture of the scientific developments regarding gamification.For statements made by GamificationGurus, their status as an expert played an important role. Back references were used to gain a more complete picture of the scientific developments regarding gamification. In turn, stories from companies where gamification had been successfully implemented were studied in such a way that they would provide insight into the process, benefits and best practices.

## 4. STATE-OF-THE-ART STUDY:

The state-of-the-art literature study first clarifies the term “gamification” and its principles. Next, the different game elements are introduced and how they are currently being applied in software engineering. The last part of the literature study examines a cross-integrated gamified solution in the software engineering domain.

### 4.1. GAMIFICATION

Games have been around for thousands of years and their main purpose is to entertain people and create pleasure. People nowadays are dedicating a great amount of their time to video games [5]. However, with gamification it is possible to engage users in solving real-world problems, which activities and outcomes are value-adding and not just a waste of time. Furthermore, many research papers have shown that when people are engaged, they are likely to be more productive, resulting in higher quality outcomes [9]. Gamificationcan be used in environments where it is difficult to influence people’s behavior and improve their motivation and engagement. The vertical axis represents the purpose and fun dimension and the horizontal axis game play [11]. The main difference is that games are just for fun and entertaining, whereas gamification has a certain purpose, without any gameplay. Figure 1 shows the different types of game thinking and where gamification is classified.

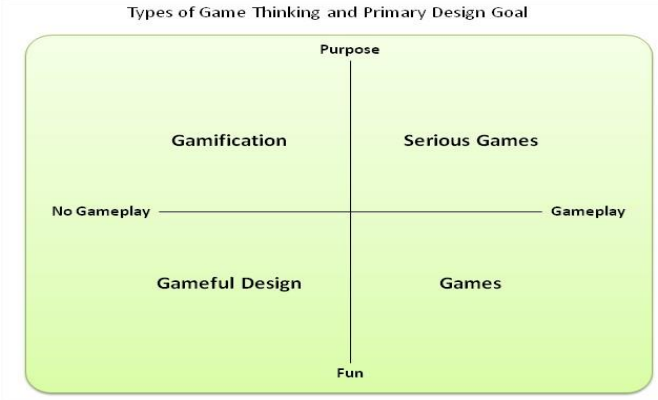


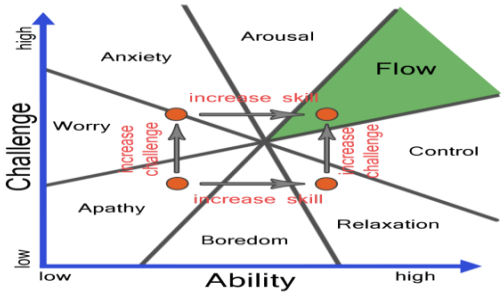

Figure 1.

4.2 **GAME ELEMENTS**:

Gamification does not create games, but rather takes certain game elements from classical video games and applies them to real-world (business) problems. These game elements are typically known as points and keeps players engaged and motivated in their activities [12]. Points are responsible to keep track of score, provide feedback, and display progress, whereas badges are representations of achievements. Leaderboards allow users to directly compare themselves in a ranking. More useful game elements that were very often encountered in the literature study are stories, quests, levels, rewards and avatars [13]. Stories are the background narratives and create the dramatics and arouse emotions. A quest is an assignment or a task with a clear goal that must be accomplished by the user. Rewards can exist in form of items or virtual currencies and are presented to the player when a certain task is achieved. Levels help to make tasks more challenging as the user gains new skills and experiences. From a motivational perspective, levels can help to maintain user engagement for an optimal experience. On the other side, if the game is too easy, the player will be bored and stop playing the game as well. Therefore, the balance of these two dimensions is important to keep players engaged. *Figure 2* shows the different dimensions and aspects of the flow model.

An avatar allows the user to create an individual virtual character with different types of attributes that can be personalized to his interests. Some attributes of the avatar might enhance while the user levels-up in a game. The usage of avatars can help to increase performance and motivation [15].

In addition to the aforementioned elements, feedback loops are an essential part of gamification. They provide information about how users are doing in their process. For instance, a score list reflects clear and distinctive information to the player about their current status. The feedback is what creates the motivation for further actions and keeps users engaged. The feedback loop is to be considered less as a game element, but more as a game mechanic, because it ties several game elements together.The Flow Model :Figure 2.




**4.3. GAMIFICATION IN SOFTWARE ENGINEERING**

Now that the potential of the different types of game elements are unveiled, this part will demonstrate how they are applied in software engineering. To incorporate different aspects of gamification in to a company's software development lifecycle (SDLC), it is important to focus on the organization's internal perspective. A successful system should increase software engineers' work performance and motivation, resulting in higher software quality, with less changes and defects. The standard process of a software development project consists of five phases. These are 'requirements definition', 'system and software design', 'implementation and unit testing', 'integration and system testing' and 'operations and maintenance', as shown in SDLC Figure 3

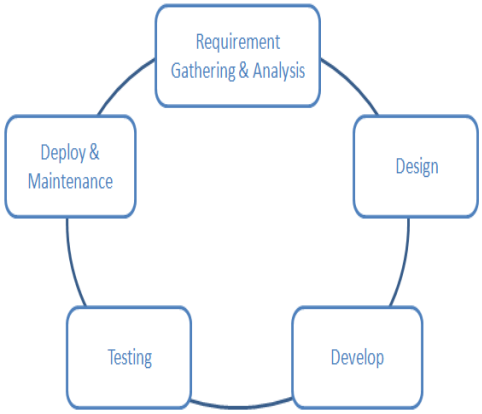



Figure 3: Software Development Life Cycle.

In the analysis phase, the system's functionalities, goals and constraints are established in form of requirements. Next, architecture is constructed based on the identified requirements. In phase three, the system is developed and tested in a set of program units. The separate programs will then be integrated as a complete system and tested against the requirements. Finally, the system will be installed in the customer's operational environment. This phase involves fixing errors and bugs, which were not detected during testing phases. Some researchers and companies have already gamified certain parts of the development process. This part will elaborate on some successful practices of the different phases.

Fig. 4: Software design Model

**4.3.1 Requirements Definition: iThink**

iThink is a gamified approach to improve the collaboration and participation in the phase of requirements definition [19]. The authors analyzed several requirement elicitation practices and then developed a system with the appropriate game elements. The tool is based on the theory of the six thinking hats [20]. The idea

behind this method is to stimulate parallel thinking in teams in order to be more effective and avoid team conflicts. Each hat has a different style of thinking and is represented by a unique color. This approach can be used to discover new requirements and discus existing ones. The main game elements used for iThink are points and progress bars. Points can be acquired through different user actions, which are mapped to one of the hats. For instance, a user can win points by submitting a new requirement, rating a requirement, commenting on a requirement and completing a discussion of a requirement. *In Figure 4,* an example screen can be seen, in which the user is asked to reflect on a requirement in four different ways, each awarding a certain amount of points.

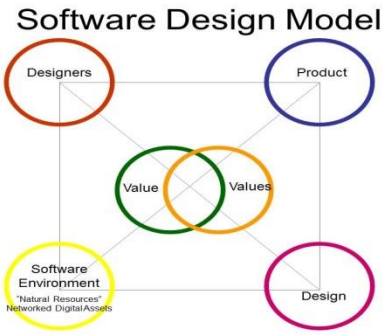



*Figure 4*

**4.3.2 Software development:**

In the coding domain, many educational web-based platforms exist, targeting students and teachers to learn and advance their programming skills. For instance, Code Hunt was release by Microsoft Research and uses puzzles to encourage players to learn how to code [21]. The system uses the feedback mechanism as a gameplay. Every puzzle must be solved with a piece of written code that is verified by the system. If the code is wrong, the system returns an error. If the code is correct, the player wins the quest and continues to the next level. As the player keeps playing, his skill rating goes up and the quests get more difficult, thereby maintaining 'flow'.

**4.3.3. TESTING: TO GET QUALITY GAME.**

The behaviors organizations want to evoke for testers, is to find as many bugs and defects as possible and report them properly. For many people this task might seem boring. However, game elements can make testing more entertaining, just as with other applications. Much of what testers do can be thought of in gaming terms.

**4.3.4 CROSS-INTEGRATED GAMIFICATION SOLUTION IN SDLC**

The software development process has already been gamified. However, these phases are still separated and not yet cross-integrated into one gamified environment in form of an ecosystem. The author introduced several game elements and analyzed how they can be utilized to make software building more productive and fun. He proposes a multi-layer architecture containing different game elements, as shown in Figure 5. The story element is located in the bottom layer and serves as the basis for the system over the entire SDLC. Business requirements, project context and organizational culture can be part of the story. Next, epic-quests, quests and achievements are built upon the story to give players a clear goal. Quests should be constructed depending on the particular software phase. The third layer consists of levels, points and rewards, which are dependent on quests and achievements. Solving quests not only increases the level, but also allows earning more points and rewards. The avatar is positioned on the top of the stack and takes part of the entire development process. The avatar's characteristics are based on the level, as well as points and rewards achieved during the project lifecycle. For instance, as the player earns more points, the stronger and better his avatar gets.

| Avatar | | |
|---|---|---|
| Level | Points | Reward |
| Epic-Quest | Quest | Achievement |
| Story | | |

Figure :5 Game Elementation

**5. CONCLUSIONS:**

1: In this research the hidden potential and increased business attention of gamificationfor software engineering was highlighted. Not only can gamification support the enhancement of user motivation and engagement, but also the overall productivity that correlates with performance of the user. Applying designated game elements to the software development lifecycle can have a positive effect on the software quality as well. Therefore, software projects are more likely to be delivered within time, budget and with fewer defects. The following conceptual model illustrates in the figure 6.shows the relationships between the different concepts of gamification.

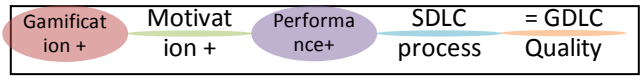
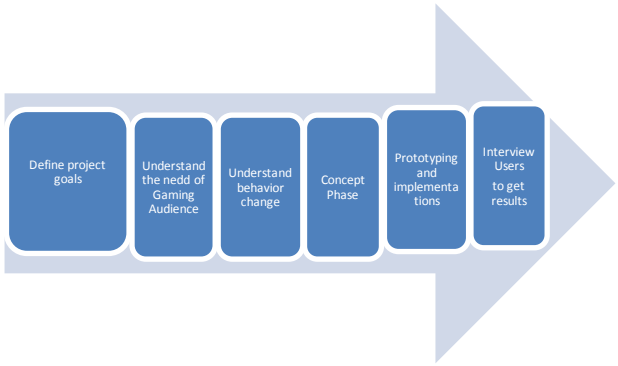



Figure 6: Self designed Conceptual Model from objectives .

2: Many game elements are obtained from classical video games and can be incorporated into existing systems. However, there is no complete answer to which elements should be used. The decision depends on the circumstances of the business environment, as well as the players interacting with the system.

3: Today, many approaches exist to apply gamification to the software development process. The results show positive effects on people's behavior and allow increasing the performance of GDLC. However, there is no proof of concept that demonstrates how a cross-integrated system can be gamified over the entire development lifecycle. But following Model suggests the best result for Gamification cycle. (Figure 7)

Figure 7: GDLC Model.




In this process, it is important to emphasize that the result of each of the phases should correspond with the rest. For instance, a company may want to implement a leaderboard, as was also described in the first case, however with the aim to improve cooperation. Obviously, this would not achieve this goal. It is crucial that the effect of each game element and mechanic are explained clearly and applied appropriately. Another important point of attention is the audience, as this defines what game elements may or may not work. This has to do with what motivates different kinds of people.

For example, older people will be less likely to pay attention to receiving a special digital badge than younger people, making this element less effective. In this case, it is a good idea to apply gamification without computers. Another factor that can change how people experience different game elements, or even simply how they cooperate, is culture. The example that was mentioned was the difference in working culture between Western and Asian countries. When designing a proposal for the system, this should definitely be taken into account. This aspect of gamification when discussing the different kinds of motivation. In explaining what the process entails, it is important to emphasize that what is asked for, is not necessarily what is needed to achieve the desired result.

Gamification is all about fine-tuning and selecting the right concepts and technologies. People's behavior must be continuously measured and improved when necessary. However, gamification is only one element of an enterprise system and cannot stand on its own and solve all problems. This research aimed at defining the term gamification and its principles together with the different types of game elements. Furthermore, an examination of literature was conducted to evaluate the current status of gamification in software engineering. The phenomenon has been successfully used in the last few years with an upward trend to change people's behavior and improve work performance .Applying game design techniques can motivate people to improve the complex software development process across the lifecycle. It is already used in all different phases of software engineering

The following table lists the summaries some of the suggestions to the companies who are interested in Gaming industry.

| **Area** | **Suggestions** |
|---|---|
| Focus | • Can have more on gamification within software engineering |
| Game Elements | ▪ Should not be limited to gaming elements on screens.<br>▪ Makes a clear distinction between extrinsic and intrinsic motivation. |
| Gamifying a system | ▪ First understand the audience and the problem & then game design.<br>▪ Elements requested by client may not be what is needed to achieve the desired effect. |
| Users of the system | ▪ Can vary greatly<br>▪ Prefers the term users instead of players.<br>▪ Uses personas to identify players' characteristics and motivations. |
| Gamification in SDLC | ▪ Do not have direct experience with gamifying SDLC.<br>▪ Would approach it in the same way as other projects, no reason to change this.<br>▪ Demonstrated more existing solutions that are in SDLC related phases, such as knowledge sharing and documentation |

**6. References :**

[1] "Essential Facts about the Computer and Video Game Industry," 2014. [Online]. Available: http://www.theesa.com/facts/pdfs/esa_ef_2014.pdf. [Accessed: 30-Sep-2014].
[2] M. Herger, Enterprise Gamification: Engaging people by letting them have fun (Volume 1), 1st ed. CreateSpace Independent Publishing Platform, 2014.
[3] T. Tsumaki and T. Tamai, "Framework for matching requirements elicitation techniques to project characteristics," Softw.Process Improv.Pract., vol. 11, no. 5, pp. 505–519, 2006.
[4] J. Hamari, J. Koivisto, and H. Sarsa, "Does Gamification Work? A Literature Review of Empirical Studies on Gamification," in System Sciences (HICSS), 2014 47th Hawaii International Conference on, 2014, pp. 3025–3034.
[5] J. McGonigal, Reality is broken: Why games make us better and how they can change the world. Penguin, 2011.
[6] "Time spent with video games in the U.S. 2013 | Statistic." [Online]. Available: http://www.statista.com/statistics/186960/time-spent-withvideogames-in-the-us-since-2002/. [Accessed: 13-Sep-2014].
[7] "Active video game players and payers 2012 | Country ranking." [Online]. Available: http://www.statista.com/statistics/195768/active-video-gameplayers-and-payers-in-selected-countries/. [Accessed: 13-Sep-2014].
[8] K. Salen and E. Zimmerman, Rules of play: Game design fundamentals. MIT press, 2004.
[9] C. M. Ellis and A. Sorensen, "Assessing employee engagement: the key to improving productivity," Perspectives (Montclair)., vol. 15, no. 1, pp. 1–9, 2007.
[10] G. Zichermann and C. Cunningham, Gamification by design: Implementing game mechanics in web and mobile apps. " O'Reilly Media, Inc.," 2011.
[11] A. Marczewski, "Gamasutra: AndrzejMarczewski's Blog - What's the difference between Gamification and Serious Games?," 2013. [Online]. Available: http://www.gamasutra.com/blogs/AndrzejMarczewski/20130311/188218/Whats_the_difference_between_Gamification_and_Serious_Games.php. [Accessed: 14-Sep-2014].
[12] K. Werbach and D. Hunter, For the win: How game thinking can revolutionize your business. Wharton Digital Press, 2012.
[13] M. Fecher, "Gamification in der Softwareentwicklung: Chancen und Möglichkeiten," Retrieved August, vol. 5, p. 2013, 2012.
[14] M. Csikszentmihalyi, Beyond boredom and anxiety. Jossey-Bass, 2000.
[15] S. F. van der Land, A. P. Schouten, F. Feldberg, M. Huysman, and B. van den Hooff, "Does Avatar Appearance Matter? How Team Visual Similarity and Member--Avatar Similarity Influence Virtual Team Performance," Hum. Commun.Res., 2014.
[16] "Game Mechanics | Gamification.org," 2014. [Online]. Available: http://badgeville.com/wiki/Game_Mechanics#Game_Mechanics_Shortcuts. [Accessed: 15-Sep-2014].
[17] C. R. Prause, "Improving the internal quality of software through reputationbasedgamification," Universit{ä}tsbibliothek, 2013.
[18] I. Sommerville, Software Engineering, 9th ed. Boston: Addison-Wesley, 2010, p. 792.
[19] J. Fernandes, D. Duarte, C. Ribeiro, C. Farinha, J. M. Pereira, and M. M. da Silva, "iThink: A Game-Based Approach Towards Improving Collaboration and Participation in Requirement Elicitation," ProcediaComput. Sci., vol. 15, pp. 66–77, Jan. 2012.
[20] E. De Bono, Six thinking hats. Taylor & Francis, 1999.
[21] N. Tillmann, J. Bishop, R. N. Horspool, D. Perelman, and T. Xie, "Code hunt: searching for secret code for fun," in Proceedings of the 7th International Workshop on Search-Based Software Testing, 2014, pp. 23–26.
[22] K. Januszewski, "Announcing Visual Studio Achievements Beta," 2012. [Online]. Available: http://channel9.msdn.com/Blogs/C9team/AnnouncingVisual-Studio-Achievements. [Accessed: 15-Sep-2014].
[23] J. Williams and S. Ross, "Score One for Quality! Using Games to Improve Product Quality," Redmond, 2010.
[24] D. J. Dubois and G. Tamburrelli, "Understanding gamification mechanisms for software development," in Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, 2013, pp. 659–662.
[25] "Gamification Gurus | Leaderboard," 2014. [Online]. Available: https://www.leaderboarded.com/gurus. [Accessed: 07-Oct-2014].
[26] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," Contemp. Educ. Psychol., vol. 25, no. 1, pp. 54–67, 2000.

# A Watermarking Technique Without Data Modification for Process Mining

Tintu Mary Xavier
Assistant Professor in CSE
College of Engineering, Poonjar

***Abstract***

**Watermark, usually applied on the existing datasets to identify authenticity. But the watermarking usually requires the modification of attributes. Here, trying to utilize the option of matching data by using common characteristics found within the data set in the relational datasets. The proposed project is to create an innovative system to identify the fake tuples through which the data owner can easily identify the fake users re-using the distributed datasets. Uncontrolled data leakage put business in a vulnerable position. Data allocation strategies are proposed to improve the probability of identifying leakages. Storing data in a de-normalized way enhance security. Amending the fake tuples relevant to the data owner is one of the main ideas to safeguard the data. This can be automated by providing an innovative model of amending this kind of data in it. Modifying the existing data is another option by creating a clear watermark on the data and it should be easily recoverable one. Privacy on the existing data by cut shorting the columns from user view makes the column looks more secure one.**

## I.INTRODUCTION

A data owner shares his datasets with mining agents to extract information from a datasets and transform it into an understandable structure for further use. During this sharing of datasets there may be a chance of dataset leakage and led the business into a vulnerable position. If cyber criminal cash out or sell this data for profit it leads to organization downfall. Therefore watermark has been applied for protecting ownership on the datasets.

The information to be embedded in a digital data in different formats such as audio, video, image, relational database, text and software is called a watermark .The signal in which the watermark is embedded is called host signal. A watermarking scheme consists of three steps, insertion, attack and detection. In insertion algorithm the host and the data to be embedded is selected and produces a watermark signal. Then the watermarked signal is transmitted to another person, and if he makes any modification then it is called an attack. Detection checks whether the watermark inserted was modified or not after the transmission. Watermarking is very useful in protecting ownership but it involves some modification of the original data and it can be destroyed if the data recipient is malicious.

The data can be secured more by the amendment of fake tuples. It improves the possibility to identify the agent that leaked the data. The fake tuples can be considered as a type of watermark for the entire set and it doesn't need modification of any individual values. If it proves an agent was given one or more fake objects that were leaked, then the distributor can be making sure that agent was guilty.

## II.NEW WATERMARKING MODEL

The core idea of proposed work is preserving the data from exposing to the end user. To achieve this an automated system is proposed which will take care of injecting the data in a fashioned manner. A recognizable watermarking pattern based on the usability constraints has the property of robustness and flexibility as well as safeguard the data based on the user preference. Data owner act in the core part to identify the fake users reusing distributed datasets thereby using the automated system that identifies the fake tuples, this in turn prevents the leakage of data. A strategy named data allocation is proposed to improve the probability of identifying the leakage.

Preserving main columns from exposing for mining is the next level of security provided on the data. To enhance security, data will be stored in a denormalized way. An automated way of handling the data selection for

distribution is proposed in this work. Morphing sensitive data and Injecting row wise tuples using an automation process with a clear picture of data changes in the system is one of the newest and proposed features in this work.

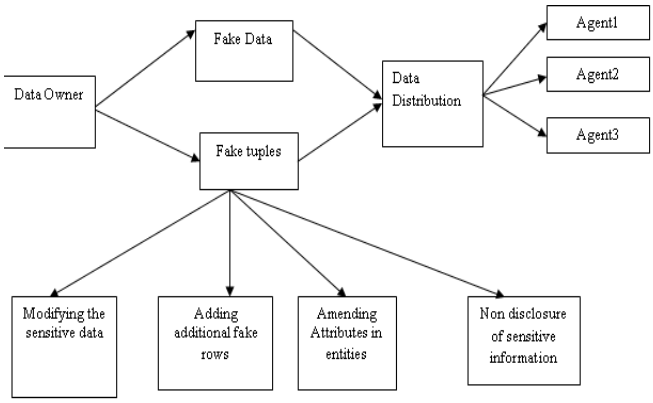



Fig. 1 Architecture diagram of the proposed system

The fig. 1 shows the architecture diagram of the proposed system. A data owner enters into the system selects a distributor from the distributor list and the data to be transferred based on the user preference. Based on the selection automatic queries will be generated to retrieve the datasets. A unique watermarking will be applied on the selected datasets. Addition of fake tuple is one of the watermarking strategies. The following algorithm explains the amendment of fake tuples to the distributed sets.

Algorithm1:

Input: - D: dataset of the agent.
N': the number of fake tuples.
A: the set of attributes that are not Candidate keys.
Output:-D: dataset with fake tuples.
1. for each $a_i$ in A
a) Read a fake data from the owner b) Generate n fake values by adding different characters to the fake data. The generated value should be distinct.
2. Set of fake records, F with n tuples is generated.
3. While n>0 do
f= select a fake tuple at random from set F
D= D U {f}

F= F-{f}
n=n-1

The data is modified in such a way that it should be easily recoverable as well as feature should not be altered much. The identification of more sensitive information and not revealing it to the agent, makes the datasets more secured one.

After applying the watermark, the data is subjected for distribution. A mode of data processing whereby a terminal or group of terminals serves as a receptacle for data supplied by the CPU. The Data Distribution Service for Real-Time Systems (DDS) is an Object Management Group (OMG) machine to machine middleware standard that aims to enable scalable real-time dependable high performance and inter operable data exchanges between publishers and subscribers. The data are sent to the data distributors with unique watermarking. The distributors proceeds with the data on passing it to the different receiver ignorance of the watermarking. The distribution is based on the fixed grid DDM algorithm. The routing space is partitioned into non-overlapping grid cells, and a multicast group is defined for each cell. A federate subscribes to the group associated with each cell that partially or fully overlaps with its subscriber regions. The result associates a region with several multicast groups in a fixed and pre-determined manner. A publish operation is realized by sending an update message to the multicast groups corresponding to the cells that partially or fully overlap with the associated publisher region. The fixed-grid approach eliminates the need to explicitly match publisher and subscriber regions.

The watermark detection algorithm checks whether the distributed datasets has been modified or not. The data is examined carefully for accuracy with the intention of verification. Robustness requirements of watermarks mandate that any attempt to remove or destroy the watermark should produce a remarkable degradation in data quality before the watermark is lost. A master copy of the data cum watermark data will be maintained by the Distributor and the watermarked original copy will be matched with the agent's copy to identify the threat in data leakage. The focus is to find the exact guilt.

The proposed model provides data security in high level and identification of leakage using a data allocation strategy is an effective way to safeguard the data. Unlike other models, this provides security at data level as well as column level. Usage of automated system drives the work more perfect than the manual process used in the existing.

## III. CONCLUSION

In this paper, we have developed a novel yet pragmatic framework for watering. It includes carefully designed, shifting and clustering in each of which handles a specific problem in watermarking. Fixed grid pre-processes tend to host images by adjusting the pixels into a reliable range for satisfactory reversibility. Datasets are included in shifting and clustering constructs of new watermark embedding and extraction processes for good robustness and low run-time complexity in the precise estimates the local sensitivity of grids and adaptively optimizes the watermark strength for a trade-off between robustness and invisibility. In contrast to representative methods, the proposed framework:

1) obtains comprehensive performance in terms of reversibility, robustness, invisibility, capacity and run-time complexity;

2) is widely applicable to different kinds of images; and

3) is readily applicable in practice.

## REFERENCES

[1] R. Agrawal, P. Haas, and J. Kiernan, "Watermarking relational data:Framework, algorithms and analysis," *The VLDB Journal*, vol. 12, no.2, pp. 157–169, 2003.

[2] J. Palsberg, S. Krishnaswamy, M. Kwon, D. Ma, Q. Shao, and Y.Zhang, "Experience with software watermarking," in *Proc. 16th Ann.Computer Security Applications Conf.*, 2000, pp. 308–316.

[3] M. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, and S. Naik, "Natural language watermarking: design,
analysis, and a proof-of-concept implementation," in *Information Hiding*. New York, NY, USA: Springer, 2001, pp. 185–200.

[4] Mayuree K.Rathva, Prof. G. J. Sahani, "Watermarking relational databases," *International Journal of Computer Science, Engineering and Applications (IJCSEA)* Vol.3, No.1, February 2013.

[5] M. Kamran, M. Farooq, "A formal usability constraints model for watermarking of outsourced datasets", *IEEE Transactions On Information Forensics And Security*, Vol. 8, No. 6, June 2013.

[6] Umamaheswari, S. Getha , H. A, "Detection of guilty agents", *Innovations in Emerging Technologies(*NCOIET*)*, 2011, pp. 23-26

[7] Cui, Yingwei, and Jennifer Widom. "Lineage tracing for general data warehouse transformations." *The VLDB Journal—The International Journal on Very Large Data Bases* 12.1 ,2003, pp. 41-58.

[8] Pankaj Gupta, Ratan K. Guha, " Design, Analysis, and Performance Evaluation of an Efficient Algorithm for Data Distribution Management in High Level Architecture," *CS-TR*-05-12, December 2005.

# Video Inpainting Forgery Detection: Blind Detection Method

**Sree lakshmi Das**[*1],
*Department of Computer Science and Engineering*
*Kerala University*
*Sree Buddha College of Engineeing*
*Pattoor, Alappuzha*
[1]sreelakshmidas89@gmail.com

*ABSTRACT*
**Video forgery, also referred as video falsifying, is a technique for generating fake videos by altering, combining or creating new video contents. Exemplar-based inpainting technique can be used to remove objects from an video/image and play visual tricks, which would affect the authenticity of videos. In this paper, a blind detection method based on zero-connectivity feature and fuzzy membership is proposed to detect the video forgery. Firstly, zero-connectivity labeling is applied on block pairs to yield matching degree feature for all blocks in the region of suspicious, and fuzzy memberships are computed by constructing ascending semi-trapezoid membership function. Then the tampered regions are identified by a cut set. A number of natural and inpainted forged videos are used to show the effectiveness of our method.**

***Keywords*—Video forgery; Exemplar-based inpainting; Zero-connectivity labelling; Fuzzy Membership**

## 1. INTRODUCTION

Although the research of altering contents in a video may create a potential sociological problem, it is interesting and challenging to investigate video forgery, also referred as video falsifying, technologies if they are used with good intension (e.g., special effects of a movie). Digital image/video tampering or forgery has become major problem lately, due to ease of artificially synthesizing photographic fakes - for promoting a story by media channels and social networking websites. This is due to significant advances in computer graphics and animation technologies, and availability of low cost off-the-shelf digital video/image manipulation and cloning tools. With lack of proper regulatory frameworks and infrastructure for prosecution of such evolving cyber-crimes, there is an increasing dissatisfaction about increasing use of such tools for law enforcement, and a feeling of cynicism and mistrust among the civilian operating environments. The availability of powerful digital image processing tools, such as photoshop, makes it relatively easy to add or remove important features from an image without leaving any obvious traces of tampering.

Digital video/ image tamper detection can use either active tamper detection techniques or passive tamper detection techniques. A significant body of work, however is available
on active tamper detection techniques, which involves embedding a digital watermark into the images when the images are captured. The problem with active tamper detection techniques is that not all camera manufacturers embed the watermarks, and in general, most of the customers have a dislike towards cameras which embed watermarks due to compromise in the image quality. So there is a need for passive and blind tamper detection techniques with no watermark embedded in the images. . On contrary, the passive schemes extract some intrinsic fingerprint traces of image/video to detect the tampered regions.

For detecting video forgery, falsified video is divided into several frames and the blind detection method based on zero-connectivity feature and fuzzy membership is proposed to detect the forgery. Since digital image forensics technology is a challenging problem and lack of any prior knowledge, it is difficult to judge the authenticity of images with a single detection algorithm, what is needed is a set of different tools that can be all applied to the image. Eventually when detecting the forgery of an image, a fused algorithm can be used to obtain a decisive answer. Therefore image forensics technology need to be continuously enriched and improved.

In contrast to above approaches, we describe a method for detecting forged video/images created by inpainting technique. Inpainting is a method to fill in undesired image portions using the information from the remaining area of the image. Region filling and object removal by inpainting technique is important both for content correction by removing unwanted objects and for image restoration by repairing damaged portions. But it is also a key tool for image manipulation, such as producing photo-montages, so it is necessary to develop a forensic algorithm to detect the inpainted forged video/images. Although this inpainted doctoring is often imperceptible, it introduces

abnormal similarity into the image block pairs, which can be used as evidence of digital tampering. In this paper we describe the form of this abnormal similarity, and how they can be automatically detected in any region of an video/image.We show the general effectiveness of proposed method.

This paper is organized as follows. Section2,we discuss existing work which is closely related to the work presented in this paper.Section 3 describes the classical image inpainting: exemplar-based inpainting technique. Section 4 describes the proposed detection method. Section 5 evaluates the performance of proposed method. In section 6 conclusions is presented.

## 2. RELAED WORK

Recently, several image forensics methods for detecting digital forgeries were proposed, in the complete absence of any form of digital watermark or signature. Fridrich established methods for detecting copy-move forgeries and double compressed JPEG images . Johnson proposed several methods for exposing digital forgeries through inconsistencies in lighting and specular highlights on the eye . Popescu developed a method for exposing digital forgeries in color filter array interpolated images . In , several statistical approaches are proposed to detect image splicing. In , binary similarity measures (BSM), image quality measures (IQM) and high order wavelet (HOW) statistics feature are extracted to detect image manipulation.

The proposed method in this paper preserves dividing falsified video into several frames and the blind detection method based on zero-connectivity feature and fuzzy membership is applied to detect the forgery.

## 3. EXEMPLAR-BASED IMAGE INPAINTING

In this section, we review the exemplar-based inpainting technique proposed by Criminisi which is well recognized for its capability of removing large objects and filling in the hole, together with analyzing the abnormal similarity caused by copy and fill operation.

Criminisi's algorithm iterates the following three steps until all pixels have been filled: (1) Computing patch priorities. Patch priorities determine the order of repair, and repair order has important implications to the ultimate outcome. (2) Propagating texture and structure information. According to texture similarity, the algorithm searches best-match patch in the source region for to-be-filled patch. (3) Filling to-be-filled patch. The value of each pixel to-be-filled is copied from its corresponding position inside best-match patch. Subsequently there have been some improvements on Criminisi's algorithm, which mostly concentrate on improving priority calculation and optimizing patch searching method.

Despite there are differences in ways of calculating priority and searching among Criminisi's algorithm and its improved algorithms, the final step of copying best-match block and replacing the block to be filled will introduce abnormal similarity between the repaired region and the rest region of the image. This abnormal similarity can be used as a basis for successful detection of inpainted forged images. The experimental results show that we can always find corresponding regions satisfying the high concentration of statistical significance in the inpainted images.

## 4. PROPOSED DETECTION METHOD

For detecting video forgery, falsified video is divided into several frames and the blind detection method based on zero-connectivity feature and fuzzy membership is proposed to detect the forgery.

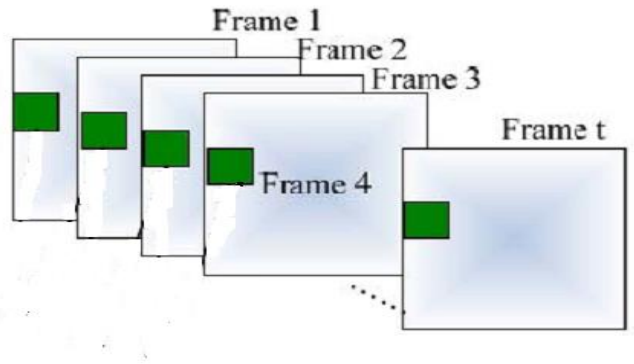



Figure. 1 A video is divided into number of frames.

For ease of understanding, we adopt notation similar to that used in the Criminisi's literature. First, given an input image I , the user selects a region of suspicious (ROS), Ω ,which may be defined as the entire image or part of it. The purpose of selecting partial region as ROS is to produce good balance between performance and complexity. Then the size of template, m*n, the parameters of membership function, a, b, and the threshold of cut set λ should be specified. Once these parameters are defined, our algorithm iterates the following three steps until all blocks in the ROS have been checked.

### 4.1. Computing block matching degrees

Due to the filling scheme of exemplar-based inpainting algorithm proposed by Criminisi, there exist a number of zeros connecting in the difference array of filling block and its best-match block. Based on this characteristic, we adopt the length of the largest zero-connectivity set in the block-pair difference array as the corresponding block matching degree. Zero connectivity should meet following conditions simultaneously: (1) the difference of the corresponding pixels in the block-pair is zero; (2) there still exists zero or zeros within its 8 neighborhoods, that is, the

location and other "zero" positions are 8-neighborhood connected.

For each block $\Psi p$ of the ROS where p is its upper left point,we search in the rest region of the image $\Phi = I - \Psi p$ for that block , which is the best similar to $\Psi p$ . Formally

$$\Psi\hat{q} = \arg\max_{\Psi q \epsilon \Phi} n(\Psi p, \Psi q) \quad (1)$$

Where $n(\Psi p, \Psi q)$ is the matching degree of block pair $(\Psi p, \Psi q)$.After having found the best-similar block $\Psi\hat{q}$, the matching degree of $\Psi p$ is the matching degree of block pair $(\Psi p, \Psi\hat{q})$:

$$\sigma\psi p = n(\Psi p, \Psi q) \quad (2)$$

The matching degree $n(\Psi p, \Psi q)$ is computed as follows: 1) two blocks are subtracted, then take their absolute values and obtain difference array Dpq, see (3); 2) in order to extract the number and location relations of "zero" appeared in Dpq, it is necessary to label all the zero-components. We use the binary image component labeling method to extract zero-connectivity features. Connectivity determines whether two binary pixels are "connected" or not . 4-connectivity and 8-connectivity are commonly used, and the pixels are marked by coordinates that are considered connected to p, as shown in Fig. 2. First Dpq should be converted into binary array by applying (4), where t denotes the threshold. Setting a suitable t can be remove noise impact while maintaining good performance. In this paper we define t=0. Having found all the connectivity components sets in , each set is marked by a unique label. Suppose the largest label is marked by "N", then only the greatest length among all labels is taken as its matching degree $n(\Psi p, \Psi q)$ , see (5).

$$D_{pq} = abs(\Psi_p - \Psi_q) \quad (3)$$

$$BW_{pq}(i,j) = \begin{cases} 1 & D_{pq}(i,j) \le t \\ 0 & D_{pq}(i,j) > t \end{cases} \quad (4)$$

$$n(\Psi_p, \Psi_q) = \begin{cases} \max\limits_{n \in \{1,\ldots N\}} length(label == n) & N \ge 1 \\ 0 & N = 0 \end{cases} \quad (5)$$

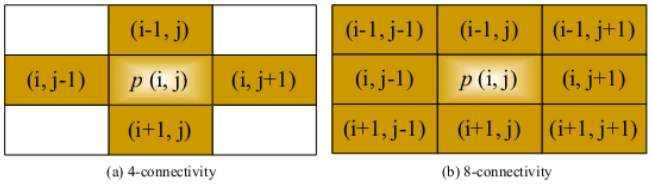



Figure 2. 4-connectivity and 8-connectivity diagrams

4.2. Computing fuzzy membership

Yet another problem in our method is how to convert the characteristics of block matching degrees into the judging factors of image doctoring. No matter how accurate our measurements (e.g. block matching degree) are some uncertainty always remains, therefore we use fuzzy set theory to describe the uncertainty in detecting tampered regions. Fuzzy set theory is the extension of conventional set theory. It handles the concept of partial truth. It was introduced by Prof. Zadeh in 1965 as a mean to model the vagueness and uncertainty in complex systems . A fuzzy set is characterized by a membership function, which maps the members of the universe into the unit interval [0, 1].

Let the universe set U denotes block matching degree, the fuzzy set $\tilde{A}$ denotes "tampered block", then $\tilde{A}$ is defined by its membership function:

$$\mu\tilde{A} : U \rightarrow [0, 1] \quad (6)$$

Because the fuzzy set described here is preferred to big fuzzy phenomenon, therefore we select an ascending semi-trapezoidal membership function. The membership function $\mu\tilde{A}$ is defined as:

$$\mu_{\tilde{A}}(\sigma_{\Psi_p}; a, b) = \begin{cases} 0 & \sigma_{\Psi_p} \le a \\ \dfrac{\sigma_{\Psi_p} - a}{b - a} & a < \sigma_{\Psi_p} \le b \\ 1 & \sigma_{\Psi_p} > b \end{cases} \quad (7)$$

4.3. Dividing the cut set

Fuzzy set is defined by the membership function, and we agree that if the membership $\mu\tilde{A}(\sigma\Psi p)$ meets or exceeds the threshold $\lambda$ , then $\sigma\Psi p$ belongs to the cut set. A cut set $A\lambda$ of a fuzzy set $\tilde{A}$ is a crisp set which defined by:

$$A\lambda = \{\sigma\Psi p \mid \mu\tilde{A}(\sigma\Psi p) \ge \lambda, \sigma\Psi p \epsilon U\},\ 0 \le \lambda \le 1 \quad (8)$$

Finally, if $A\lambda \neq \Phi$ , for all matching degrees in $A\lambda$ , the corresponding blocks that contributed to that specific matching degree are colored with the white color and thus identified as regions that might have been forged. In addition, another assumption that should be accepted is that the forged region will likely be a connected component rather than a collection of very small blocks or individual pixels. Therefore in our method, morphological operations are employed for removing isolated blocks which may be introduced by noise or compression [13]. Morphological operations used here include dilation and erosion, and they can be defined as:

$$f \oplus B = \{x \mid B(x) \cap x \neq \phi\} \quad (9)$$

$$f \Theta B = \{x \mid B(x) \subseteq f\} \quad (10)$$

Where f denotes 2-dimension image, B denotes structuring element, x denotes pixel value, $\oplus$denotes dilation operation, and $\Theta$ denotes erosion operation.

A pseudocode description of the algorithmic steps is shown in below.

- Given an input image I.
- Select ROS Ω, set parameters Ψ, a, b,λ
- For ⊎p ϵ Ω, repeat the following steps:
  1a.Compute zero-connectivity length n(Ψp , Ψq), ⊎Ψq ϵ Φ = I – Ψp
  1b.Find the patch Ψq̂ with the maximum zero-connectivity length,i.e. Ψq̂ = arg max n(Ψp , Ψq )
  ΨqϵΦ
  1c. Compute Match Degree σψp = n(Ψp , Ψq)
  2.Compute fuzzy membership μÃ(σΨp).
  3.Extract cut-set Aλ.
- If Aλ ≠ Φ, locate tampered region; else exit.

## 5. EXPERIMENTAL RESULTS

Here we apply our method to a variety of inpainted forged images, including images provided by Criminisi's paper, images created by Zhu's improved inpainting algorithm and images created by Tang's improved inpainting algorithm . In all of these experiments, all parameters are set to: m*n =9*9, a=4, b=78, λ = 0.5 .

First we test on inpainted forged images provided by Criminisi's paper. Fig. 3 is one of the detection results. Fig. 3(a) is the original natural image; Fig. 3(b) is the inpainted image created by Criminisi's algorithm; Fig. 3(c) is the detection result by running our algorithm to the selected ROS. Note that only the filling region is correctly deemed as forged because both the filling region and the corresponding best-match blocks contain high matching degrees that bring the memberships above 0.5.

Figure 3. The detection result of an inpainted forged image
provided by Criminisi's paper.

Our algorithm can also be applied to other improved inpainting approaches, such as Tang's algorithm and Zhu's algorithm. Then we test on inpainted forged images created by Tang's improved inpainting algorithm. Fig. 4(a) is the natural image; Fig. 4(b) is the inpainted image created by Tang's algorithm; Fig. 4(c) is our detection result to the entire image.

Figure 4. The detection result of an inpainted forged image created by Tang's improved inpainting algorithm.

Figure 5. The detection result of an inpainted forged image created by Zhu's improved inpainting algorithm.

Thirdly, we test on inpainted forged images created by Zhu's improved inpainting algorithm. In Fig. 5(a) the target region and searching region are selected and marked with black and red respectively, and in Fig. 5(b) the corresponding region is filled by Zhu's algorithm; In Fig. 5(c) the forged region is colored with the white color.

To analyze the difference of fuzzy membership between natural images and inpainted images, Fig. 6 compares the membership maps of natural images with those of inpainted images. The ROS is set to the entire image. In the top row of Fig. 6 are natura"man" image, inpainted "man" image, natural "woman" image and inpainted "woman" image (from left to right). In the bottom row of Fig. 6 are their membership maps respectively. These results in Fig.6 demonstrate that the proposed membership measure can successfully distinguish inpainted forged images from natural images.

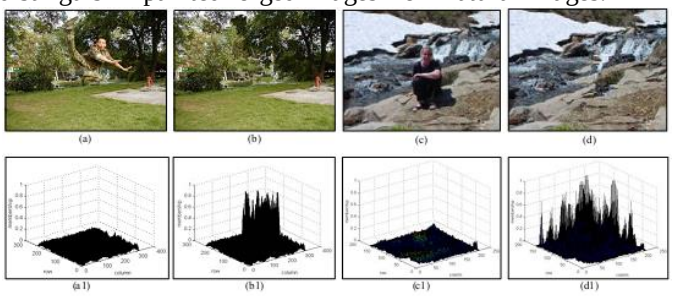
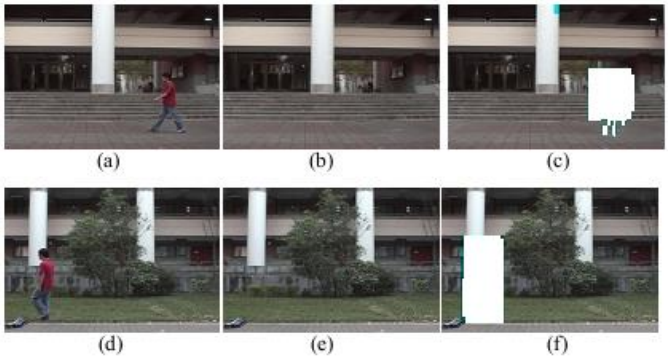


Figure 6. Comparison the membership maps of natural images with those of inpainted images.



(a) (b) (c)

Figure7.(a)(d) Source video frame. (b)(e) Correspondingfalsified result of (a) .(c)(f)Detection result of an inpainted forged video.

## 6.CONCLUSIONS

For detecting video forgery, falsified video is divided into several frames and the blind detection method based on zero-connectivity feature and fuzzy membership is proposed to detect the forgery. The exemplar-based inpainting process typically leaves behind no perceptual artifacts, it does introduce abnormal similarity between image block pairs. We have presented an efficient technique that automatically detects forged regions in an inpainted video/image. The technique works by first applying zero-connectivity labeling on block pairs to yield matching degree feature of all blocks in ROS. Fuzzy memberships are then computed by constructing ascending semi-trapezoid membership function, and tampered regions are identified by a cut set. We have shown the effectiveness of the technique on a variety of inpainted forged video/images. The major weakness of our approach is that it is currently only applicable to uncompressed images, and JPEG images with high compression rate.

Our future work will proceed to detect forged video/images created by other inpainting techniques, so that blind image forensics can be used in various forgery circumstances.

## REFERENCES

[1] J. Fridrich, D. Soukal, J. Lukáš, "Detection of copy-move forgery in digital images," In: Proc. Digital Forensic Research Workshop, Cleveland, OH, 2003.

[2] J. Lukáš, J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," In: Proc. of DFRWS, Cleveland, OH, USA, 2003.

[3] M. Johnson, H. Farid, "Exposing digital forgeries by detecting inconsistencies in lighting," In: Proc. of ACM Multimedia and Security Workshop, New York, NY, 2005.

[4] M. Johnson, H. Farid, "Exposing digital forgeries through specular highlights on the eye," In: Proc. of 9th International Workshop on Information Hiding, Saint Malo, France, 2007.

[5] A. Popescu, H. Farid, "Exposing digital forgeries in color filter array interpolated images," IEEE Transactions on Signal Processing, 2005, 53(10): 3948-3959.

[6] T.-T. Ng, S.-F. Chang, "A model for image splicing," In: IEEE International Conference on Image Processing, Singapore, 2004.

[7] Y. Q. Shi, C. H. Chen, "A natural image model approach to splicing detection," In: Proc. of MM&Sec, 2007, 51-62.

[8] S. Bayram, I. Avcibas, B. Sankur, "Image manipulation detection," Journal of Electronic Imaging, 2006, 15(4).

[9] Patrick P´erez, Michel Gangnet, Andrew Blake, "Patchworks: example-based region tiling for image editing," Microsoft Research Technical Report, MSR-TR-2004-04.

[10] A. Criminisi, P. Pérez, and K. Toyama, "Object removal by exemplar based inpainting," In: Proc. Conf. Computer Vision and Pattern Recognition, Madison, WI, June 2003.

[11] Wei Zhu, Guo-Hui Li, Dan Tu, "Application of texture synthesis in old photograph completion," Computer Engineering and Application, 2007, 28(10): 220-222.

[12] F. Tang, Y. T. Ying, et al, "A novel texture synthesis based algorithm for object removal in photographs," In: Proc. of 9th Asian Computing Science Conference, Thailang, 2004, 248-258.

[13] R. C. Gonzalez, R. E. Woods, S. L. Eddins, Digital Image Processing Using Matlab. Prentice Hall, 1st edition, 2003.

[14] G. J. Klir, U. St. Clair, B. Yuan, Fuzzy Set Theory:Foundations and Applications, Prentice Hall, 1997.

[15] F. Nielsen and R. Nock, "ClickRemoval: Interactive pinpoint image objectremoval," in Proc. 13th Annu. ACM Int. Conf. Multimedia 2005, pp. 315–318.

[16] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in Proc. 2005 IEEE Int. Conf. ImageProcess., Genova, pp. II, pp. 69–72.

[17] K. A. Patwardhan, G. Sapiro, and M. Bertalm´ıo, "Video inpainting underconstrained camera motion," IEEE Trans. Image Proc., pp. 545–553, Feb2007.

[18] S. Peleg, "Video mosaicing for non-chronological time editing," in pre-sented at the 2007 IAPR Conf. Mach. Vision Applications, Tokyo, Japan,May 16–18, 2007.

[19] L.-M. Po and W. C. Ma, "A novel four-step search algorithm for fast blockmotion estimation," IEEE Trans. Video Tech., vol. 6, no. 3, pp. 313–317,Jun. 1996.

[20] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg "Evolving timefronts: Spatio-Temporal video warping," Tech. Rep. 2005-10, The HebrewUniversity of Jerusalem.

[21] T. K. Shih, N. C. Tang, W.-S. Yeh, T.-J. Chen, and W. Lee, "Video inpaint-ing and implant via diversified temporal continuations," in Proc. 2006ACM Multimedia Conf., Santa Barbara, CA, Oct. 23–27, 2006, pp. 133–136.

[22] T. K. Shih, N. C. Tang, and J.-N. Hwang, "Ghost shadow removal in multi-layered video inpainting," in proc. IEEE 2007 Int. Conf. Multimedia Expo,Beijing, China, Jul. 2–5, pp. 1471–1474.

[23] J. Sun, L. Yuan, J. Jia, and H. Y. Shum, "Image completion with structure propagation," in Proc. ACM SIGGRAPH 2005, vol. 24, no. 3, pp. 861–868.

[24] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua, "Priors for peopletracking from small training sets," in in Proc. 10th IEEE Int. Conf. Comput.Vis., Oct. 17–21, 2005, vol. 1, pp. 403–410.

# Classification Model for Intrusion Detection

Dheeraj Tandon
Associate Professor and HOD – MCA
(Sacred Heart Institute of Management and Technology, Sitapur)
Prof. (Dr.) Harsh Dev
Dean – PSIT, Kanpur

**Abstract**

**As network-based computer systems play increasingly vital roles in modern society, they have become the targets of our enemies and criminals. Therefore, we need to find the best ways possible to protect our systems. The security of a computer system is compromised when an intrusion takes place. An intrusion can thus be defined as "*any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource*". IntroductionIntrusion prevention techniques, such as user authentication (e.g., using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weakness in the systems due to design and programming errors, or various "socially engineered" penetration techniques. For example, after it was first reported many years ago, exploitable "buffer overflow" still exists in some recent system software due to programming errors. The policies that balance convenience versus strict control of a system and information access also make it impossible for an operational system to be completely secure. Intrusion detection is therefore needed as another wall to protect computer systems. The primary assumptions of intrusion detection are: user and program activities are observable, for example, via system auditing mechanisms; and more importantly, normal and intrusion activities have distinct behavior.**

## Our Approach

We first give a brief overview of the classification problem, and explain the rationale of using classification rules for intrusion detection. We then describe in detail our experiments in constructing classification rules for anomaly detection. These experiments, motivated our research in developing CM ID. The first set of experiments is on the *sendmail* system call data, and the second is on the network *tcpdump* data. We demonstrate the effectiveness of using classification models for anomaly detection, and discuss the difficulties in building these models.

## Classification

In many applications, e.g., pattern recognition, we need to classify data items into one of a discrete set of possible categories. These *classification* tasks typically require the construction of a *classifier*, a function that assigns a *class label* to each data item described by a set of *attributes*.

More formally, let $A$ be a vector of attributes $[A0, A1, \ldots, An]$, where each attribute is either one of two types, discrete or numerical. Let $C$ be the set of class labels. Let $\alpha$ denote the universe of all possible data items in the problem domain. We write each data item $x \in \alpha$ as a vector $[v0, v1, \ldots, vn]$ such that each $vi$ is a value of $Ai$. Since each data item belongs to a category, it can be assigned a class label. That is, there is a function $f$ from $\alpha$ to $C$ such that

$$f(x) = c$$

where $c$ is a valid value of $C$. Function (or model) $f$ is normally unknown, and is the subject of inductive learning.

## Classification Rules

In many applications, it is desirable to learn a set of propositional if-then rules that jointly define the target function. Rule sets are relatively easy for humans to understand, and can easily be incorporated into (existing) general rule execution engines. One solution is to learn a model, e.g., a decision tree, and then translate the model

into an equivalent set of rules, e.g., one rule for each leaf node in the tree.

A better solution is to directly apply a rule learning algorithm. Rule learners can outperform decision tree learners and other algorithms on many problems, and can easily incorporate certain types of prior knowledge into the learning process.

The rule learning algorithm used in our study. Each rule consists of a conjunction of conditions, i.e., attribute-value tests, and a consequence, i.e., a class label. A condition is of the form *Ad = v*, *An* _ _ or *An* _ _, where *Ad* is a discrete attribute and *v* a legal value of *Ad*, or *An* is a numerical attribute and _ is some value of *An*. A data item is *covered* by a rule if its attribute values satisfy the conditions of the rule. To a rule's target class, i.e., its consequence, data items that belong to this class are called the *positive* examples, and those in other classes are called the *negative* examples.

A classification rule is learned in two phases, a *growing* phase and a *pruning* phase. The training data set *D* is also partitioned into a growing set and a pruning set for these two phases. A rule is "grown", from an empty conjunction, by repeatedly adding a condition that maximizes the *FOIL information gain* criterion until the rule covers no negative data items in the the growing data set. The FOIL information gain is measured based on the number of positive and negative examples covered before and after adding the condition to the rule. It can be interpreted as the reduction in the total number of bits required to encode the classification of all positive examples covered by the rule . After the growing phase, a rule is immediately pruned, i.e., simplified, by repeatedly deleting a condition (or a conjunction of conditions) that can lead to a more accurate rule on the pruning data set. The pruning stage is thus a means to improve both the generalization accuracy and the simplicity of the rule.

Classification rule can handle multiple classes. Given training data set *D* with *n* classes. The algorithm first orders the classes in increasing order of prevalence. That is, among the ordered *c*1, *c*2, . . . , *cn*, *c*1 is the least prevalent class and *cn* is the most prevalent. Classification first learns a rule set that separates *c*1 from the rest of the classes, removes from *D* the data items covered by the learned rules, and then learns another rule set that separates *c*2 from *c*3, *c*4, . . . , *cn*. This process continues until there is a single class *cn*, which is called the *default class*.

**Classification Rules as Intrusion Detection Models**

CM ID produces classification rules, as intrusion detection models. Since these learned rules have the standard if-then format, with minimum processing, they can be used in many rule-based IDSs in the same manner as hand-coded rules. Further, the classification rules are concise and intuitive, and can be inspected and edited by security experts when needed. That is, they can be checked for "sanity" before being employed in the actual systems.

Classification rules have two very desirable properties for intrusion detection: good generalization accuracy and concise (simplified) conditions. There will always be "new" intrusions, in the forms of slight variations of "known" intrusions or completely new breed of attacks, after the intrusion detection models are built. The ability to detect these new intrusions, that is, the generalization accuracy of the rules on the unseen data, is therefore critical for an IDS. Real-time IDSs require simple models for the sake of efficiency. Rule execution entails computing and testing the attribute values for rule condition checking. Having concise conditions therefore contributes to the real-time performance of IDSs.

**Experiments on *tcpdump* Data**

There are two approaches for network intrusion detection: one is to analyze the audit data on each host of the network and correlate the evidence from the hosts. The other is to monitor the network traffic directly using a packet capturing pro- gram such as *tcpdump* . In this section, we describe how classifiers can be induced from *tcpdump* data to distinguish network attacks from normal traffic.

**The *tcpdump* Data**

We obtained a set of *tcpdump* data, available at http://iris.cs.uml.edu:8080/ network.html, that is part of an Information Exploration Shootout (see http:// iris.cs.uml.edu:8080). The *tcpdump* was executed on the gateway that connects the enterprise LAN and the external networks. It captured the headers (not the user data) of the network packets that passed by the network interface of the gateway. Network traffic between the enterprise LAN and external networks, as well as the broadcast packets within the LAN were therefore collected. For the purposes of the shootout, filters were used so that *tcpdump* only collected Internet Transmission Control Protocol (TCP) and Internet User Datagram Protocol (UDP) packets. The data set consists of 3 runs of *tcpdump* on generated network intrusions and one *tcpdump* run on normal network traffic (with no

intrusions). The output of each *tcpdump* run is in a separate file. The traffic volume (number of network connections) of these runs are about the same. Our experiments focused on building an anomaly detection model from the normal dataset.

Since *tcpdump* output is not intended specifically for security purposes, we had to go through multiple iterations of data pre-processing to extract meaningful features and measures. We studied TCP/IP and its security related problems, for example, for guidelines on the protocols and the important features that characterize a connection.

**Data Pre-processing**

Our goal is to build a model that can predict a network connection as normal or an intrusion. *tcpdump* output contains packet data of network connections, in the order of packet appearances in the network. Packet data therefore needs to be preprocessed in order to gather connection-level information. Figure 8.1 depicts the data preprocessing process. Here, binary *tcpdump* data is first converted into ASCII packet level data, where each line contains the information of one network packet. The data is ordered by the timestamps of the packets. Therefore, packets belonging to different connections may be interleaved. For example, the 3 packets shown in the figure are from different connections. The packet data is then processed into connection records with a number of features (i.e. attributes), e.g., *time* (the starting time of the connection, i.e., the timestamp of its first packet), *dur* (the duration of the connection), *src* and *dst* (source and destination hosts), *bytes* (number of data bytes from source to destination), *srv* (the service, i.e., port, in the destination), and *flag* (how the connection conforms to the network protocols, e.g., *SF* is normal, *REJ* is "rejected"), etc. These intrinsic features essentially summarize the packet level information within a connection. We developed a script to scan each *tcpdump* ASCII packet data file and extract the "connection-level" information to form connection records. For each TCP connection, the script processes packets between the two ports of the participating hosts, and: checks whether 3-way handshake has been properly followed to establish

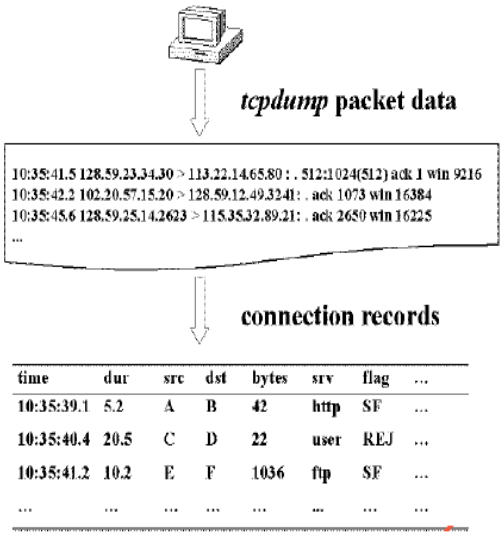



Figure : Processing Packet-level Network Audit Data into Connection Records

the connection. The following errors are recorded: connection rejected, connection attempted but not established (i.e., the initiating host never receives a SYN acknowledgment), and unwanted SYN acknowledgment received (i.e.,no connection request, a SYN packet, was sent first),

- monitors each data packet and ACK packet, keeps a number of counters in order to calculate these statistics of the connection: resent rate, wrong resent rate, duplicate ACK rate, hole rate, wrong data packet size rate, total number of data bytes sent in each direction, percentage of data packets, and percentage of control packets, and

- watches how the connection is terminated: normal (i.e., both sides properly send and receive FINs), abort (i.e., one host sends RST to terminate, and all data packets are properly ACKed), half closed (i.e., only one host sends FIN), and disconnected.

Since UDP is connectionless (i.e., no connection state), we simply treat each packet as a connection.

A connection record, in preparation of machine learning, now has the following attributes (features): start time, duration, participating hosts, ports, the statistics of

the connection (e.g., bytes sent in each direction, resent rate, etc.), flag (i.e., "normal" or one of the recorded connection/termination errors), and protocol type (i.e., TCP or UDP). From the ports, we know whether the connection is to a well-known service, e.g., *http* (port 80), or to a user application.

We call the host that initiates the connection, i.e., the one that sends the first SYN, as the source, and the other as the destination. Depending on the direction from the source to the destination, a connection is one of the three types: *out-going* - from the LAN to the external networks; *in-coming* - from the external networks to the LAN; and *inter-LAN* - within the LAN. Taking the topology of the network into consideration is important in network intrusion detection. Intuitively, intrusions, which come from outside, may first exhibit some abnormal patterns (e.g., penetration attempts) in the *in-coming* connections, and subsequently in the *inter-LAN* (e.g., doing damage to the LAN) and/or the *out-going* (e.g., stealing/uploading data) connections. Analyzing these types of connections and constructing corresponding detection models separately may improve detection accuracy.

**Experiments and Results**

For each type (i.e., direction) of the connections, we formulated the classification experiments as the following:

- Each connection record uses the destination service (i.e., destination port) as the class label, and all the other connection features as attributes;
- The training data is 80% of the connections from the normal *tcpdump* data file, while the test data includes the remaining 20% from the normal *tcpdump* data file, and all the connections from the 3 *tcpdump* data files marked as having embedded attacks;
- 5-fold cross validation evaluation is reported here. The process of training and testing is repeated 5 times, each time using a different 80% of the normal data as the training data, and accordingly the different remaining 20% of the normal data as part of the test data. The averaged accuracy of the classifiers from the 5 runs is reported.

We applied classification to the connection data. The resulting classifier characterizes the normal patterns of each service in terms of the connection features. When using the classifier on the testing data, the percentage of misclassifications on each *tcpdump* data set is reported. Here a misclassification is the situation where the classifier predicts a destination service, according to the connection features, that is different from the actual. This misclassification rate should be very low for normal connection data and high for intrusion data. The intuition behind this classification model is straightforward: when intrusions take place, the features (i.e.

| | % misclassification (by traffic type) | | |
|---|---|---|---|
| Connection Data | out-going | in-coming | inter-LAN |
| normal | 3.91% | 4.68% | 4% |
| intrusion1 | 3.81% | 6.76% | 22.65% |
| intrusion2 | 4.76% | 7.47% | 8.7% |
| intrusion3 | 3.71% | 13.7% | 7.86% |

Table : Misclassification Rate on Normal and Intrusion Data. Separate classifiers were trained and tested on connection data of each traffic type. "normal" is the 20% data set aside from the training data. No intrusion data was used for training. characteristics) of connections to certain services, for example, *ftp*, are different from the normal traffic patterns of the same service.

The results from the first round of experiments, as shown in Table 8.1, were not very good: the differences in the misclassification rates of the normal and intrusion data were small, except for the *inter-LAN* traffic of some intrusions.

We then redesigned our set of features by adding some continuous and intensity measures into each connection record:

- Examining all connections in the past *n* seconds, and counting the number of: connection establishment errors (e.g., "connection rejected"), all other types of errors (e.g., "disconnected"), connections to designated system services (e.g., *ftp*), connections to user applications, and connections to the same service as the current connection;
- Calculate for the past *n* seconds, the per-connection average duration and data bytes (on both directions) of all connections, and the same averages of connections to the same service.

These temporal and statistical features provide additional information of the network activity from a continuous perspective, and provide more insight into
anomalies. For example, a low rate of error due to innocent attempts and network glitches in a short time span is expected, but an excess beyond the averaged norm indicates anomalous activity. Table 8.2 shows the improvement of adding these features. Here, using a time interval of 30 seconds (i.e., $n$ = 30*s*), we see that the misclassification rates on the intrusion data are much higher than the normal data, especially for the *in-coming* traffic. The classification rule set (i.e., the classifier) has just 9 rules and 25 conditions. For example, one rule says "if the average number of bytes from source to destination of the connections to the same service is 0, and the percentage of control packets in the current connection is 100%, then the service is *auth*".

To understand the effects of the time intervals on the misclassification rates, we ran the experiments using various time intervals: 5s, 10s, 30s, 60s, and 90s. The effects on the *out-going* and *inter-LAN* traffic were very small. However, as Figure 8.2 shows, for the *in-coming* traffic, the misclassification rates on the intrusion data increase dramatically as the time interval goes from 5s to 30s, then stabilizes or tapers off afterwards.

| | % misclassification (by traffic type) | | |
|---|---|---|---|
| Connection Data | out-going | in-coming | inter-LAN |
| normal | 0.88% | 0.31% | 1.43% |
| intrusion1 | 3.04% | 27.42% | 5.63% |
| intrusion2 | 2.54% | 27.37% | 20.48% |
| intrusion3 | 2.32% | 42.20% | 6.80% |

**Conclusions**

This Paper described a framework, CM ID, for Classification Models for Intrusion Detection. CM ID consists of classification and meta-classification programs. Using CM ID, frequent patterns computed from large amount of system audit data are used to construct predictive features, from which classification rules are inductively learned to detect intrusions.

**Closing Remarks**

This paper documented our research in developing and applying data mining techniques to the important and challenging problem of building intrusion detection models. A set of algorithms and tools, collectively called CM ID, have been designed, implemented, and evaluated for this thesis research. While CM ID has shown great promises, there are open problems for future research.

**References**

R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 207–216,1993.

P. K. Chan and S. J. Stolfo. Toward parallel and distributed learning by meta-learning. In *AAAI Workshop in Knowledge Discovery in Databases*, pages 227–240, 1993.

W. W. Cohen. Fast effective rule induction. In *Machine Learning: the 12th International Conference*,Lake Taho, CA, 1995. Morgan Kaufmann.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process of extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

K. Ilgun, R. A. Kemmerer, and P. A. Porras. State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering*, 21(3):181–199, March 1995.

V. Jacobson, C. Leres, and S. McCanne. tcpdump. available via anonymous ftp to ftp.ee.lbl.gov, June 1989.

C. Ko, G. Fink, and K. Levitt. Automated detection of vulnerabilities in privileged programs by execution monitoring. In *Proceedings of the 10th Annual Computer Security Applications Conference*, pages 134–144, December 1994.

# COURSES OFFERED

| Graduate Courses | Post Graduate Courses | Professional Courses |
|---|---|---|
| B.A. (Eco) | M.Com | B.B.A |
| B.Sc. (Comp. Sc) | M.Sc. (Comp. Sc.) | B.C.A |
| B.Com | MCA (Commerce) | B.B.M |

SAP | ACCA | CFA | CIMA | TALLY