IITM Journal of Management and IT

SOUVENIR

National Conference on Emerging Trends in Information Technology-Advances in High Performance Computing, Data Sciences & Cyber Security

Volume 8 **Issue 1** **January-June, 2017**

CONTENTS

Research Papers & Articles

Intelligent Cyber Security Solutions through High Performance Computing and Data Sciences : An Integrated Approach

Sandhya Maitra*
Dr. Sushila Madan**

Abstract

The recent advances in Data Sciences and HPC despite transforming the ongoing digitization to have a positive impact on the social and economic aspect of our lives, have at the same time, given birth to several security issues. Thus the face of Cyber security has changed in the recent times with the advent of new technologies such as the Cloud, the internet of things, mobile/wireless and wearable technology. The technological advances in data science which help develop contemporary cyber security solutions are storage, computing and behavior. On the other hand high performance computing power facilitates the usage of sophisticated machine learning techniques to build innovative models for identification of malware. Big data holds vital importance in building analytical models which identify cyber attacks. Besides High performance computing is necessary for supporting all aspects of data-driven research. An integrated approach combining the technological benefits provided by predictive power of data sciences and the aggregated parallel processing power of high performance computing would help devise intelligent and powerful cyber security solutions supporting proactive and dynamic approach to threat management to counteract the multitude of potentially new emerging cyber attacks.

Keywords: High Performance computing, Data Sciences, Machine Learning, Cyber Security

I. Introduction

The researchers all over the world face challenges related to upsurge of voluminous data of many areas such as Bioinformatics, Medicine, Engineering & Technology, GIS and Remote Sensing, Cognitive science and Statistical data. Advanced algorithms, visualization techniques, data streaming methodologies and analytics are the need of the hour. These have to be developed within the constraints of storage and computational power, algorithm design, visualization, scalability, distributed data architectures, data dimension reduction and implementation to name a few. The other issues to be considered include optimization, uncertainty quantification, systems theory, statistics and types of model development methods. This requires contextual problem solving based on multidisciplinary approaches. The scale, diversity, and complexity of Big Data necessitates the advent of new architecture, techniques, algorithms, and analytics to manage it and extract value or hidden knowledge from it. Analytics research encompasses a large range of problems of data mining research[1].

Data is increasingly becoming cheap and ubiquitous. The rapid growth in computer science and information technology in the recent times has led to the generation of massive amount of data. This avalanche of data has made a strong impact on almost all aspects of human life and fundamentally changed every field in science and technology. A multitude of new types of data is collected from web logs, sensors, mobile devices, transactions and various instruments. The emerging technologies such as data mining and machine learning enable us to interpret this massive data. The High Performance Computing (HPC) techniques are increasingly being used by organizations to efficiently and effectively deal with processing and storage challenges thrown by explosive growth of such

Sandhya Maitra*
Research Scholar
Banasthali Vidyapith
Dr. Sushila Madan**
Professor
Lady Shri Ram College for Women

enormous data. Advances in Networking, High End Computers, Distributed and Grid computing, Large-scale visualization and data management, Systems reliability, High-performance software tools and techniques, and compilation techniques are taking a new era of high performance, parallel and distributed computing. Over the past few decades security concerns are becoming increasingly important and extremely critical in the realm of communication and information systems as they become more indispensable to the society. With the continuous growth of cyber connectivity and the ever increasing number of applications, remotely delivered services, and networked systems digital security has become the need of the hour. Today government agencies, financial institutions, and business enterprises are experiencing security incidents and cyber-crimes, by which attackers could generate fraudulent financial transactions, commit crimes, perform an industrial espionage, and disrupt the business processes. The sophistication and the borderless nature of the intrusion techniques used during a cyber security incident, have generated the need for designing new active cyber defense solutions, and developing efficient incident response plans. With the number of cyber threats escalating worldwide, there is a need for comprehensive security analysis, assessment and actions to protect our critical infrastructures and sensitive information[1].

II. Cyber Security

The spectacular growth of cyber connectivity and the monumental increase of number of networked systems, applications and remotely delivered services cyber security has taken top precedence amongst other issues. Attackers are able to effect fraudulent financial transactions, perform industrial espionage, disrupt business processes and commit crimes with much ease. Additionally government agencies are also experiencing security incidents and cyber-crimes of dangerous proportions which can compromise on Nations Security. The sophisticated intrusion techniques used in the cyber security incidents and their borderless nature have provided the impetus to design new active cyber defense solutions, and develop efficient and novel incident response plans. The number of cyber threats are escalating globally, necessitating comprehensive security analysis, assessment and action plans for protecting our critical infrastructures and sensitive information[1].

Cyber security in recent times demand secure systems which help in detection of intrusions, identification of attacks, confinement of sensitive information to security zones, data encryption, time stamping and validation of data and documents, protection of intellectual property, besides others. The current security solutions require a mix of software and hardware to augment the power of security algorithms, real time analysis of voluminous data, rapid encryption and decryption of data, identification of abnormal patterns, checking identities, simulation of attacks, validation of software security proof, patrol systems, analysing video material and many more innumerable actions [2].

Analysis of new and diverse digital data streams can reveal potentially new sources of economic value, fresh insights into customer behavior and market trends. But this influx of new data creates challenges for IT Industry. We need to have Information Security measures to ensure a safe, secure and reliable cyber network, for the transmission and flow of information[1].

III. High Performance computing

The re-emergence of need for supercomputers for cyber security stems from their computing capacity ability to perform large number of checks in an extremely short time particularly in the case of financial transactions for the identification of cyber crimes using techniques featuring cross-analysis of data coming from several different sources[2]. The knowledge gained through HPC analysis and evaluation can be instrumental providing comprehensive cyber security as it helps interpret the multifaceted complexities involved in cyber space comprising complex technical, organizational and human systems[3].

A combined system of Distributed sensor networks and HPC cybersecurity systems such as exascale computing helps in real-time fast I/O HPC accelerated processing. This covers various issues such as data collection, analysis and response to takes care of the

issues of data locality, transport, throughput, latency, processing time and return of information to defenders and defense devices.

An important set of HPC jobs has involved analytics, discovering patterns in the data itself as in cryptography. The data explosion fueling the growth of high performance data analysis originates from the following factors:

1. The efficiency of HPC systems to run data-intensive modeling.
2. Advent of larger, more complex scientific instruments and sensor networks such as "smart" power grids.
3. Growth of stochastic modeling (financial services), parametric modeling (manufacturing) and iterative problem-solving methods, whose cumulative results are large volumes of data.
4. Availability of newer advanced analytics methods and tools: MapReduce/Hadoop, graph analytics, semantic analysis, knowledge discovery algorithms and others the escalating need to perform advanced analytics by commercial applications in near-real-time such as cloud.

Data-driven research necessitates High performance computing. Big Data fuels the growth of HP data analysis[3]. Research on High Performance Computing includes mainly networks, parallel and high performance algorithms, programming paradigms and run-time systems for data science apart from other areas. High-performance computing (HPC) refers to systems that can rapidly solve difficult computational problems across a diverse range of scientific, engineering, and business fields by virtue of their processing capability and storage capacity. HPC being at the forefront of scientific discovery and commercial innovation, holds leading competitive edge for nations and their enterprises[4]. India in an endeavour to meet its stated research and education goals is making every effort towards doubling up its high performance computing capacity and is exploring opportunities to integrate with global research and education networks.

Cyber Security and Data Sciences

The challenge of protecting sensitive data increased exponentially in recent times because of the non existence of a secure perimeter as before where it was confined to secure data centers as data leaks out of massive data centers into cloud, mobile devices and individual PCS . Most companies do not have policies prohibiting storage of data in mobiles while people on the other hand prefer storing them on to their mobiles with huge computing and storage power for convenience and efficiency of operations.

Cloud-based data mostly exists in commercial data centers, on shared networks, on multiple disk devices in the data center, and multiple data centers for the purpose of replication. The extremely difficult task of developing Cloud security is now made possible with new technologies such as HPC and machine learning.

Data from data centers should be moved to cloud only for business reasons with benefits outweighing the costs of providing cloud security to protect it. Data Inventories should be maintained in encrypted form, tracked and managed well on mobile devices to prevent theft of data. Additionally Cloud networks should be subjected to thorough penetration testing[5].

The value of cyber security data plays a major role in constructing machine learning models. Value of a data is the predictive power of a given data model as well as the type of hidden trends which reveal as a result of meticulous data analysis. The value of cyber security data refers to the nature of data which can be positive or negative. Positive data such as malicious network traffic data either from malware or varied set of cyber attacks hold higher value than data science problems as it can be used to build machine learning based network security models. From cyber security view point the predictive power of effective data models lies in the ability to differentiate normal network traffic from abnormal malicious traffic indicating active cyber attack. Machine learning builds classifiers to identify network traffic as good or bad based on the analysis. The spam filters are based on these techniques to identify normal emails from ad's, phishing and other types of spam. Big Data helps build Classifiers to train a machine learning algorithm and also helps evaluate the classifiers performance. The positive data that a spam classifier needs to detect is behavior exhibited

by a spam email. Similarly the network traffic exhibiting behavior of real cyber attacks is positive data for a network security model. Negative data refers to normal data such as legitimate emails in case of spam classifier and normal traffic data for a network security model. In both the cases the classifier should be able to detect bad behavior without incorrectly classifying genuine mails or network traffic to be harmful. The various cyber security problems differ on the basis of quick availability of positive data. In the case of spam emails positive data is easily available in abundance for building a classifier. On the other hand despite increased cyber attacks across various organizations positive data from real cyber attacks and malware infections can seldom be accessed. This is true for especially targeted attacks. The pace at which the hackers modify their techniques to create increasingly sophisticated attacks render libraries of malware samples quickly obsolete. In case of targeted attacks malware is custom built to steal or destroy data in a secret manner. The predictive power of a machine learning model relies on the high value of positive samples in terms of its general nature for identifying potentially new cyber attacks. Additionally performance on these models is highly influenced by the choice of features used to build them. The prerequisites for interpreting huge amount of positive samples are feature selection and appropriate training techniques. The highly unbalanced nature of training data for a machine learning model is owing to negative samples always being many orders of magnitude more abundant than positive data samples. The application of proper evaluation metrics, sophisticated sampling methods and proper training data set balancing helps us find out if we have the appropriate quantity of positive samples or not. The lengthy process of collecting positive samples is one of the first and most important tasks for building machine learning based cyber security models. This is how big data is relevant to cyber security[6].

Intelligent Cyber Security Solutions Powered by HPC and Data Sciences

The advances in Data Sciences and HPC have extended innumerable benefits and conveniences to our day to day activities and transformed the ongoing digitization to deeply impact the social and economic aspects of our lives. At the same time these dependencies have also given rise to many security issues. The attackers in the cyber world are also getting more creative and ambitious in exploitation of techniques and causing real-world damages of major dimensions by making even proprietary as well as personally identifiable information equally vulnerable. The problem is further compounded as designing effective security measures in a globally expanding digital world is a demanding task. The issues to be addressed include defining the core elements of the cyber security, Virtual private network security solutions, Security of wireless devices, protocols and networks, Security of key internet protocols, protection of information infrastructure and database security. The advent of the Internet of Things (IoT) also increased the need to step up cyber security. The Io T is a network of physical objects with embedded technology to communicate, sense or interact with their internal states or the external environment where a digitally represented object becomes something greater than the object by itself or possesses ambient intelligence. Despite its manifold advantages the rapid adoption of IoT by various types of organizations escalated the importance of security and vulnerability. The computing world underwent a major transformation in terms of increased reliability, scalability, quality of services and economy with emergence of cloud computing. Nevertheless, remote storage of data in cloud away from owner can lead to loss of control of data. The success and wide spread usage of cloud computing in future depends on effective handling of data security issues such as accountability, data provenance and identity and risk management. The face of Cyber security has changed in the recent times with the advent of new technologies such as the Cloud, the internet of things, mobile/ wireless and wearable technology[1].

The static data once contained within systems have now become dynamic and travel through a number of routers, hosts and data centers. The hackers in cyber criminals have started using Man-in-the-Middle attacks to eavesdrop on entire data conversations Spying software and Google Glass to track fingerprint movements on touch screens, Memory-scraping malware on point-of-sale systems, theft of specific data by Bespoke attacks.

Context-aware behavioral analytics treats unusual behavior as a symptom of an ongoing nefarious activity in the computer system.

These cases can no longer be handled by tool based approaches fire walls or antivirus machines. The previous solutions no more succeed in managing risk in recent technologies, there is an imperative need for brand new solutions. Analytics help in identifying unusual or abnormal behaviors. Behavior based analytics approaches include Bio Printing, mobile location tracking, behavioral profiles, third party Big Data and external threat intelligence. Now a days hackers carefully analyze a system defenses and use Trojan horses and due to the velocity volume and variety of big data security breaches cannot be identified well in time. Solutions based on new technologies combining machine learning and behavioral analytics help detect breaches and trace the source. User profiling is built and machine behavior pattern studied to detect new type of cyber attacks, the emphasis is on providing rich user interfaces which help in interactive exploration and investigation. These tools can detect strange behavior and changes in data.

This problem can be solved by Virtual dispersive technologies which split the message into several encrypted parts and routed on different independent servers, computers and/or mobile phones depending on the protocol.

This problem can be solved by Virtual dispersive technologies which split the message into several encrypted parts and routed on different independent servers, computers and/or mobile phones depending on the protocol.

The traditional bottlenecks are thus completely avoided. The data dynamically travels on optimum random paths also taking into consideration network congestion and other issues as well. Hackers find it difficult to find data parts. Furthermore in order to prevent cyber criminals exploiting the weak point of the technology which is the place where two endpoints must connect to a switch to enable secure communication, hidden switches are used by VDN making them hard to find.

Critical infrastructures can be protected by security measures and standards provided by Smart Grid technologies. The cloud based applications which are beyond the realm of firewalls and traditional security measures can be secured by using a combination of encryption and intrusion detection technologies to gain control of corporate traffic. Cloud data can be protected by Security assertion Markup language, an XML based open standard format, augmented with encryption and intrusion detection technologies. This also helps control corporate traffic.

Proxy based systems designed through SAML secure access and traffic, log activity, watermark files by embedding security tags into documents and other files for tracking their movement and redirect traffic through service providers. Such solutions neither require software to load on endpoints nor changes to end user configurations. Any kind of suspicious activity such as failed or unexpected logins etc are alerted by notifications. The security administrators can instantaneously erase corporate information without effecting personal data of users. Active defense measures such as counter intelligence gathering, sink holing, honey pots and retaliatory hacking can be adopted to track and attack hackers. Counter intelligence gathering is a kind of reverse malware analysis in which a cyber expert secretly finds information about hackers and their techniques. Sink holing servers hand out non routable addresses for all domains within sink hole. Malicious traffic is intercepted and blocked for later analysis by experts. Isolated systems called Honey pots such as computer, data or network sites are set up to attract hackers. Cyber security analysts to catch spammers to prevent attacks etc.. Retaliatory hacking is most dangerous security measure which usually considered illegal as it may require infiltration into a hacker community, build a hacking reputation to prove the hacking group of your credentials. None of these things being legal raises debate over active defense measures. Early warning systems forecast sites and server likely to be hacked using machine learning algorithms. These systems are created with the help of machine learning and data mining techniques. Most of the algorithms take into the account a website software, traffic statistic, file system structure or webpage structure. It uses a variety of other signature features to determine the presence of known hacked and malicious websites.

Notifications can be sent to website operators and search engines to exclude the results. Classifiers should be designed to adapt to emerging threats. Such security measure is growing in its scope. The more data that absorbs the better will be its accuracy[7].

The cyber threats in recent times necessitate state of the art dynamic approach to threat management. The Cyber security threats rapidly changing with technological advancements. An application vulnerability free today may be exposed to a major unanticipated attack tomorrow. A few of recent examples are of Adobe Flash vulnerability allowing remote code execution, NTP (Network Time Protocol) issue allowing denial-of-service attacks, Cisco ASA firewall exposure allowing for denial-of-service attacks, and Apple, thought for a long time to be invulnerable, releasing iOS 9, quickly followed by additional releases to correct newly discovered exposures. The dynamic threats are the key challenges to information security and necessitate dynamic security approaches for their mitigation. Neither were these a resultant of negligence on the part of affected parties nor was it the result of a change affected by these parties in the products. The information security programs should be proactive, agile and adaptive. A few of the strategies for moving from static to a dynamic is by making vulnerability checks a regular and frequent task with monthly external scans and internal scans conducted on same schedule or when software or configuration changes are made, whichever happens first, paying attention to fundamentals such as checking logs and auditing access rights. Firmware updates should be top priority as many of the exposures we face today result from issues found in the firmware of devices attached to our network score devices such as routers and firewalls, or Internet of Things devices, such as printers and copiers. Threat sources should be studied on a regular basis[8].

Data science techniques help in the prediction of types of security threats decides reacting to these threats. Data sciences and cyber security were highly isolated disciplines until recent times. The cyber security solutions are usually based on signatures which use pattern matching with prior identified malware to capture cyber attacks. But these signature based solutions could not prevent zero day attacks for unidentified malware as they lack predictive power of data science. Data science effectively uses scientific techniques to draw knowledge from data. The ongoing security breaches accentuate the need for new approaches for identification and prevention of malware. The technological advances in data science which help develop contemporary cyber security solutions are storage, computing and behavior. The storage aspect eases the process of collection and storage of huge data on which analytic techniques are applicable. On the other hand high performance computing power assists machine learning techniques to build novel models for identification of malware. The behavioral aspect had shifted from identification of malware with signatures to identify the specific kind of behaviors exhibited by an infected computer. Big data plays a key role analytical models which identify cyber attacks. Any rule based model based on machine learning requires large number of data samples to be analyzed in order to unearth the set of characteristics of a model. Subsequently data is required to cross check and assess the performance of a model.

Application of machine learning tools to enterprise security gives rise to a new set of solutions. These tools can analyze networks, learn about them, detect anomalies and protect enterprises from threats[9].

Machine learning increased in its popularity with the advent of high performance computing resources. This has resulted in the development of off-the-shelf machine learning packages which allow complex machine learning algorithms to be trained and tested on huge data samples. The aforementioned characteristics render machine learning as an indispensable tool for developing cyber security solutions. Machine learning is a broader data science solution for detecting cyber attacks. Minor changes in malware can leave Intrusion Prevention Systems and Next-generation Fire wall perimeter security solutions performing signature matching in network traffic ineffective. The rigorous analytical methods of data sciences differentiate abnormal behavior defining an infected machine after identifying normal behavior through repetitive usage. Therefore contemporary cyber security solutions require big data samples and

advanced analytical methods to build data-driven solutions for malware identification and detection of cyber attacks. This results in spectacular improvement of cyber security efficacy[10].

Conclusions

- Cyber Security Solutions should be more proactive and dynamic.
- Effective Cyber Security Solutions for future threats can be achieved by exploiting the processing and storage power of High Performance Computing.
- Intelligent Cyber Security Solutions can be built by exploring the predictive power of machine learning and data mining approaches.
- Machine learning approaches require Big Data for training models.
- Big Data can be efficiently processed in real time using High Performance Computing.
- Cloud Computing, IoT can be highly risk prone in the absence of effective security framework.
- The Solution to Future security needs lies in integrating the processing and storage power of High Performance Computing with predictive power of machine learning and data mining techniques.

References

1. S. Maitra, "NCETIT'2017", *iitmipu.ac.in*, 2017. [Online]. Available: http://iitmipu.ac.in/wp-content/uploads/2017/02/NCETIT-2017-Brochure.pdf. [Accessed: 14- Feb- 2017].
2. "HPC solutions for cyber security", *Eurotech.com*, 2017. [Online]. Available: https://www.eurotech.com/en/hpc/industry+solutions/cyber+security. [Accessed: 11- Feb- 2017].
3. C. Keliiaa and J. Hamlet, "National Cyber Defense High Performance Computing and Analysis: Concepts, Planning and Roadmap", Sandia National Laboratories, New Mexico, 2010.
4. S. Tracy, "Big Data Meets HPC", *Scientific Computing*, 2014. [Online]. Available: http://www.scientificcomputing.com/article/2014/03/big-data-meets-hpc. [Accessed: 11- Feb- 2017].
5. R. Covington, "Risk Awareness:The risk of data theft — here, there and everywhere", *IDG Contributor Network*, 2016.
6. D. Pegna, "Cybersecurity, data science and machine learning: Is all data equal?", *Cybersecurity and Data Science*, 2015.
7. "Hot-technologies-cyber-security", *cyberdegrees*, 2017. [Online]. Available: http://www.cyberdegrees.org/resources/hot-technologies-cyber-security/. [Accessed: 04-Feb- 2017].
8. R. Covington, "Risk Awareness:Is your information security program giving you static?", *: IDG Contributor Network*, 2015.
9. B. Violino, "Machine learning offers hope against cyber attacks", *Network World*, 2016.
10. D. Pegna, "Cybersecurity and Data Science:Creating cybersecurity that thinks", *IDG Contributor Network*, 2015.

Applications of Machine Learning and Data Mining for Cyber Security

Ruby Dahiya*
Anamika**

Abstract

Security is an essential objective in any digital communication. Nowadays, there is enormous information, lots of protocols, too many layers and applications, and massive use of these applications for various tasks. With this wealth of information, there is also too little information about what is important for detecting attacks. Methods of machine learning and data mining can help to build better detectors from massive amounts of complex data. Such methods can also help to discover the information required to build more secure systems, free of attacks. This paper will highlight the applications of machine learning and data mining techniques for securing data in huge network of computers. This paper will also present the review of applications of data mining and machine learning in the field of computer security. The papers which will be reviewed here, present the results of various techniques of data mining and machine learning on different performance parameters.

Keywords: Data mining, Machine Learning, Artificial Neural Networks, Classification, Clustering, Inductive Learning, Evolution Learning, Support Vector Machine.

I. Introduction

As technology moves forward user become more technical aware then before. People communicate and corporate efficiently through the internet using their PC's, PDs or mobile phones. Through these digital devices link by the internet, hacker also attack personal privacy using a variety of weapons such as virus, worms, botnet attacks, spam and social engineering platforms. These forms of attack can be categorized into three groups- Stilling confidential information, manipulating the components of cyber infrastructures and denying the functions of infrastructure. There are three approaches to deal with these attacks: signature-based, anomaly-based and hybrid. The signature based detection system use the particular signature of an attack, hence are unable to detect unknown attacks. The anomaly-based system detects the anomalies as the deviation from the normal behavior so they can detect unknown attacks as well. The main disadvantage of these systems is high false alarm rates (FAR). The hybrid approach uses the combination of both signature-based and anomaly-based techniques. These types of system have high detection rate of known attacks and low false positive rates for unknown attacks. The literature review shows that most of the techniques were actually hybrid. The security mechanisms are also categorized as: network based and host based. A network-based system monitors the traffic through the network devices. A host based system monitors the processes and the file related activities associated with a specific host. However building a defense system for discovered attacks is not easy because of constantly evolving cyber attacks. The figure 1 depicts the cyber security mechanism.

This paper is intended for readers who wish to begin research in the field of machine learning and data mining for cyber security. This paper highlights ML and DM techniques used for cyber security. The paper describes ML and DM techniques in reference to anomaly method and signature based hybrid methods however the in depth description of these methods is in the paper of Bhuyan et al. [1]. This paper focuses on cyber intrusion detection for both wired and wireless networks. The paper Zhang el al. [2] focuses more on dynamic networking.

Ruby Dahiya*
Associate Professor (IT)
Institute of Information Technology & Management

Anamika**
Assistant Professor (IT)
Institute of Information Technology & Management

Figure1. Cyber Security System

The paper is organized as follow: section II highlights the procedure of Machine Learning and Data Mining. Section III describes the techniques of ML and DM. Section IV presents and discusses the comparative analysis of individual technique and related work. Section V presents the conclusion.

II. Machine Learning and Data mining Procedure

The ML and DM are two terms that are often confused because generally, they both have same techniques. Machine Learning, a branch of artificial intelligence, was originally employed to develop hniques to enable computers to learn. Arthur Samuel in 1959 defined Machine Learning as a "field of study that gives computers the ability to learn without being explicitly programmed"[3]. ML algorithm applies classification followed by prediction, based on known properties learned from the training data. ML algorithms need a well defined problem from the domain where as DM focuses on the unknown properties in the data discovered priory. DM focuses on finding new and interesting knowledge. An ML approach consists of two phases: training and testing. These phases include classification of training data, feature selection, training of the model and use of model for testing unknown data.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The following are areas in which data mining technology may be applied or further developed for intrusion detection

- Development of data mining algorithms for intrusion detection: Data mining algorithms can be used for misuse detection and anomaly detection. The techniques must be efficient and scalable, and capable of handling network data of high volume, dimensionality and heterogeneity.
- Association and correlation analysis and aggregation to help select and build discriminating attributes: Association and correlation mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection.
- Analysis of stream data: Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment. It is necessary to study what sequences of events are frequently encountered together, finding sequential patterns, and identify outliers.
- Distributed data mining: Intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to analyze network data from several network locations in order to detect these distributed attacks.

- Visualization and querying tools: Visualization tools should be available for viewing any anomalous patterns detected. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results.

III. Techniques of ML and DM

This section focuses on the various ML/DM techniques for cyber security. Here, each technique is elaborated with references to the seminal work. Few papers of each technique related to their applications to cyber security.

*A. Artificial Neural Networks:*Neural Networks follow predictive model which are based on biological modeling capability and predicts data by a learning process. The Artificial Neural Networks (ANN) is composed of connected artificial neurons capable of certain computations on their inputs [4]. When ANN is used as classifiers, the each layer passes its output as an input to the next layer and the output of the last layer generates the final classification category.

ANN are widely accepted classifiers that are based on perceptron [5] but suffer from local minima and lengthy learning process. This technique of ANN is used for as multi-category classifier for signature-based detection by Cannady [6]. He detected 3000 simulated attacks from a dataset of events. The findings of the paper reported almost 93% accuracy and error rate 0.070 root mean square. This technique is also used by Lippmann and Cunningham [27] for anomaly detection. They used keyword selection based on statistics and fed it to ANN which provides posterior probability of attack as output. This approach showed 80% detection rate and hardly one false alarm per day. Also, a five-stage approach for intrusion detection was proposed by Biven et al. [8] that fully detected the normal behavior but FAR is 76% only for some attacks.

B. Association Rules and Fuzzy Association Rules: Association Rule Mining was introduced by Agarwal et.al. [9], as a way to find interesting co-occurrences in super market data to find frequent set of items which bought together. The traditional association rule mining works only on binary data i.e. an item was either present in the transaction will be represented by 1 or 0 if not. But, in the real world applications, data are either quantitative or categorical for which Boolean rules are unsatisfactory. To overcome this limitation, Fuzzy Association Rule Mining was introduced [10], which can process numerical and categorical variables.

An algorithm based on Signature Apriori method was proposed by Zhengbing et al. [11] that can be applied to any signature based systems for the inclusion of new signatures. The work of Brahmi [12] using multidimensional Association rule mining is also very promising for creating signatures for the attacks. It showed the detection rate of attacks types DOS, Probe, U2R and R2L as 99%, 95%, 75% and 87% respectively. Association rule mining is used in NETMINE [35] for anomaly detection. It applied generalization association rule extraction based on Genio algorithm for the identification of recurring items. The fuzzy association rule mining is used by Tajbakhsh et al. [38] to find the related patterns in KDD 1999 dataset. The result showed good performance with 100 percent accuracy and false positive rate of 13%. But, the accuracy falls drastically with fall of FPR.

C. Bayesian Networks: A Bayesian is a graphical model based on probabilities which represents the variables and their relationships [15], [16]. The network is designed with nodes as the continuous or discrete variables and the relationship between them is represented by the edges, establishing a directed acyclic graph. Each node holds the states of the random variable and the conditional probability form.

Livadas et al. [17] presented comparative results of various approaches to DOS attack. The anomaly detection approach is mainly reactive whereas signature-based is proactive. They tried to detect botnets in Internet Relay Chat (IRC) traffic data. The analysis reported the performance of Bayesian networks as 93% precision and very low FP rate of 1.39%.Another IDS based on Bayesian networks classifiers was proposed by Jemili et al. [18] with performances of 89%, 99%, 21% and 7% for DOS, Probe, U2R and R2L respectively. Benferhat [19] also used this approach to build IDS for DOS attack.

D. Clustering: Clustering is unsupervised technique to find patterns in high-dimensional unlabeled data. It is used to group data items into clusters based on a similarity measure which are not predefined.

This technique was applied by Blowers and Williams [20] to detect anomaly in KDD dataset at packet level. They used DBSCAN clustering technique. The study highlighted various machine learning techniques for cyber security. Sequeira and Zaki [21] performed detection over shell commands data to identify whether the user is a legitimate one or intruder. Out of various approaches for sequence matching, the longest common sequence was the most appropriate one. They stated the performance in terms of 80% accuracies and 15% false alarm rate.

E. Decision Trees: It is a tree like structure where the leaf node represents or predicts the decision and the non-leaf node represents the various possible conditions that can occur. The decision tree technique has simple implementation, high accuracy and intuitive knowledge expression. This expression is large for small trees and less for deeper and wider trees. The common algorithms for creating decision tree are ID3 [22] and C4.5 [23].

Kruegel and Toth [24] proposed clustering along with decision tree approach to build a signature detection system and compared its performance to SNORT2.0.The speed up varies from 105% to 5 %, depending on the traffic. This paper showed that the combination of decision trees with clustering technique can prove an efficient IDS approach. The decision tree approach using WEKA J48 program was also used in EXPOSURE [25] to detect the malicious domains like botnet command, scam hosts, phishing sites etc. Its performance is satisfactory in terms of accuracy and FAR.

F. Ensemble Learning: It is a supervised machine learning paradigm where multiple learners are trained to solve the same problem. As compared with ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use.

An outlier detector was designed to classify data as anomaly as well as to classify it to one of the attack labels of KDD dataset by Zhang et al. [26] with the use of Random Forests. The Random forest was used as the proximity measure. The accuracy for the DOS, Probe, U2R and R2L attacks were 95%, 93%, 90% and 87% respectively. The FAR is 1%.

G. Evolutionary Computation: It is the collective name for a range of problem-solving techniques like Genetic Algorithms, genetic programming, particle swarm optimization, ant colony optimization and evolution strategies based on principles of biological evolution.

The signature-based model was developed by Li [27] with genetic algorithms used for evolving rules. Abraham et al. [28] also used genetic programming techniques to classify attacks in DARPA 1998 intrusion detection dataset.

H. Inductive Learning: It is a learning method where learner starts with specific observations and measures, begins to detect patterns and regularities, formulates some tentative hypothesis to be explored and ends up with development of some general conclusion and theories. Inductive learning moves from bottom-up that is from specific observations to broader generalizations and theories. Repeated Incremental Pruning to Produce Error Reduction RIPPER [29] applies separate and conquer approach to induce rules in two-class problems. Lee et al. [31] provided a framework for signature-based model using various machine learning and data mining techniques like inductive learning, association rules, sequential pattern mining etc.

I. Naïve Bayes: It is a simple probabilistic classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.Panda and Patra [31] presented the comparison of Naïve Bayes with NN classifier and stated that Naïve Bayes performed better in terms of accuracy but not in FAR. Amor et. al. [32] used Bayesian network as naïve bayes classifier. The paper stated accuracy of 98% with less than 3% false alarm rate.

J. Support Vector Machine: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. *In other* words, given

Table1. Analysis of ML and DM techniques

ML/DM Technique	Method	Data Set	Evaluation Metric	Work
ANN	Signature based	Network Packet level	Acc., RMS	Cannady
ANN	Anomaly	DARPA 1998	DR, FAR	Lippmann & Cunningham
ANN	Anomaly	DARPA 1999	DR, FAR	Bivens et. al.
Association Rules	Signature based	DARPA 1998	DR	Brahmi et. al.
Association Rules	Signature based	Signature attacks	Runtime	Zhengbing et. al.
Association Rules - Fuzzy	Hybrid	KDD 1999 (corrected)	Acc., FAR	Tajbakhsh et. al.
Bayesian Network	Signature based	Tcpdump- botnet traffic	Precision, FAR	Livadas et. al.
Bayesian Network	Signature based	KDD 1999	DR	Jemili et. al.
Clustering- density based	Anomaly	KDD 1999	DR but no actual FAR	Blowers and Williams
Clustering – Sequence	Anomaly	Shell Commands	Acc., FAR	Sequeira and Zaki
Decision Tree	Signature based	DARPA 1999	Speedup	Kruegel and Toth
Ensemble – Random Forest	Hybrid	KDD 1999	Acc., FAR	Zhang et. al.
Evolutionary Computing (GA)	Signature based	DARPA 2000	Acc.	Li
Evolutionary Computing (GP)	Signature based	DARPA 1998	FAR	Abraham et. al.
Inductive Learning	Signature based	DARPA 1998	Acc.	Lee et. al.
Naïve Bayes	Signature based	KDD 1999	Acc., FAR	Panda & Patra
Naïve Bayes	Anomaly	KDD 1999	Acc., FAR	Amor et. al.
Support Vector Machine	Signature based	KDD 1999	Acc.	Li et. al.
Support Vector Machine	Anomaly	DARPA 1998	Acc., FAR	Hu et. al.

labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples.

An SVM classifier was built to classify KDD 1999 dataset by Li et. al.[33] using ant colony optimization for the trainee. This study showed 98% accuracy, however it is not performing well for U2R attacks. RSVM(Robust Support Vector Machine) was used as anomaly classifier by Hu et. al.[34] which showed a better performance with noise having 75% accuracy with no false alarms.

IV. Comparative Analysis And Discussion

The analysis of the work using of ML and DM for cyber security highlights few facts about the growing research area in this field. From the comparative analysis presented in Table 1, it is obvious that the DARPA 1998, DARPA 1999, DARPA2000 KDD 1998, KDD 1999 are the favorite choices of most of the researchers for the dataset for IDS. Most of the researches have used accuracy, detection rate, false alarm rate as the evaluation criteria. There have been multiple approaches that are applied for both anomaly and signature-based detection. Several approaches are appropriate for signature-based others are for anomaly detection. But, the answer to the question about determination of most appropriate approach depends on multiple factors like the quality of the training data, properties of that data, working of the system(online or offline) etc.

V. Conclusions

In this paper, we survey a wide spectrum of existing studies on machine learning and data mining techniques applied for the cyber security. Based on this analysis we then outline key factors that need to be considered while choosing the technique to develop an IDS. These are the quality and properties of the training data, the system type for which the IDS has to be devised and the working nature and environment

of the system. There is a strong need to develop strong representative dataset augmented by network data level. There is also a need to regular updating of the models for the cyber detection using some fast incremental learning ways.

References

1. M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detection:Methods, systems and tools," *IEEE Commun. Surv. Tuts.*, vol. 16, no. 1, pp. 303–336, First Quart. 2014.
2. Y. Zhang, L. Wenke, and Y.-A. Huang, "Intrusion detection techniques for mobile wireless networks," *Wireless Netw.*, vol. 9, no. 5, pp. 545–556, 2003.
3. J. McCarthy, "Arthur Samuel: Pioneer in Machine Learning," AI Magazine, vol. 11, no. 3, pp. 10-11, 1990.
4. K. Hornik,M. Stinchcombe, and H.White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359–366,1989.
5. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6,pp. 386–408, 1958.
6. J. Cannady, "Artificial neural networks for misuse detection," in *Proc 1998 Nat. Inf. Syst. Secur. Conf.*, Arlington, VA, USA, 1998, pp. 443–456.
7. R. P. Lippmann and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Comput.Netw.*, vol. 34, pp. 597–603, 2000.
8. A. Bivens, C. Palagiri, R. Smith, B. Szymanski, and M. Embrechts, "Network-based intrusion detection using neural networks," *Intell. Eng.Syst. Artif. Neural Netw.*, vol. 12, no. 1, pp. 579–584, 2002.
9. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. Int. Conf. Manage. Data Assoc. Comput. Mach. (ACM)*, 1993, pp. 207–216.
10. C. M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," *ACM SIGMOD Rec.*, vol. 27, no. 1, pp. 41–46, 1998.
11. H. Brahmi, B. Imen, and B. Sadok, "OMC-IDS: At the cross-roads of OLAP mining and intrusion detection," in *Advances in Knowledge Discovery and Data Mining*. New York, NY, USA: Springer, 2012, pp. 13–24.
12. H. Zhengbing, L. Zhitang, and W. Junqi, "A novel network intrusion detection system (NIDS) based on signatures search of data mining," in *Proc. 1st Int. Conf. Forensic Appl. Techn. Telecommun. Inf. Multimedia Workshop (e-Forensics '08)*, 2008, pp. 10–16.
13. D. Apiletti, E. Baralis, T. Cerquitelli, and V. D'Elia, "Characterizing network traffic by means of the NetMine framework," *Comput. Netw.*, vol. 53, no. 6, pp. 774–789, Apr. 2009.
14. A. Tajbakhsh, M. Rahmati, and A. Mirzaei, "Intrusion detection using fuzzy association rules," *Appl. Soft Comput.*, vol. 9, pp. 462–469, 2009.
15. D. Heckerman, *A Tutorial on Learning with Bayesian Networks*. New York, NY, USA: Springer, 1998.
16. F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York, NY, USA: Springer, 2001.
17. C. Livadas, R.Walsh, D. Lapsley, andW. Strayer, "Usingmachine learning techniques to identify botnet traffic," in *Proc 31st IEEE Conf. Local Comput. Netw.*, 2006, pp. 967–974.
18. F. Jemili, M. Zaghdoud, and A. Ben, "A framework for an adaptive intrusion detection system using Bayesian network," in *Proc. IEEE Intell. Secur. Informat.*, 2007, pp. 66–70.

19. S. Benferhat, T. Kenaza, and A. Mokhtari, "A Naïve Bayes approach for detecting coordinated attacks," in *Proc. 32nd Annu. IEEE Int. Comput. Software Appl. Conf.*, 2008, pp. 704–709.

20. M. Blowers and J. Williams, "Machine learning applied to cyber operations," in *Network Science and Cybersecurity*. New York, NY, USA: Springer, 2014, pp. 55–175.

21. K. Sequeira and M. Zaki, "ADMIT: Anomaly-based data mining for intrusions," in *Proc 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2002, pp. 386–395.

22. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

23. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

24. C. Kruegel and T. Toth, "Using decision trees to improve signature based intrusion detection," in *Proc. 6th Int. Workshop Recent Adv. Intrusion Detect.*, West Lafayette, IN, USA, 2003, pp. 173–191.

25. L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding malicious domains using passive DNS analysis," presented at the 18th Annu. Netw. Distrib. Syst. Secur. Conf., 2011.

26. J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," *IEEE Trans. Syst. Man Cybern. C: Appl. Rev.*, vol. 38, no. 5, pp. 649–659, Sep. 2008.

27. W. Li, "Using genetic algorithms for network intrusion detection," in *Proc. U.S. Dept. Energy Cyber Secur. Group 2004 Train. Conf.*, 2004, pp. 1–8.

28. A. Abraham, C. Grosan, and C. Martin-Vide, "Evolutionary design of intrusion detection programs," *Int. J. Netw. Secur.*, vol. 4, no. 3, pp. 328–339, 2007.

29. W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, Lake Tahoe, CA, USA, 1995, pp. 115–123.

30. W. Lee, S. Stolfo, and K. Mok, "A data mining framework for building intrusion detection models," in *Proc. IEEE Symp. Secur. Privacy*, 1999, pp. 120–132.

31. M. Panda and M. R. Patra, "Network intrusion detection using Naïve Bayes," *Int. J. Comput. Sci. Netw. Secur.*, vol. 7, no. 12, pp. 258–263, 2007.

32. N. B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. decision trees in intrusion detection systems," in *Proc ACMSymp. Appl. Comput.*, 2004, pp. 420–424.

33. Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 424–430, 2012.

34. W. J. Hu, Y. H. Liao, and V. R. Vemuri, "Robust support vector machines for anomaly detection in computer security," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 282–289.

Fingerprint Image Enhancement Using Different Enhancement Techniques

Upender Kumar Agrawal*
Pragati Patharia**
Swati Kumari***
Mini Priya****

Abstract

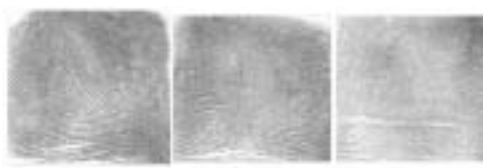
Fingerprint identification is one of the most reliable biometrics technologies. It has applications in many fields such as voting, ecommerce, banking military etc for security purposes. In this paper, we have apllied the Histogram Equalization and Adaptive Histogram Equalization. We have evaluated the performance of the enhancement image method by testing it with fingerprint images.

Keywords: HE, AHE, DNA, CLAHE

I. Introduction

Image Enhancement is one of the necessary step for better analysis. There are various methods to improve the contrast of images [1-3]. Fingerprints are unique patterns, made by friction ridges (raised) and furrows (recessed), which appear on the pads of the fingers and thumbs. They form pressure on a baby's tiny, developing fingers in the womb. The fingerprints are unique. No two persons have been found to have the same fingerprints — Fingerprints are even more unique than DNA, the genetic material in each of our cells. Although identical twins can share the same DNA - or at least most of it -they can't have the same fingerprints. Friction ridge patterns are grouped into three distinct types—loops, whorls, and arches—each with unique variations, depending on the shape and relationship of the ridges:

Upender Kumar Agrawal*
upeagrawal@gmail.com
Pragati Patharia**
pathariapragati@gmail.com
Swati Kumari***
swati.kumari3661@gmail.com
Mini Priya
minipriya9496@gmail.com
Guru Ghasidas Viswavidyalya, Bilaspur



Loops - prints that recurve back on themselves to form a loop shape. Divided into radial loops (pointing toward the radius bone, or thumb) and ulnar loops (pointing toward the ulna bone or pinky), loops account for approximately 60 percent of pattern types.

Whorls - form circular or spiral patterns, like tiny whirlpools. There are four groups of whorls: plain (concentric circles), central pocket loop (a loop with a whorl at the end), double loop (two loops that create an S-like pattern) and accidental loop (irregular shaped). Whorls make up about 35 percent of pattern types.

Arches - create a wave-like pattern and include plain arches and tented arches. Tented arches rise to a sharper point than plain arches. Arches make up about five percent of all pattern types.

2. Histogram Eqalization

Histogram equalization (HE) is one of the popular technique for contrast enhancement of images. It is one of the well-known methods for enhancing the

contrast of a given image in accordance with the samples distribution. HE is a simple and effective contrast enhancement technique which distributes pixel values uniformly such that enhanced image have linear cumulative histogram. HE has been widely applied when the image need enhancement, such as medical image processing radar image processing, texture synthesis and speech recognition.

It stretches the contrast of high histogram regions and compresses the contrast of low histogram region. The goal of histogram equalization is to remap the image grey levels so as to obtain a uniform (flat) histogram in the other words to enhance the image quality .HE based methods are reviewed and compared with image quality measurement (IQM) tools such as Peak Signal to Noise Ratio (PSNR) to evaluate contrast enhancement.

Peak Signal to Noise Ratio (PSNR)

Let, X(i,j) is a source image that contains M by N pixels and a reconstructed image Y(i,j), where Y is reconstructed by decoding the encoded version of X(i,j). In this method, errors are computed only on the luminance signal; so, the pixel values X(i,j) range between black (0) and white (255)[6-7]. First, the mean squared error (MSE) of the reconstructed image is calculated. The root mean square error is computed from root of MSE. Then the PSNR in decibels (dB) is computed as;

PSNR = 20log10 (Max(Y(i,j) RMSE)

Greater the value of PSNR better the contrast enhancement of the image.

3. Adaptive Histogram Equalization

Adaptive histogram equalization (AHE) is a image processing technique used to improve contrast in images [1-3]. It differs from ordinary histogram equalization in the respect that the adaptive method computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the lightness values of the image. It is therefore suitable for improving the local contrast and enhancing the definitions of edges in each region of an image. However, AHE has a tendency to over amplify noise in relatively homogeneous regions of an image. A variant of adaptive histogram equalization called contrast limited adaptive histogram equalization (CLAHE) prevents this by limiting the amplification. The size of the neighbourhood region is a parameter of the method. It constitutes a characteristic length scale: contrast at smaller scales is enhanced, while contrast at larger scales is reduced [4-5].Due to the nature of histogram equalization, the result value of a pixel under AHE is proportional to its rank among the pixels in its neighbourhood. This allows an efficient implementation on specialist hardware that can compare the centre pixel with all other pixels in the neighbourhood.

4. Original Data Of Fingerprint Thumb Impression :

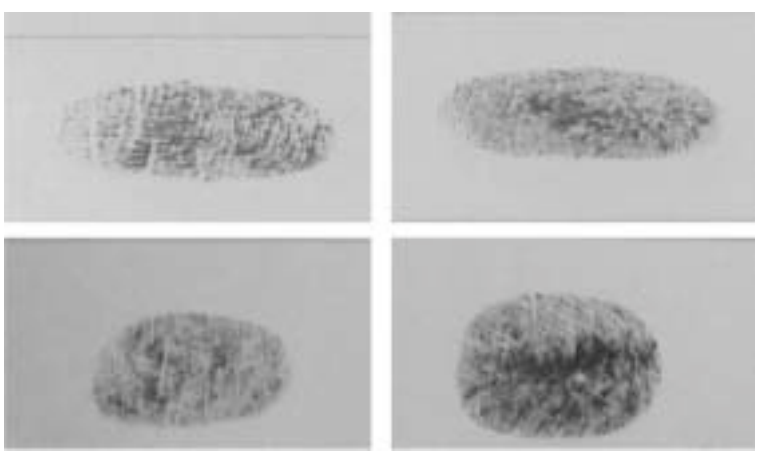


Fig 1: Sample variations of individual left hand thumb impression showing arches, loops and whorls.

5. Results And Comparision

The above discussed methodologies have been implemented by using Matlab. For the testing purpose we have created two Image Database. At first we captured fingerprint image using mobile camera then we enhance the fingerprint image using histogram and adaptive histogram techniques. Results from the above implementation are in described in the following section.

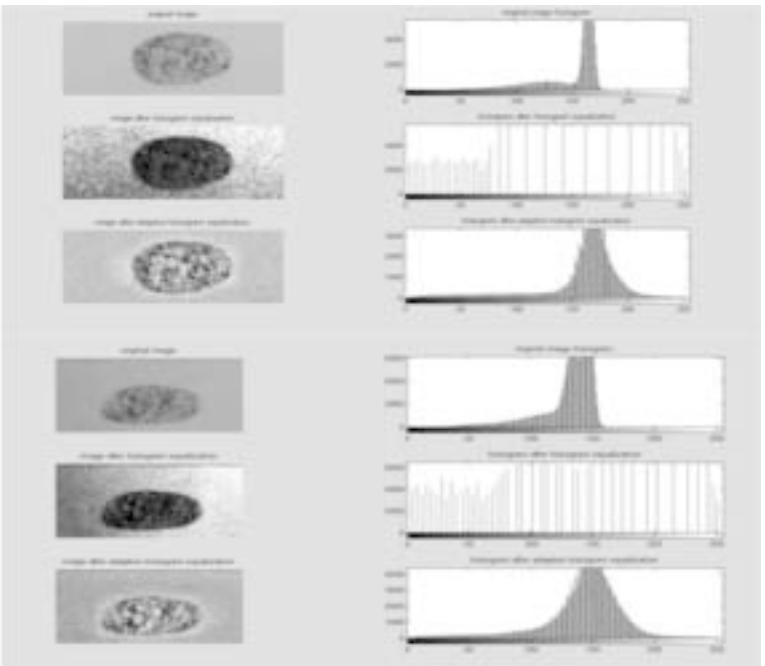
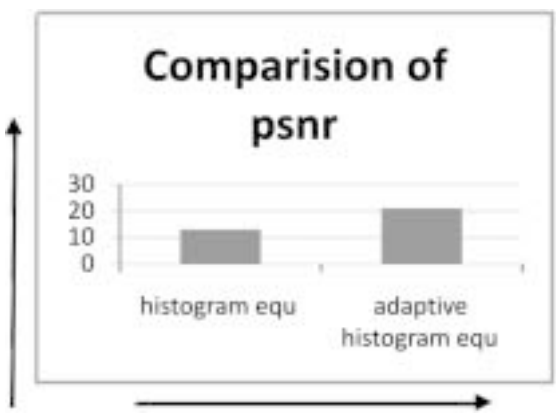


Fig 2. Original image and its histogram, Histogram equalization and its histogram, Adaptive histogram equalization and its histogram.

Comparision of PSNR



(Variation of histogram technique)

6. Conclusion

Based on the result of the experiment phase in this research we found. Firstly, the use of Histogram Equalization enable to increase fingerprint contrasts and for brightness preserving .Secondly by using Adaptive Histogram Equalization (AHE) is an excellent contrast enhancement method for both natural images and medical and other initially non-

visual images. As conclusion, the proposed Technique produces a fine fingerprint image quality. This graph shows the comparison of PSNR. The output shows that the PSNR of adaptive histogram equalization is more than histogram equalization.

References

1. Z. M. Win and M. M. Sein, ¯Fingerprint recognition system for low quality images, presented at the SICE Annual Conference, Waseda University, Tokyo, Japan, Sep. 13-18, 2011.
2. Dr. Muna F. Al-Samaraie, "A New Enhancement Approach for Enhancing Image of Digital Cameras by Changing the Contrast", International Journal of Advanced Science and Technology Vol. 32, July, 2011.pp.-13-22.
3. Mustafa Salah Khalefa 1, Zaid Amin Abduljabar 2 and Huda Ameer Zeki, "Fingerprint Image Enhancement by Develop Mehtre Technique", Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011,pp.-171-182.
4. D. Ezhilmaran and M. Adhiyaman, "A Review Study on Fingerprint Image Enhancement Techniques", International Journal of Computer Science & Engineering Technology (IJCSET)Vol. 5 No. 06 Jun 2014, ISSN : 2229-3345,625-631.
5. Darshan Charan Nayak, "Comparative Study of Various Enhancement Techniques for Finger Print Images", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, ISSN :0975-9646, 1900-1905.
6. C.Nandini and C.N.Ravikumar, "Improved fingerprint image representation for recognition," International journal of computer science and information technology, MIT Publication, Vol. 01-no.2, 2011,pp.59-64.
7. J.Choudhary, Dr.S.Sharma, J.S.Verma, "A new framework for improving low quality fingerprint images," international journal of computer technology and application. Vol.2, no.6, pp.1859 -1866,2011.

Data Mining in Credit Card Frauds: An Overview

Vidhi Khurana*
Ramandeep Kaur**

Abstract

With the increasing awareness of customers amongst plastic money and internet banking, the number of frauds in transactions have also emerged. In order to detect these frauds, various data mining techniques can be applied. Financial Fraud Detection(FFD) has been a major concern among the leading organizations and the banks. Hence a framework has been proposed so as to detect the fraud in the early stages as well as forecast which transactions are prone to fraudulent activities. This paper reviews the previous research conducted by the leading researchers in their areas with a focus on credit card fraud detection and prevention using data mining approaches.

Keywords: Credit Card, Data mining, Financial Fraud Detection, Fraud Prevention

I. Introduction

Data Mining has been a very vibrant and upcoming field in all the prevailing industries. From a small and independent IT firm, banking organizations, convenience stores, to leading industries, the implications of data mining can be felt. It may be defined as the logical process of extraction of hidden and interesting information from the huge databases[1]. It is a methodology of mining of knowledge from the given data sources. Hence may aid in Knowledge discovery.

Data Mining can be categorized into three identifiable steps: (i) Exploration (ii) Pattern Identification and (iii) Deployment. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining, viz.,Descriptive and Classification and Prediction[27]. Mined knowledge can be used in various domains like: fraud detection, production control, science exploration and market analysis. Financial Fraud Detection(FFD) is of high priority at present. Data Mining help in detection of financial frauds by analysing patterns hidden in the transaction data [8]. FFD is vital for the prevention of the often devastating consequences of financial fraud. According to the 2008 Javelin fraud survey report, victims who detected the fraud within 24 hours were defrauded for an average of $428. Victims who did not discover the fraud up to a month later suffered an average loss of $572[6].

Financial Fraud can be classified into various categories as depicted in Table 1.

Bank Frauds are very devastating and have a severe repercussion on the organizations. It comprises of all the fraudulent activities involved in the banking sector. It is broadly classified into two categories: i) External: here the assassin are outside the bank ii) Internal: here bank personnel commits the fraud. Card fraud, mortgage fraud and money laundering are few instances of bank fraud. Insurance Fraud is an activity of obtaining fraudulent outcomes from an insurance company[8]. It can be committed by consumer, broker and agents, insurance company employees and others. Automobile fraud and healthcare fraud are in top category of this classification [2,13]. Securities and commodities fraud is a type of white collar crime that can be committed by individuals. [investopedia] The types of misrepresentation involved in this crime include providing false information, withholding key information, offering bad advice, and offering or acting on inside information. Other related financial frauds include corporate and mass marketing fraud. Mass communication media such as telephones and internets are used in mass market fraud [14]. Mass-marketing fraud schemes generally fall into two broad

Vidhi Khurana*
Pursuing MCA from Institute of Information Technology & Management
Ramandeep Kaur**
Assistant Professor
Institute of Information Technology & Management

Table 1: Classification for Financial Fraud based on FBI, 2007

Financial fraud based categories	Fraudulent activities
Bank fraud	Mortgage fraud, Asset forfeiture/money laundering
Insurance fraud	Healthcare fraud, Insurance fraud
Securities and commodities fraud	Securities and commodities fraud
Other related financial fraud	Corporate fraud, Mass marketing fraud

categories: (1) schemes that defraud numerous victims out of comparatively small amounts, such as several hundred dollars, per victim; and (2) schemes that defraud comparatively less numerous victims out of large amounts, such as thousands or millions of dollars per victim.

The objective of this paper is to describe generalized architecture of Financial Fraud detection as well as the techniques of preventing the frauds. Special focus has been laid on Credit Card Financial Frauds. The remainder of the paper is divided in the following sections: Section II deals with a detailed review of literature. Section III deals with a framework for Financial Fraud Detection. Section IV deals with Fraud detection in Credit Cards. Section V gives a concluding remark on the review carried out.

II. Literature Review

Vast research has been carried out in the field of data mining and fraud detection but the challenge in dealing with the increasing number of frauds remains the same. Data mining enables a user to seek valuable information and their interesting relationships [24]. A number of data mining techniques are available such as decision trees, neural networks (NN), Bayesian belief networks, case based reasoning, fuzzy rule-based reasoning, hybrid methods, logistic regression, text mining, feature selection etc. Financial fraud is a serious problem worldwide and more so in fast growing countries like China[21]. According to Kirkos et al. [7], some estimates stated that fraud cost US business more than $400 billion annually. An innovative fraud detection mechanism was developed on the basis of

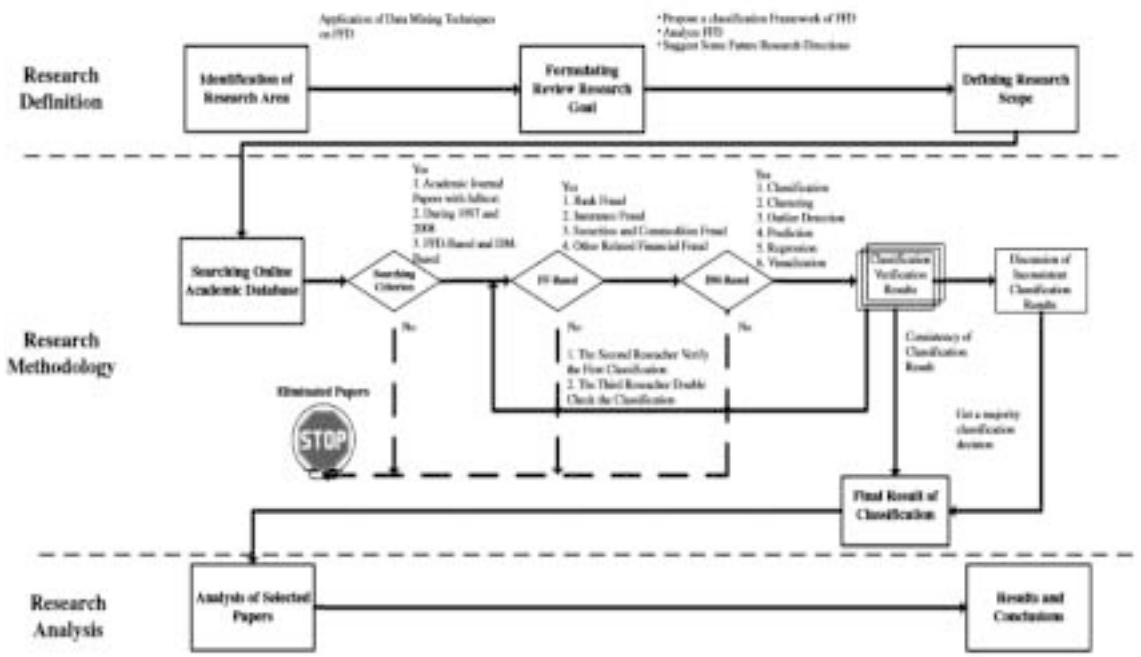


Fig 1: Methodological Framework for research[8]

Table 2: Research on data mining techniques in FFD[8]

FF-based categories	Fraudulent activities	Data mining application class	Data mining techniques
Bank fraud	Credit card fraud	Classification	Ada boost algorithm, decision trees, CART, RIPPER, Bayesian Belief Network, Neural networks, discriminant analysis
			K-nearest neighbor, logistic model, discriminant analysis, Naïve Bayes, neural networks, decision trees
			Support vector machine, evolutionary algorithms
		Clustering	Hidden Markov Model
			Self-organizing map
	Money laundering	Classification	Network analysis
Insurance fraud	Crop insurance fraud	Regression	Yield-switching model
			Logistic model, probit model
	Healthcare insurance fraud	Classification	Association rule
			Polymorphous (M-of-N) logic
			Self-organizing map
		Visualization	Visualization
		Outlier detection	Discounting learning algorithm
	Automobile insurance fraud	Classification	Logistic model
			Neural networks
			Principal component analysis of RIDIT(PRIDIT)
			Logistic model
			Logistic model, decision trees, neural networks, support vector machine, K-nearest neighbor, Naïve Bayes, Bayesian belief network
			Fuzzy logic
			Logistic model
			Logistic model, Bayesian belief network
			Self-organizing map
			Naïve Bayes
		Prediction	Evolutionary algorithms
			Logistic model
		Regression	Probit model
			Logistic model
			Probit model
Other related financial fraud	Corporate fraud	Classification	Neural networks, decision trees, Bayesian belief network
			Multicriteria decision aid (MCDA), Utilite's Additives DIScriminantes (UTADIS)
			Evolutionary algorithms
			Fuzzy logic
			Neural networks
			Neural networks, logistic model
			Logistic model
			CART
			Decision trees, neural networks, Bayesian belief network, K-nearest neighbor, RIPPER, support vector machine, stacking variant methodology
		Clustering	Naïve Bayes
		Prediction	Neural networks
		Regression	Logistic model
			Logistic model
			Logistic model

Zipf's Law with a purpose of assisting the auditors in reviewing the bulbous volumes of datasets while at the same time intending to identify any potential fraud records[26]. The study of Bolton and Hand [22] provides a very good summary of literature on fraud detection problems. Some researchers used methods such as ID3 decision tree, Bayesian belief, back-propagation Neural Network to detect and report the financial frauds[7,12]. Fuzzy logic based techniques based on soft computing were also incorporated to deal with the frauds [15, 16]. Panigrahi et. al.[25] suggested a four component fraud detection solution with an idea to determine a set of suspicious transactions and then predict the frauds by running Bayesian learning algorithm. Further, a set of fuzzy association rules were extracted from a data set containing genuine and fraudulent transactions w.r.t credit cards to analyze and compare the frauds. It was suggested that novel combination of meta-heuristic approaches, namely the genetic algorithms and the scatter search when applied to real time data, may yield fraudulent transactions which are classified correctly[5]. Padhy et al (2012) provided a detailed survey of data mining applications and its feature scope. A number of researchers also discussed the application of data mining in anomaly detection [17, 19, 20, 23].

III. Framework of FFD

Methodological framework for review is a three step process: i) Research Definition ii) Research Methodology and iii) research analysis. Research definition is a phase mining technique.Goal of the research is to create a classification framework for data mining techniques applicable to FFD. Research scope here is the literature comprising application of data

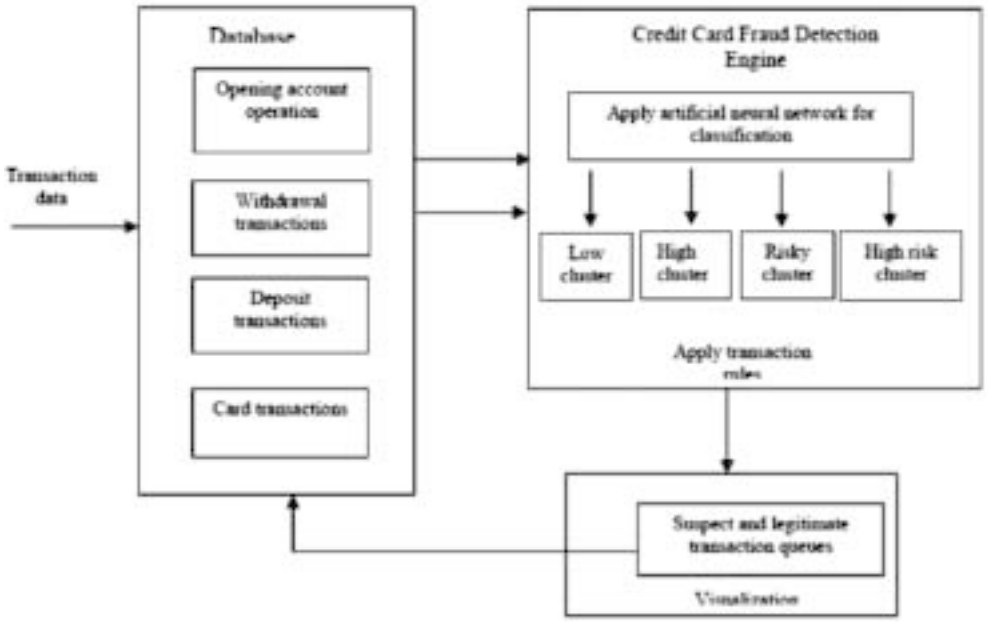


Fig 2: Architecture for Credit Card Fraud Detection[10]

mining techniques on FFD published from 1997 to 2008. Phase to is the research methodology. In this phase the online academic databases are searched for FFD. In each iteration these databses are filtered out to obtain the articles that were published in the academic journals(1997-2008) and should present data mining techniques along with application to FFD. A detailed process for FFD has been depicted in Fig 1. All the obtained articles consistency are verified and final result of classification is passed to third phase of the framework. Research analysis phase includes the analysis of the selected where the topic or area of research is identified for formulating the research goal and definingg the scope of the performed research. Here identified research area: the academic reserch on FFD that applies data papers to formulate conclusion and results based on the analysis of paper[8].

IV. Fraud Detection in Credit Cards

Credit card fraud is sort of identity theft, where an unauthorized person makes fraudulent transactions. It can be classified into: Application fraud and Behaviour fraud. Application fraud occurs when a fraudster gets a credit card issued from companies by providing false information[3]. It is very serious because victim may learn about the fraud too late.

Various data mining techniques used in credit card fraud detection are logistic regression, support vector machine and random forests. Credit card fraud detection scheme scans all the transactions inclusive of fraudulent transactions[10]. Data obtained from the data warehouse is divided into various dataset. Dataset comprises of primary attributes (account number, sale, purchase, date name and many others) and derived attributes (for instance transactions grouped monthly). Derived attributes are not precise, which causes approximation of results and therefore not accurate information. Therefore derived attributes are limitation to the credit card fraud detection scheme. The implemented architecture [Fig2] comprises of database interface subsystem and credit card fraud (CCF) detection engine. The former enables the reading of transactions, i.e. it acts as an interface for banking software.

In the CCF detection subsystem, the host server checks every transaction rendered to it using neural networks and transactions business rules.

V. Conclusion

Data mining gained weightage in the areas where finding the patterns, forecasting, discovery of knowledge etc., is required and becomes obligatory in

different industrial domains. Various techniques and algorithms such as feature selection, classification, memory based reasoning, clustering etc., aids in fraud detection in areanas of insurance, financial frauds etc.. Financial sector has been majory affected ny fradulent activities due to increase in conversion rate of non-internet users to internet users. A detailed review was conducted to understand how these financial frauds can be detected and avoided using data mining techniques. A special reference to Credit card frauds was mentioned to understand the architecture of credit card fraud detection.

References

1. Bose, R.K. Mahapatra, "Business data mining — a machine learning perspective", *Information Management,* vol.39, no.3, pp.211–225, 2001.
2. Coalition against Insurance Fraud, "Learn about fraud," http://www.insurancefraud.org/learn_about_fraud.htm, Last accessed 23 January 2017.
3. Credit Card Fraud: An Overview, Legal Information Institute, web: https://www.law.cornell.edu/wex/credit_card_fraud, Last Accessed: 23 January 2017.
4. D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "Association rules applied to credit card fraud detection," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3630–3640, 2009.
5. E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13057–13063, 2011.
6. E. Joyner, "Enterprisewide Fraud Management", *Banking, Financial Services and Insurance*, Paper 029, 2011
7. E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statement", *Expert Systems with Applications*, vol.32, pp.995–1003, 2007.
8. E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 2592–2602, 2009.
9. FBI, Federal Bureau of Investigation, Financial Crimes Report to the Public Fiscal Year, Department of Justice, United States, 2007, http://www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm.
10. F. N. Ogwueleka, "Data Mining Application In Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp.311 – 322, 2011.
11. F.N. Ogwueleka, and H.C. Inyiama, "Credit card fraud detection using artificial neural networks with a rule-based component', The IUP Journal of Science and Technology, vol.5, no.1, pp.40-47, 2009.
12. J.E. Sohl and A.R. Venkatachalam, "A neural network approach to forecasting model Selection", *Information & Management*, vol.29, no.6, pp. 297–303, 1995.
13. J.L. Kaminski, "Insurance Fraud", OLR Research Report, http://www.cga.ct.gov/2005/rpt/2005-R-0025.htm. 2004
14. "Mass Marketing Fraud(MMF)", Strategy, Policy & Training Unit, Department of Justice, http://www.justice.gov/criminal-fraud/mass-marketing-fraud, Last Accessed: 23 January 2017.
15. M. Delgado, D. Sa´nchez, and M.A. Vila, "Fuzzy cardinality based evaluation of quantified sentences", *International Journal of Approximate Reasoning*, vol.23, pp.23–66, 2000.
16. M. Delgado, N. Marý´n, D. Sa´nchez, and M.A.Vila, "Fuzzy association rules: General model and applications", *IEEE Transactions on Fuzzy Systems*, vol.11, no.2, pp.214–225, 2003.

17. N. Kaur, "Survey paper on Data Mining techniques of Intrusion Detection", *International Journal of Science, Engineering and Technology Research*, vol. 2, no. 4, pp. 799-804, 2013.

18. N. Padhy, P. Mishra, and R. Panigrahi, "The Survey of Data Mining Applications and Feature Scope", *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 3,pp. 43-58, 2012.

19. P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava and P.N.Tan, "Data mining for network intrusion detection", *Proceedings of NSF Workshop on Next Generation Data Mining*, pp. 21-30, 2002.

20. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", *Computers and security*, vol.28, no. 1, pp. 18-28, 2009.

21. P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *", Decision Support Systems*, vol. 50, no. 2, pp. 491–500, 2011.

22. R. Bolton, and D. Hand, 'Statistical fraud detection: A review", *Statistical Science*, vol.17, pp.235–255, 2002.

23. S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Comput. Sci.*, vol. 60, no. 1, pp. 708–713, 2015.

24. S. H. Weiss, and N. Indurkhya, "Predictive Data Mining: A Practical Guide", , *CA: Morgan Kaufmann Publishers*, 1998.

25. S. Panigrahi, A. Kundu, S. Sural, and A. Majumdar, "Credit card fraud detection a fusion approach using Dempster–Shafer theory and bayesian learning", *Information Fusion*, pp.354–363, 2009.

26. S.-M. Huang, D.C. Yen, L.-W. Yang and J.-S. Hua, "An investigation of Zipf's Law for fraud Detection*", Decision Support Systems*, vol.46, no. 1, pp. 70–83, 2008.

27. Tutorialspoint, "Data mining Tasks", http://www.tutorialspoint.com/ data_mining/ dm_tasks.htm, Last Accessed: 24 January 2017.

Review of Text Mining Techniques

Priya Bhardwaj*
Priyanka Khosla**

Abstract

Data mining is a process of discovering potential and practical, previously unknown patterns from large pre existing databases. Text mining is a realm of data mining in which large amount of structured and unstructured text data is analyzed to produce information of high commercial value. Analyzing textual data requires context analysis. This paper represents the current research status of text mining. Association rules, a novel technique in text mining is gaining increasing currency among research scholars is discussed. Based on studied attempts, the potential future research activities have been proposed.

Keywords: component; formatting; style; styling; insert (key words)

I. Introduction

With the evolution of internet and rapid developments in information technology enormous amount of textual data is generated in the form of blogs, tweets and discussion forums. The data potentially has a lot of hidden information which can intuitively predict human behavior. The major challenge is to uncover relationships and associations in the data which is in various formats i.e. unstructured data [1]. Text mining aims at revealing the concealed information by using various techniques that are capable of coping up with large amount of structured data on one hand and handling the vagueness, fuzziness and uncertainty of the unstructured data on the other. Text mining or knowledge discovery from text (KDT) — for the first time mentioned in Feldman et al. [2] — deals with the computational analysis of textual data. It is an interdisciplinary field involving techniques from information extraction, information retrieval as well as Natural Language Processing (NLP) and integrates them with the algorithms and methods of data mining, statistics and machine learning.

The most convenient way of storing information is believed to be text. In the recent surveys it is considered that 80% of company's information is contained in text [4] and analysis of this information is required for making strategic decisions.

This paper introduces the current research status of text mining. Section III describes some general models used for mining text. The applications of text mining and the related techniques are discussed in Section IV followed by a conclusion.

II. State of the Art

Hans Peter Luhn[6] in 1958, published an article in journal of IBM which discusses about the automatic extraction by data processing machine and classifies the document on the word frequency statistics. This was considered to be one of the primitive definitions of business intelligence.

The research in the field of text mining continued and many scholars carried prolific research in the field. In the 1st International Conference on Data Mining and Knowledge Discovery in 1995 Feldman et al. [5] proposed Knowledge Discovery in Database (KDT). Supervised [7] and Unsupervised [8][9] learning algorithms are used to uncover hidden patterns in the textual documents.

Subsequently, other outstanding work done is in the field including dimensionality reduction on the basis of correlation in feature extraction [13]-[14]; soft set approach using association rule mining [15] by introducing SOFTAPRIORI that discovers relationships more accurately; sentiment analysis for online forums hotspot detection and forecast [16];

Priya Bhardwaj*
Assistant Professor
Institute of Information Technology and Management, Delhi, India
Priyanka Khosla**
Assistant Professor
Institute of Information Technology and Management, Delhi, India

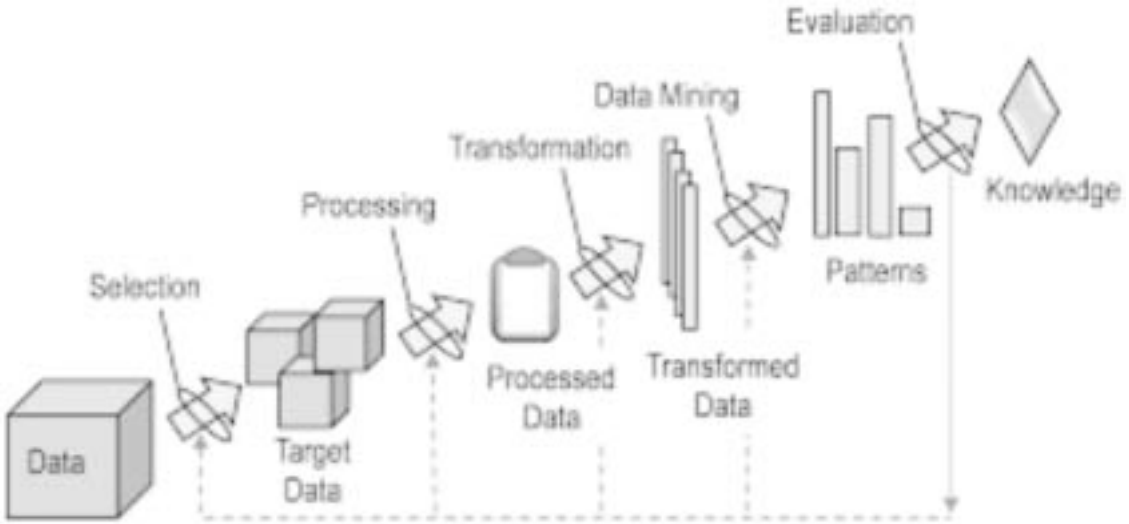


Figure 1: Knowledge Discovery Process

sentiment analysis using self organizing maps and ant clustering [17]; and text mining in various other fields such as stock prediction [18], web mining [19], digital library [20] and so on.

III. Text mining Models

Generally text mining is a four step process which is text preprocessing, data selection, data mining and post processing..

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Data Cleaning

The textual data available for mining is generally collected over web from the tweets, discussion forums and blogs. The data set available from these sources is in various formats i.e. "unstructured". We need to "clean" the data by performing parsing of data, missing value treatment, removing inconsistencies. After performing the desired operations the data set should be consistent with the system.

B. Data selection and transformation

The textual data available for mining is generally collected over web from the tweets, discussion forums and blogs. The data set available from these sources is in various formats i.e. "unstructured". We need to "clean" the data by performing parsing of data, missing value treatment, removing inconsistencies. After performing the desired operations the data set should be consistent with the system.

C. Data Mining

After the document being converted into the intermediate form data mining techniques can be applied to different type of data according (structured, semi- structured and unstructured) to recognize relationships and patterns. The various data mining techniques are discussed in detail in section IV.

D. Data Post processing

It includes the tasks of evaluation and visualization of the knowledge coming out after performing text mining operations.

IV. Techniques Used in Data Mining

The progress of Information Technology has produced large amount of data and data repositories in diverse areas. The research made in databases has further given rise to the techniques used to store and process the data for decision making. Thus, Data mining is a process of finding useful patterns from large amount of data and is also termed as knowledge discovery process which states the knowledge mining or extraction from large amount of data.

Machine Learning Algorithms

- Unsupervised Machine Learning :It is a type of machine learning algorithm that is used to draw

conclusion from datasets that consists of input data without the labeled responses. The most familiar unsupervised learning method is cluster analysis, that is used for exploratory data analysis to find hidden patterns or grouping in data.

- Supervised Machine Learning Algorithm: It is a type of machine learning algorithm that uses a identified dataset (called the training dataset) in order to make predictions. The training data set comprises of input data and response values. From this dataset, the supervised learning algorithm searches for a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets.

A. Classification Technique:

Classification is the commonly used data mining technique that employs training dataset or pre-classified data to generate a model that is used to classify records according to rules. This technique of data mining is used to find out in which group each data instance is related within a given dataset using the training dataset. It is used for classifying data into different classes according to some constraints. Credit Risk analysis and fraud detection are the application of this technique. This algorithm employs decision tree or neural network-based classification algorithms. Classification is a Supervised learning that involves the following steps:

Step 1: Rules are extracted using the learning algorithm from (create a model of) the training data. The training data are pre classified examples (class label is known for each example).

Step 2: Evaluation of the rules on test data. Usually split known data into training sample (2/3) and test sample (1/3).

Step 3: Apply the generated rules on new data.

Thus, the classifier-training algorithm uses the pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called as a classifier. Rules are generated from it that further helps in making decisions.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. Clustering RulesTechnique:

It is the task of grouping objects in such a way that objects in the same group or cluster are similar in one sense or another to each other than to those objects present in another groups. Thus it is an identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Types of clustering methods involves

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

C. Association Rules Technique:

Association is a data mining technique that discovers the probability of the co-occurrence of items in a collection. The relationships between co-occurring items are expressed as association rules. These rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk." Therefore both eggs and milk together are associated with each other and are likely to be placed together to increase the sales of both the product. Thus association rules helps industries and businesses to make certain decisions, such as cross marketing, customer shopping, designing of catalogue etc. Association Rule algorithms should be able to generate rules with confidence values less than one. Although the number of possible

Table I. Tasks With Algorithms

Examples of tasks	Algorithms to use
Predicting a discrete attribute	**Decision Tree Algorithm**
Flag the customers in a prospective buyers list as good or poor prospects.	
Calculate the probability that a server will fail within the next 6 months.	**Clustering Algorithm**
Categorize patient outcomes and explore related factors.	**Neural Network Algorithm**
Predicting a continuous attribute	**Decision Tree Algorithm**
Forecast next year's sales.	
Predict site visitors given past historical and seasonal trends.	
Generate a risk score given demographics.	
Predicting a sequence:	**Clustering Algorithm**
Perform click stream analysis of a company's Web site.	
Analyze the factors leading to server failure.	
Capture and analyze sequences of activities during outpatient visits, to formulate best practices around common activities.	
Finding groups of common items in transactions:	**Association Algorithm**
Use market basket analysis to determine product placement.	**Decision Tree Algorithm**
Suggest additional products to a customer for purchase.	
Analyze survey data from visitors to an event, to find which activities or booths were correlated, to plan future activities.	
Finding groups of similar items:	**Clustering Algorithm**
Create patient risk profiles groups based on attributes such as demographics and behaviors.	
Analyze users by browsing and buying patterns.	
Identify servers that have similar usage characteristics.	

Association Rules for a given dataset is generally very large and among that a high proportion of the rules are usually of little value. Types of association rules are:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

V. Choosing an Algorithm by Task

To help you select an algorithm for use with a specific task, the following table provides suggestions for the types of tasks for which each algorithm is traditionally used.

VI. Conclusion

The paper has provided a concise introduction about the state of the art of text mining. In the next section the steps required to extract valuable information from the data set are described. Consequent section summarized various data mining techniques such as classification, clustering and association rule. Text mining gives a direction to the upcoming fields like artificial intelligence, therefore it needs the continuous improvement in order to grow its application areas.

References

1. Ah Hwee Tan et al., "Text Mining: The state of the art and the challenges", *Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases*, pp. 65-70, 2000.
2. R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117, 1995.
3. Marti A. Hearst, Untangling text data mining, pp. 3-10, 1999, University of Maryland.
4. S.Grimes. "Unstructured data and 80 percent rule." Carabridge Bridgepoints, 2008
5. H. P. Luhn, "A Business Intelligence System", *Ibm Journal of Research & Development*, vol. 2, no. 4, pp. 314-319, 1958.
6. M. E. Maron, J. L. Kuhns, "On Relevance Probabilistic Indexing and Information Rctrieval", *Journal of the Acm*, vol. 7, no. 3, pp. 216-244, 1960.
7. Larsen, Bjornar, and Chinatsu Aone. "Fast and effective text mining using linear-time document clustering." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.
8. Jiang, Chuntao, et al. "Text classification using graph mining-based feature extraction." Knowledge-Based Systems 23.4 (2010): 302-308.
9. Liu, Wei, and Wilson Wong. "Web service clustering using text mining techniques." International Journal of Agent-Oriented Software Engineering 3.1 (2009): 6-26.
10. Ronen Feldman, I. Dagan, H. Hirsh, "Mining Text Using Keyword Distributions", Journal of Incelligent Information Systems, vol. 10, no. 3, pp. 281-300, 1998.
11. J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret, "Information mining: use of the document dimensions to analyse interactively a document set", European Colloquium on IR Research: ECIR, pp. 66-77, 2001.
12. M. Ghanem, A. Chortaras, Y. Guo, A. Rowe, J. Ratcliffe, "A Grid Infrastructure For Mixed Bioinformatics Data And Text Mining", Computer Systems and Applications 2005. The 3rd ACS/IEEE International Conference, vol. 29, pp. 41-1, 2005.
13. Haralampos Karanikas, C. Tjortjis, B. Theodoulidis, "An Approach to Text Mining using Information Extraction", Proc. Workshop Knowledge Management Theory Applications (KMTA 00, 2000.
14. Qinghua Hu et al., "A novel weighting formula and feature selection for text classification based on rough set theory", *Natural Language Processing and Knowledge Engineering 2003. Proceedings. 2003 International Conference on IEEE*, pp. 638-645, 2003.
15. Nahm, Un Yong, and Raymond J. Mooney. "Mining soft-matching association rules." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.
16. Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." Decision support systems 48.2 (2010): 354-368.
17. Chifu, Emil ªt, Tiberiu ªt Leþia, and Viorica R. Chifu. "Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps." Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. IEEE, 2015.
18. Nikfarjam, Azadeh, Ehsan Emadzadeh, and Saravanan Muthaiyah. "Text mining approaches for stock market prediction." Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. Vol. 4. IEEE, 2010.
19. Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.
20. Fuhr, Norbert, et al. "Digital libraries: A generic classification and evaluation scheme." International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2001.

Security Vulnerabilities of Websites and Challenges in Combating these Threats

Dhananjay*
Priya Khandelwal**
Kavita Srivastava***

Abstract

The public use of Internet started in 1990s. Since then, billions of websites have been developed. Also the technology has caused development of websites easier and less costly. It has enabled people to make their online presence quickly and easily through the use of websites. In recent years a number of Open Source CMS (Content Management Systems) have developed which enabled creation of websites in minutes. This large number of adoption of website by people also led to the growth of unskilled website administrators and developers. As a result almost 75% of websites are found to be infected with malware.

Google reported in March 2015 that around 17 million websites either have installed malicious software or trying to steal information. This number is increased to 50 million in March 2016. Google blocks nearly 20000 websites per week for malware and phishing. Most of these blocked websites are found to be implemented with WordPress, Joomla, and Magento.

This paper addresses various security vulnerabilities found in websites implemented with different technologies, methods of combating these vulnerabilities and research and development in this direction.

Keywords:

Introduction

Security is one of the critical phases of quality of any software or any application. Security testing of web applications attempts to figure out various vulnerabilities, attacks, threats, viruses etc related to the respective application. Security testing should attempt to consider as many as potential attacks as possible.

Increase in usage of web applications has opened the doors for hackers around the world for penetrating these applications. Hackers and attacker try to find out loop holes in coding of web applications to harm them in a number of ways such as applying Denial of Service (DoS) attack, spreading malware, illegal redirection to another website access or posting malicious content by gaining access into the application.

In order to prevent such attacks, the most effective method is to develop web applications by applying good and secure coding skills. Most of the web applications which suffer from security vulnerabilities have common coding problems such as improper input field validations, wrong or no session management, poor configuration settings in web applications as well as the web server which runs these applications.

We can organize the threats to web applications in a number of classes like Inadequate Authentication, Cross-Site Scripting, SQL Injection and so on. In the next sections all these web security vulnerability classes are elaborated.

Security Issues in Websites

In this section we discuss the classification of website security vulnerabilities.

Dhananjay*
BCA, IV, IITM
Priya Khandelwal**
BCA, IV, IITM
Kavita Srivastava***
Associate Professor, IITM

(1) Poor Access Grant and Lack of Sufficient Authorization

Authorization it is a process where a requester is allowed to perform an authorized action or to receive a service. Often a web application grants the access of some of its features to specified users only. The web application verifies the credentials of users trying to access these features through a Login page. This type of vulnerability exists in an application if users can access these features without verification through certain links or tabs and access other users' accounts also.

(2) Poorly Implemented Functionality

This kind of vulnerability exists in a website due to its own code which results in harmful consequences such as password leak, consuming large amount of resources and giving access to administrative features. The security breaches may lead to the disclosure of any confidential or sensitive data from any web application.

(3) Inadequate Exception and Error Handling Mechanisms

The error messages and exception handling code should return only limited amount of information which prevents an attacker to identify a place for SQL Injection. For Instance consider the following code.

…catch(Exception e) {Console.WriteLine(e.Message);} If it is an SQL exception, this code cn display information related too database.

(4) Brute Force Attack

This is the process of trial and error in order to guess users' credentials such as user name, password, security questions for the purpose of hacking a user's account.

(5) Data/Information Leak

This kind of security breache may lead to the disclosure of any confidential or sensitive data from any web application. This vulnerability exists in web applications as a result of improper use of technology for developing application. It can cause revealing of developer's comments, source code, etc. It can give enough information to hacker for exploiting the system.

(6) Inadequate Authentication

Authentication this involves confirming the identity of an entity/person claiming that it is a trusted one. Sometimes a developer doesn't provide a link for administrative access. Yet administrative access is provided through another folder on the server. If a hacker identifies its path it becomes very easy to exploit the application.

(7) Spoofing

This is an attack where an attacker tries to masquerades another program or user by falsifying the content/data. Hacker injects malicious piece of code to replace the original content.

(8) Cross-Site Scripting

This type of attack is possible when a website containing input fields accepts scripts as well and leads to the phishing attack. The script gets stored in the database and executed every time the page is attacked. For example, <script>alert(message)</script>. Message could be a cookie also. When any user visits the page and application searches for username or password, the script will be executed.

(9) Denial of Service Attack

This kind of attack prevents normal users to access a website. The attacker attempts to access database server and performs SQL injections on it so that database becomes inaccessible. The attacker may also try to gain access as normal user with wrong password. After few attempts the user is locked out. The attacker may also gain access to web server and sends specially crafted requests so that web server is crashed.

(10) SQL Injection

It is an attack where any malicious script/code is inserted into an instance of SQL server/database for execution which eventually will try to fetch any database information.

(11) Poor Session Management

If an attacker can predict a unique value that identifies a particular user or session (session hijacking) he can use it to enter in the system as a genuine user. This problem also occurs when logout activity just redirects

the user to home page without termination of current session. The old session IDs can be used for authorization.

(12) Application Configuration Settings

Certain configuration settings exist in a web application by default such as debug settings, permissions, hardcoded user names, passwords and admin account information. An attacker may use this information to obtain unauthorized access.

(13) Cross site request forgery [6,7]:

It is a vulnerability which includes exploitation of a website by transmitting unauthorized commands from a user that a website trusts. Thus it exploits the trust of a website which it has on its user browser.

(14) Xml injection [1]:

It is an attack where an attacker tries to inject xml code with aim of modifying the xml structure thus violating the integrity of the application.

(15) Malicious file execution [3]:

Web applications are often vulnerable to malicious file execution and it usually occurs the code execution occurs from a non trusted source.

(16) Cookie cloning [11]:

Where an attacker after cloning the user/browser cookies tries to change the user files or data or may even harm the injected code.

(17) Xpath injection [3]:

It occurs when ever a website uses the information provided by the user so as to construct an xml query for xml data.

(18) Cookie sniffing [11]:

It is a session hijacking vulnerability with the aim of intercepting the unencrypted cookies from web applications.

(19) Cookie manipulation [5]:

Here an attacker tries to manipulate or change the content of the cookies and thus can cause any harm to the data or he may even change the data.

(20) Sidejacking [11]:

It is a hacking vulnerability where an attacker tries to capture all the cookies and may even get access to the user mailboxes etc.

(21) Social vulnerability (hacking), session hijacking [4, 5, 10, 11]:

It is a popular hijacking mechanism where an attacker gains unauthorized access to the information. xviii. Mis-configuration [24]: in appropriate or inadequate configuration of the web application may even lead to the security breaches.

(22) Absence of secure network infrastructure [9]:

Absence of any intrusion detection or protection system or failover systems etc may even lead to violation of the security breaches.

(23) Off the shelf components [9, 11]:

These components are purchased from third party vendors so there occurs a suspicion about their security aspect.

(24) Firewall intrusion detection system [8, 9,10]:

A firewall builds a secured wall between the outside/ external network and the internal network which is kept to be trusted.

(25) Path traversal [3]:

It is a vulnerability where malicious untrusted input causes non desirable changes to the path.

(26) Command injection [3]:

It is the injection of any input value which is usually embedded into the command to be executed.

(27) Parameter manipulation [5]:

It is similar to XSS where an invader inserts malicious code/script into the web application.

(28) LDAP injection [3]:

It is similar to SQL and Xpath injection where queries are being targeted to LDAP server.

(29) Bad code or fault in implementation [2]:

Iimproper coding or fault in the implementation of the web application may even lead to the violation of the security of the web application.

(30) Clickjacking [6]:

It is an attack where a user's click may be hijacked so that the user would be directed to some other link which may contain some malicious code.

(31) Content injection [8, 6]:

It is a vulnerability where an attacker loads some static content that may be some false content into the web page.

(32) File injection [8]:

It refers to the inclusion of any unintended file and is a typical vulnerability often found in web applications. Example: remote file inclusion.

Challenges faced by security testing of web applications

One of the concerns of security testing of web applications is the development of automated tools for testing the security of web applications [3]. Increase in the usage of Rich Internet Applications (RIAs) also poses a challenge for security testing of web application. This is due to the fact that the crawling techniques which are used for exploration of the web applications used for earlier web applications do not fulfil the requirements for RIAs [3]. RIAs being more users friendly and responsive due to the usage of AJAX technologies. Another challenge could be the usage of unintended invalid inputs which may result in security attacks [1]. And these security breaches may lead to extensive damage to the integrity of the data. While working the mutants, one should be sincere enough to incorporate them as injecting && (and) instead of || (or) or any such other modification may lead to fault injection which could result in a security vulnerability as vulnerabilities do not take semantics into consideration [1]. This may even pose a challenge to the security testing of any such web application. Usage of insecure cryptographic storage may even pose a challenge to the web application security testing [1]. Security testing of web applications may face repudiation attacks where any receiver is not able to prove that the data received came from a specific sender or from any other unintended source [1]. Also the web development languages which we use may lack in enforcing the security policy which may even violate the integrity and confidentiality of the web application [11]. This may even pose a security risk. At times it is also possible that an invader is able to launder more information than intended, in such a case again this may lead to the set back to the integrity of the data which could be another challenge for a security tester.

Conclusion

In this paper we have describes various kinds of security vulnerabilities that may exist in a website if proper consideration is not taken during development. A website developer must employ all possible measures to combat any known threats during the whole development cycle of a website from its design, implementation to testing. If any security loop hole remains undetected hackers can use it for exploiting the system.

References

1. An Approach Dedicated for Web Service Security Testing, S´ebastienSalva, Patrice Laurencot and IssamRabhi. 2010 Fifth International Conference on Software Engineering Advances.
2. Security Testing of Web Applications: a Search Based Approach for Cross-Site Scripting Vulnerabilities, Andrea Avancini, Mariano Ceccato , 2011- 11th IEEE International Working Conference on Source Code Analysis and Manipulation.
3. SUPPORTING SECURITY TESTERS IN DISCOVERING INJECTION FLAWS. Sven T¨urpe, Andreas Poller, Jan Trukenm¨uller, J¨urgenRepp and Christian Bornmann, Fraunhofer-Institute for Secure Information Technology SIT, Rheinstrasse 75,64295 Darmstadt, Germany, 2008 IEEE,Testing: Academic & Industrial Conference - Practice and Research Techniques.
4. A Database Security Testing Scheme of Web Application, Yang Haixia ,Business College of Shanxi University, Nan Zhihong, Scholl of Information Management,Shanxi University of Finance &Economics,china. Proceedings of 2009 4th International Conference on Computer Science & Education.

5. Mapping software faults with web security vulnerabilities. Jose Fonseca and Marco Vieira. International conference on Dependable Systems &Networks : Anchorage, Alaska,june 2008 IEEE.

6. D-WAV: A Web Application Vulnerabilities Detection Tool Using Characteristics of Web Forms. Lijiu Zhang, Qing Gu, Shushen Peng, Xiang Chen, Haigang Zhao, Daoxu Chen State Key Laboratory of Novel Software Technology, Department of Computer Science and Technology, Nanjing University. 2010 Fifth International Conference on Software Engineering Advances.

7. Enhancing web page security with security style sheets Terri Oda and Anil Somayaji (2011) IEEE.

8. Security Testing of Web Applications: a Search Based Approach for Cross-Site Scripting Vulnerabilities, Andrea Avancini, Mariano Ceccato , 2011- 11th IEEE International Working Conference on Source Code Analysis and Manipulation.

9. Assessing and Comparing Security of Web Servers. Naaliel Mendes, AfonsoAraújoNeto, JoãoDurães, Marco Vieira, and Henrique Madeira CISUC, University of Coimbra. 2008 14th IEEE Pacific Rim International Symposium on Dependable Computing.

10. Firewall Security: Policies, Testing and Performance Evaluation. Michael R. Lyu and Lorrien K. Y. Lau. Department of computer science and engineering. The Chinese University of Hong kong, Shatin, HK. 2000 IEEE.

11. Top 10 Free Web-Mail Security Test Using Session Hijacking Preecha Noiumkar, Thawatchai Chomsiri, Mahasarakham University, Mahasarakham, Thailand. Third 2008 International Conference on Convergence and Hybrid Information Technology. Development of Security Engineering Curricula at US Universities.Mary Lynn Garcia, Sandia National Laboratories.1998 IEEE.

Security Analytics: Challenges and Future Directions

Ganga Sharma*
Bhawana Tyagi**

Abstract

The frequency and type of cyber attacks are increasing day by day. However, well-known cyber security solutions are not able to cope with the increasing volume of data that is generated for providing security solutions. Therefore, current trend in research on cyber security is to apply Big Data Analytics (BDA) techniques to cyber security. This field, called security analytics (SA), can help network managers in the monitoring and surveillance of real-time network streams and real-time detection of malicious and/or suspicious patterns. Researchers believe that an SA system can assist in enhancing all traditional security mechanisms. Nonetheless, there are certain issues related to incorporating big data analytics to cyber security. This paper presents an analysis on the issues and challenges faced by Security Analytics, and further provides future directions in the field.

Keywords: cyber-security, big data, security analytics, big data analytics

I. Introduction

Big data analytics (BDA) is the large scale analysis and processing of information [1,14]. It uses advanced analytic and parallel techniques to process very large and diverse records that include different types of contents. BDA tools allow getting enormous benefits and valuable insights by dealing with any massive volume of mixed unstructured, semi-structured and structured data that is fast changing and difficult to process using conventional database techniques.

In recent years, BDA has gained popularity in the security community as it promises efficient processing and analysis of security-related data at large scale [3]. Corporate research is now focusing on Security Analytics, i.e., the application of Big Data Analytics techniques to cyber-security. Analytics can assist network managers particularly in the monitoring and surveillance of real-time network streams and real-time detection of both malicious and suspicious (outlying) patterns. Over the past ten years, enterprise security has gone incrementally more difficult as new and unanticipated threats/attacks surface. The existing security infrastructures collect, process and analyze terabytes of security data on monthly basis. This data is too large to be handled efficiently by the existing data storage architectures, algorithms, and query mechanisms. Therefore the application of Big data analytics (BDA) to security is the need of the hour.

This paper provides an overview of how big data analytics can help in enhancing the traditional cyber security mechanisms and thus provide a means for better security analysis. Rest of the paper is organized as follows: section 2 gives a brief overview of literature work, section 3 describes the basic BDA process, section 4 and 5 respectively provide the challenges and fututre directions in security analytics while section 6 concludes the paper.

II. Literature Review

Security analytics is a new technology and concept, therefore much research has not been conducted in this area. However, there are some significant contributions by several authors in this field. For e.g., Mahmood and Afzal[14] have presented a comprehensive survey on the state of the art of Security Analytics, i.e., its description, technology, trends, and tools. Gahi et al [1] highlight the benefits of Big Data Analytics and then provide a brief overview of challenges of security and privacy in big data environments itself. Further, they present some available protection techniques and propose some

Ganga Sharma*
Assistant Professor (IT Dept)
IITM Janakpuri

Bhawana Tyagi**
Assistant Professor (IT Dept)
IITM

possible tracks that enable security and privacy in a malicious big data context. Cybenko and Landwehr[7] stud-ied historical data from a variety of cyber- and national security domains in United state such as computer vulner-ability databases, offensive and defense, co-evolution of wormbots such as Conficker etc. They claim that security analytics can provide the ultimate solution for cyber-security. Cardenas et al[9]provide details of how the security analytics landscape is changing with the introduction and widespread use of new tools to leverage large quantities of structured and unstructured data. It also outlines some of the fundamental differences between security analytics and traditional analytic. Camargo et al[10] research on the use of big data analytics for security and analyze the perception of people for security. They found that big data can indeed provide a long-term solution for citizen's security, in particular cyber security.

III. Big Data And The Basic Bda Process

Big data is data whose complexity hinders it from being managed, queried and analyzed efficiently by the existing database architectures[4]. The "complexity" of big data is defined through 4V's: 1) volume – referring to terabytes, petabytes, or even exabytes (10006 bytes) of stored information, 2) variety – referring to the co-existence of unstructured, semi-structured and structured data, and 3) velocity – referring to the rapid pace at which big data is being generated and 4) veracity- to stress the importance of maintaining quality data within an organization.

The domain of Big Data Analytics (BDA) is concerned with the extraction of value from big data, i.e., insights which are nontrivial and previously unknown, implicit and potentially useful. These insights have a direct impact on deciding or manipulating the current business strategy [14]. The assumption is that patterns of usage, occurrences or behaviors exist in big data. BDA attempts to fit mathematical models on these patterns through different data mining techniques such as Predictive Analytics, Cluster Analysis, Association Rule Mining, and Prescriptive Analytics [13]. Insights from these techniques are typically represented on interactive dashboards and help corporations maintain the competitive edge, increase profits, and enhance their CRM.

Fig. 1 shows the basic stages of BDA process[14] . Initially, data to be analyzed is selected from real-time streams of big data and is pre-processed (i.e. cleaned). This is called ETL (Extract Transform Load). It can take up to 60% of the effort of BDA, e.g., catering for inconsistent, incomplete andmissing values, normalizing, discretizing and reducing data, ensuring statistical quality of data through boxplots, cluster analysis, normality testing etc., and understanding data through descriptive statistics (correlations, hypothesis testing, histograms etc.). Once data is cleaned, it is stored in BDA databases (cloud, mobile, network servers etc.) and analyzed with analytics. The results are then shown in interactive dashboards using computer visualization.

IV. Challenges in Security Analytics

The big data is a recent technology and has been widely adopted to provide solutions to organsational decision making[11]. One of the most important area to benefit from the advancements in big data analytics is cyber security. This area is now being stated as security analytics. An important goal for security analytics is to enable organisations to identify unknown indicators of attack, and uncover things like when compromised credentials are being used to bypass defenses[2]. However, handling unstructured data and combing it with structured data to arrive at an accurate assessment is one of the big challenges in security analytics.

In the past, information security was really based on event correlation designed for monitoring and detecting known attack patterns[9]. This model alone is no longer adequate as multidimensional cyber-attacks are dynamic and can use different tactics and techniques to find their way into and out of an organization. In addition, the traditional set of security devices is designed and optimized to look for particular aspects of attacks: a network perspective, an attack perspective, a malware perspective, a host perspective, a web traffic perspective, etc[12]. These different technologies see isolated aspects of an attack and lack the bigger picture.

1. Cyber-attacks are extremely difficult to distinguish or investigate, because until all the event data is combined, it's extremely hard to determine what an attacker is trying to accomplish[6,8].

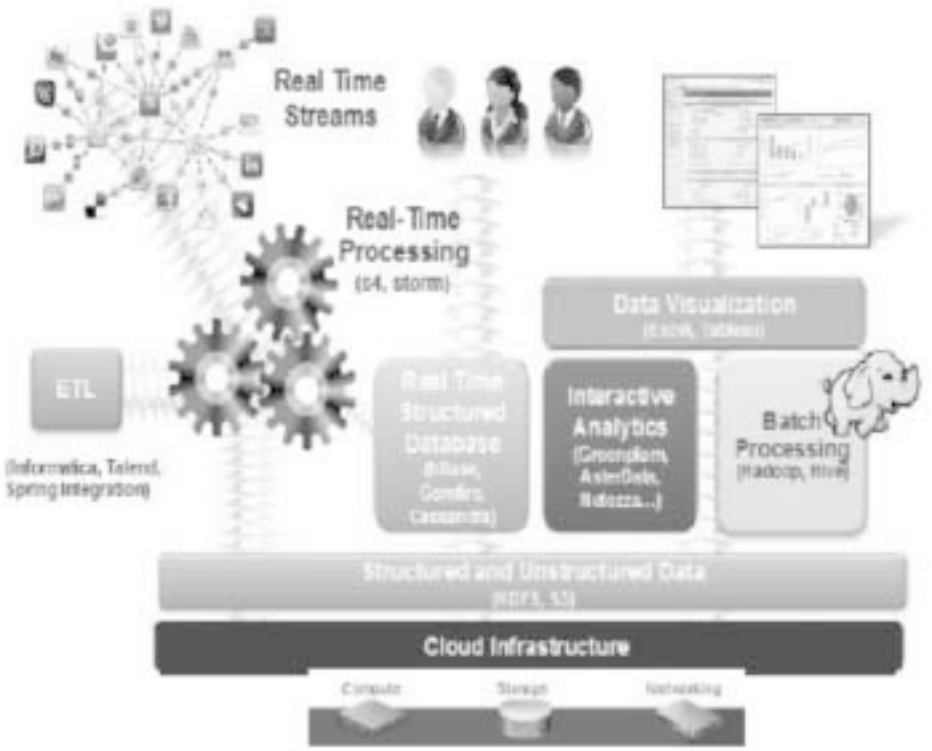


Fig1. Basic BDA process[14]

Addressing new types of cyber-threats requires a commitment to data collection and processing as well as much greater diligence on security data analytics.

2. The main idea behind big data is to extract useful insights by performing specific computations. However, it is important to secure and protect these computations to avoid any risk or attempt to change or skew the extracted results. It is also important to protect the systems from any attempt to spy on the nature or the number of performed computations.

3. In an open context, large volume of content collected through big data is not always a good metric for the quality of extracted results. Therefore, it may not always be possible to achieve good threat detection and prevention.

4. Since cyber-attacks can be multidimensional can happen over long periods of time, historical analysis must also be incorporated so that analysts can perform root cause analysis and attack scoping to determine the breadth of a compromise or data breach.

5. While original data formats should be preserved, security analysts must also have the ability to tag, index, enrich, and query any data element or group of data elements together to get a broader perspective for threat detection/response. Otherwise, security data will remain a black hole if it can't be easily queried and understood by security professionals .

6. Systems must provide a simple interface and search-based access to broaden and simplify access to data. This will empower security analysts to investigate threats and gain valuable experience. Systems should also allow for straightforward ways to create dashboards and reports to streamline security operations.

V. Future Directions

It is no longer a matter of if, but when, attackers will break into your network. They'll use zero-day attacks, stolen access credentials, infected mobile devices, a vulnerable business partner, or other tactics. Security success is not just about keeping threats out of your network. Instead it's about quickly responding to and

thwarting an attack when it happens[4,5]. According to a very reputed organization providing security solutions "Organizations are failing at early breach detection, with more than 92 percent of breaches undetected by the breached organization." It is clear that we need to play a far more active role in protecting our organizations[8]. We need to constantly monitor what is going on within our infrastructure and have an established, cyclical means of responding before attacks wreak havoc on our networks and reputations. Therefore, some of the primary requirements for the security analytics solution are:

1. Secure sensitive data entering Big database systems and then provide control access to Protected data by monitoring which applications and which users gets access to which original data.
2. Protection of sensitive data that maintains usable, realistic values for accurate analytics and modeling on data in its encrypted form.
3. Assure global regulatory compliance. Securely capture, analyze and store data from global sources, and ensure compliance with international data security, residency and privacy regulations. Address compliance comprehensively, not system-by-system.
4. Optimize performance and scalability.
5. Integrate data security, with quick implementation and an efficient, low-maintenance solution that should scale up. Leverage IT investments by integrating with the existing IT environment and extending current controls and processes into Big Databases.
6. As far as possible provide block layer encryption, which will improve security but also enable big data clusters to scale and perform[7,8].
7. Leverage security tools or third-party products. Tools may include SSL/TLS for secure communication, Kerberos for node authentication, transparent encryption for data-at-rest[13].

VI. Conclusion

Security analytics is the new technical foundation of an informed, reliable detection and response strategy for cyber attacks. Mature security organizations recognize this and are leading with building their security analytics capabilities today. A security analytics system combines and integrates the traditional ways of cyber threat detection to provide security analysts a platform with both enterprise-scale detection and investigative capabilities. It will not only help identify events that are happening now, but will also assess the state of security within the enterprise in order to predict what may occur in the future and enable more proactive security decisions.

References

1. Gahi, Y., Guennoun, M., & Mouftah, H. T. (2016, June). Big Data Analytics: Security and privacy challenges. In *Computers and Communication (ISCC), 2016 IEEE Symposium on* (pp. 952-957). IEEE.
2. Verma, R., Kantarcioglu, M., Marchette, D., Leiss, E., & Solorio, T. (2015). Security analytics: essential data analytics knowledge for cybersecurity professionals and students. *IEEE Security & Privacy*, *13*(6), 60-65.
3. Oltsik, J. (2013). The Big Data Security Analytics Era Is Here. White Paper, Retrieved *from https://www.emc.com/collateral/analyst-reports/security-analytics-esg-ar.pdf* on on 30th December, 2016
4. Shackleford D. (2013). SANS Security Analytics Survey, WhitePaper, SANS Institute InfoSec Reading Room. Downloaded on 3^{0th} December, 2016.
5. Gawron, M., Cheng, F., & Meinel, C. (2015, August). Automatic detection of vulnerabilities for advanced security analytics. In *Network Operations and Management Symposium (APNOMS), 2015 17th Asia-Pacific* (pp. 471-474). IEEE.
6. Gantsou, D. (2015, August). On the use of security analytics for attack detection in vehicular ad hoc networks. In *Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC), 2015 International Conference on* (pp. 1-6). IEEE.

7. Cybenko, G., & Landwehr, C. E. (2012). Security analytics and measurements. *IEEE Security & Privacy*, *10*(3), 5-8.

8. Cheng, F., Azodi, A., Jaeger, D., & Meinel, C. (2013, December). Multi-core Supported High Performance Security Analytics. In *Dependable, Autonomic and Secure Computing (DASC), 2013 IEEE 11th International Conference on* (pp. 621-626). IEEE.

9. Cardenas, A. A., Manadhata, P. K., & Rajan, S. P. (2013). Big data analytics for security. *IEEE Security & Privacy*, *11*(6), 74-76.

10. Camargo, J. E., Torres, C. A., Martínez, O. H., & Gómez, F. A. (2016, September). A big data analytics system to analyze citizens' perception of security. In *Smart Cities Conference (ISC2), 2016 IEEE International* (pp. 1-5). IEEE.

11. Alsuhibany, S. A. (2016, November). A space-and-time efficient technique for big data security analytics. In *Information Technology (Big Data Analysis)(KACSTIT), Saudi International Conference on* (pp. 1-6). IEEE.

12. Rao, S., Suma, S. N., & Sunitha, M. (2015, May). Security Solutions for Big Data Analytics in Healthcare. In *Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on* (pp. 510-514). IEEE.

13. Marchetti, M., Pierazzi, F., Guido, A., & Colajanni, M. (2016, May). Countering Advanced Persistent Threats through security intelligence and big data analytics. In *Cyber Conflict (CyCon), 2016 8th International Conference on* (pp. 243-261). IEEE.

14. T. Mahmood and U. Afzal, "Security Analytics: Big Data Analytics for cyber-security: A review of trends, techniques and tools," 2nd National Conference on Information Assurance (NCIA), 2013

A Survey of Multicast Routing Protocols in MANET

Ganesh Kumar Wadhwani*
Neeraj Mishra**

Abstract

Multicasting is a technique in which a sender's message is forwarded to a group of receivers. Conventional wired multicast routing protocols do not perform well in mobile ad hoc wireless network (MANET) because of the dynamic nature of the network topology. Apart from mobility aspect there is bandwidth restriction also which must be addressed by the multicasting protocol for the MANET. In this paper, we give a survey of classification of multicast routing protocol and associated protocols. In the end, a comparison is also made among different classes of multicast routing.

Keywords: Multicast routing, mobile ad hoc network, tree based protocol, mesh based protocol, source-initiated multicast, receiver initiated multicast, soft state, hard state

I. Introduction

MANET is a collection of autonomous mobile nodes communicating with each other without a fixed infrastructure. MANET find applications in areas where setting up and maintaining a communication infrastructure may be difficult or costly like emergency search and rescue operation, law enforcement and warfare situations.

Multicasting is a technique for data routing in networks that allows the same message is forwarded to a group of destinations simultaneously. Multicasting is intended for group oriented computing like audio/ video conferencing, collaborative works, etc. Multicasting is an essential technology to efficiently support one to many or many to many applications. Multicast routing has attracted a lot of attention in the past decade, due to it allows a source to send information to multiple destinations concurrently. Multicasting is the transmission of packets to a group of zero or more hosts called multicast group that is identified by a single destination address. A multicast group is a set of network clients and servers interested in sharing a specific set of data. A typical example of multicast groups is a commander and his soldiers in a battlefield. There are other examples in which multicast groups need to be established. Typically, the membership of a host group is dynamic: that is, the hosts may join and leave groups at any time. There is no restriction on the location or number of members in a host group. A host may be a member of more than one group at a time. A host does not have to be a member of a group to send packets to it. A multicast protocol has the objective of connecting members of the multicast group in an optimal way, by reducing the amount of bandwidth necessary but also considering other issues such as communication delays and reliability [1].

In MANET Multicast routing plays an important role in ad hoc wireless networks to provide communication among nodes which are highly dynamic in terms of their location. It is advantageous to use multicast rather than multiple unicast especially in the ad hoc environment where bandwidth is an issue. Conventional wired network multicast routing protocols such as DVMRP, MOSP, CBT and PIM don't perform well in MANET because of the dynamic nature of the network topology. The dynamically changing topology, coupled with relatively low bandwidth and less reliable wireless links, causes long convergence times and may give rise to formation of transient routing loops that rapidly consume the already limited bandwidth.

II. Multicast Routing Classification

One of the most popular methods to classify multicast routing protocols for MANETs is based on how distribution paths among group members are

Ganesh Kumar Wadhwani*
Computer Science,
IITM

Neeraj Mishra**
Computer Science,
IITM

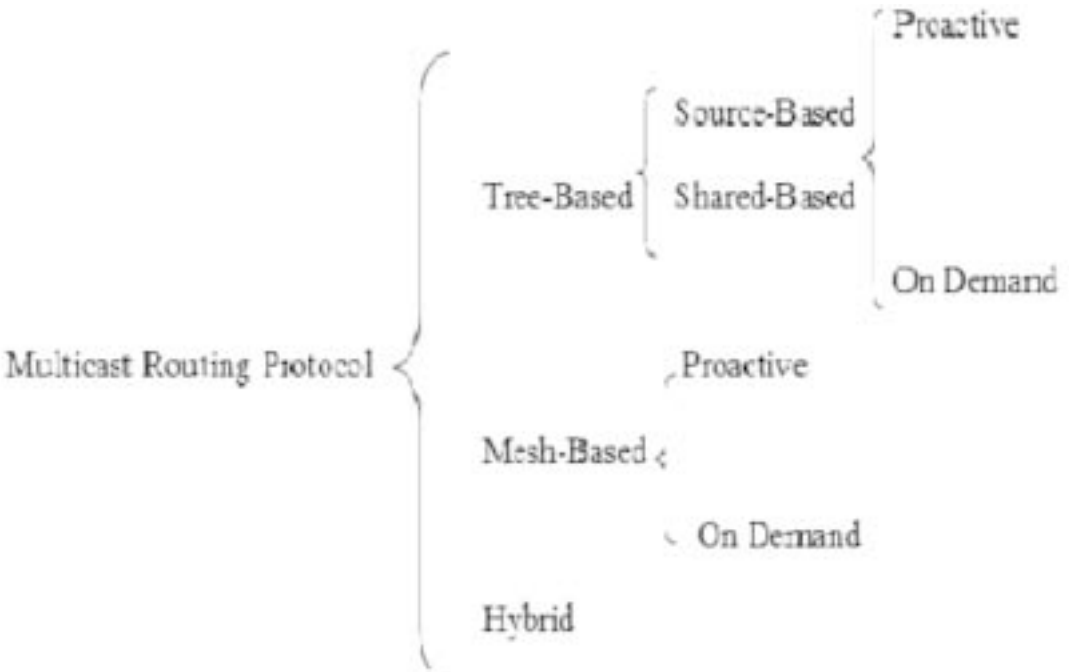


Figure I: Classification of Multicast Routing Protocols

constructed (the underlying routing structure). According to this method, existing multicast routing approaches for MANETs can be divided into tree based multicast protocols, mesh based multicast protocols and hybrid multicast protocols.

In tree-based protocols, there is only one path between a source-receiver pair. It is efficient but main drawback of these protocols is that they are not robust enough to operate in highly mobile environment. [2]

Depending on the number of trees per multicast group, tree based multicast can be further classified as source based multicast tree and group shared multicast tree. In source tree based multicast protocols, the tree is rooted at the source, whereas in shared-tree-based multicast protocols, a single tree is shared by all the sources within the multicast group and is rooted at a node referred to as the core node. The source tree based multicast perform better than the shared tree based protocol at heavy load because of efficient traffic distribution, But the latter type of protocol are more scalable. The main problem in a shared tree based multicast protocol is that it heavily depends on the core node, and hence, a single point failure at the core node affects the performance of the multicast protocol.

Some of the tree based multicast routing protocols are, bandwidth efficient multicast routing protocol (BEMRP) [3], multicast zone routing protocol (MZRP) [4], multicast core extraction distributed ad hoc routing protocol (MCEDAR) [5], differential destination based multicast protocol (DDM) [6], ad hoc multicast routing protocol utilizing increasing id numbers (AMRIS) [7], and ad hoc multicast routing protocol (AMRoute) [8].

Bandwidth-Efficient Multicast Routing Protocol (BEMRP)

It tries to find the nearest forwarding nodes, rather than the shortest path between source and receiver. Hence, it reduces the number of data packet transmissions. To maintain the multicast tree, it uses the hard state approach in which control packets are transmitted (to maintain the routes) only when a link breaks, resulting in lower control overhead, but at the cost of a low packet delivery ration. In BEMRP, the receiver initiates the multicast tree construction. When a receiver wants to join the group, it initiates flooding of Join control packets the existing members of the multicast tree, on receiving these packets, respond with Reply packets. When many such Reply packet reach

the requesting node, it chooses one of them and sends a Reserve packet on the path taken by the chosen Reply packet.

Multicast Operation of the Ad-hoc On-Demand Distance Vector Routing Protocol (MAODV)

MAODV [9] is a shared-tree-based protocol that is an extension of AODV [10] to support multicast routing. With the unicast route information of AODV, MAODV constructs the shared tree more efficiently and has low control overhead. In MAODV, the group leader is the first node joining the group and announces its existence by Group Hello message flooding. An interested node P sends a join message toward the group leader. Any tree node of the group sends a reply message back to P. P only answers an MACT message to the reply message with minimum hop count to the originator. Then a new branch to the shared tree is set up.

Ad Hoc Multicast Routing Protocol Utilizing Increasing Id-numbers (AMRIS)

AMRIS [12] is an on-demand shared-tree-based protocol which dynamically assigns every node in a multicast session an id- number. The multicast tree is rooted at a special node called Sid and the id- numbers of surrounding nodes increase in numerical value as they radiate from the Sid. These id-numbers help nodes know which neighbours are closer to the Sid and this reduces the cost to repair link failures.

Sid initially floods a NEW-SESSION message associated with its id -number through the network. Each node receiving the NEW- SESSION message generates its own id- number by computing a value that is larger than and not consecutive to the received one. Then the node places its own id-number and routing metrics before rebroadcasting the message. Each node sends a periodic beacon for exchanging information (like its own id- number) with its neighbours. When a new node P wants to join the session, it sends a join message to one of its potential parent nodes (i.e., those neighbouring nodes having smaller id-numbers) Q. If Q is a tree node, it replies a message to P; otherwise, Q forwards this join message to one of its own potential parent nodes. This process is repeated until a tree node is found (see Figure. 2). If no reply message returns to P, a localized broadcast is used.

Adaptive Demand-Driven Multicast Routing (ADMR)

ADMR [13] is an on-demand sender-tree-based protocol which adapts its behaviour based on the application data sending pattern. It does not require periodic floods of control packets, periodic neighbour sensing, or periodic routing table exchanges. The application layer behaviour allows efficient detection of link breaks and expiration of routing state. ADMR temporarily switches to the flooding of each data packet if high mobility is detected.

A multicast tree is created when a group sender originates a multicast packet for the first time. Interested nodes reply to the sender's packet to join the group. Each multicast packet includes inter -packet time which is the average packet arrival time from the sender's application layer. The inter-packet time lets tree nodes predict when the next multicast packet will arrive and hence no periodic control messages are required for tree maintenance. If the application layer does not originate new packets as expected, the routing layer of the sender will issue special keep-alive packets to maintain the multicast tree. The sender occasionally uses network floods of data packets for finding new members.

The Differential Destination Multicast Protocol (DDM)

DDM [14] is a sender-tree-based protocol that is designed for small group. DDM has no multicast routing structure. It encodes the addresses of group members in each packet header and transmits the packets using the underlying unicast routing protocol. If a node P is interested in a multicast session, it unicast a join message to the sender of the session. The sender adds P into its member list (ML) and unicasts an ACK message back to P. DDM has two operation modes: stateless mode and soft-state mode. In stateless mode, the sender includes a list of all receivers' addresses in each multicast packet. According to the address list and the unicast routing table, each node receiving the packet determines the next hop for forwarding the

packet to some receivers, and will partition the address list to distinct parts for each chosen next hop.

In order to reduce the packet size, DDM can operate in soft-state mode. Each node in soft-state mode records the set of receivers for which it has been the forwarder. Each multicast packet only describes the change of the address list since the last forwarding by a special DDM block in the packet header. For instance, if R4 moves to another place and loses connection to R3, the DDM block in the packet header describes that R4 is removed. Then B knows that it only has to forward the packet to R3.

Multicast Core-Extraction Distributed Ad Hoc Routing (MCEDAR)

MCEDAR is a multicast extension to the CEDAR architecture which provides the robustness of mesh structures and the efficiency of tree structures. MCEDAR uses a mesh as the underlying infrastructure, but the data forwarding occurs only on a sender-rooted tree. MCEDAR is particularly suitable for situations where multiple groups coexist in a MANET.

At first, MCEDAR partitions the network into disjoint clusters. Each node exchanges a special beacon with its one hop neighbors to decide that it becomes a dominator or chooses a neighbor as its dominator. A dominator and those neighbors that have chosen it as a dominator form a cluster. A dominator then becomes a core node and issues a message to nearby core nodes for building virtual links between them. All the core nodes form a core graph.

When a node intends to join a group, it delegates its dominating core node P to join the appropriate mgraph instead of itself. An mgraph is a subgraph of the core graph and is composed of those core nodes belonging to the same group. P joins the mgraph by broadcasting a join message which contains a joinID. Only those members with smaller joinIDs reply an ACK message to P (see Figure. 6). Other nodes receiving the join message forward it to their nearby core nodes. An intermediate node Q only accepts at most R ACK messages where R is a robustness factor. Q then puts the nodes from which it receives the ACK message into its parent set and the nodes to which it forwards the ACK message into its child set.

When a node has less than R/2 parents, it periodically issues new join messages to get more parents. When a data packet arrives at an mgraph member, the member only forwards the packet to those nearby member core nodes that it knows.

Mesh-based protocols may have more than one path between a source-receiver pair thereby provide redundant routes for maintaining connectivity to group members. Because of the availability of multiple paths between the source and receiver mesh based protocols are more robust compared to tree based.[2]

On-Demand Multicast Routing Protocol (ODMRP)

ODMRP provides richer connectivity among group members and builds a mesh for providing a high data delivery ratio even at high mobility. It introduces a "forwarding group" concept to construct the mesh and a mobility prediction scheme to refresh the mesh only necessarily.

The first sender floods a join message with data payload piggybacked. The join message is periodically flooded to the entire network to refresh the membership information and update the multicast paths. An interested node will respond to the join message. Note that the multicast paths built by this sender are shared with other senders. In other words, the forwarding node will forward the multicast packets from not only this sender but other senders in the same group (see Figure. 7).

Due to the high overhead incurred by flooding of join messages, a mobility prediction scheme is proposed to find the most stable path between a sender-receiver pair. The purpose is to flood join messages only when the paths indeed have to be refreshed. A formula based on the information provided by GPS (Global Positioning System) is used to predict the link expiration time between two connected nodes. A receiver sends the reply message back to the sender via the path having the maximum link expiration time.

A Dynamic Core Based Multicast Routing Protocol (DCMP)

DCMP aims at mitigating the high control overhead problem in ODMRP. DCMP dynamically classifies

the senders into different categories and only a portion of senders need issue control messages. In DCMP, senders are classified into three categories: active senders, core senders, and passive senders. Active senders flood join messages at regular intervals. Core senders are those active senders which also act as the core node for one or more passive senders. A passive sender does not flood join messages, but depends on a nearby core sender to forward its data packets. The mesh is created and refreshed by the join messages issued by active senders and core senders.

All senders are initially active senders. When a sender S has packets to send, it floods a join message. Upon receiving this message, an active sender P delegates S to be its core node if P is close to S and has smaller ID than S. Afterwards, the multicast packets sent by S will be forwarded to P first and P relays them through the mesh.

Adaptive Core Multicast Routing Protocol (ACMRP)

ACMRP presents an adaptive core mechanism in which the core node adapts to the network and group status. In general mesh-based protocols, the mesh provides too rich connectivity and results in high delivery cost. Hence, ACMRP forces only one core node to take responsibility of the mesh creation and maintenance in a group. The adaptive core mechanism also handles any core failure caused by link failures, node failures, or network partitions.

A new core node of a group emerges when the first sender has multicast packets to send. The core node floods join messages and each node stores this message into its local cache. Interested members reply a JREP message to the core node. Forwarding nodes are those nodes who have received a JREP message. If a sender only desires to send packets (it's not interested in packets from other senders), it sends an EJREP message back to the core node. Those nodes receiving this EJREP message only forward data packets from this sender. If a new sender wishes to send a packet but has not connected to the mesh, it encapsulates the packet toward the core node. The first forwarding node strips the encapsulated packet and sends the original packet through the mesh.

ACMRP proposes a novel mechanism to re-elect a new core node which is located nearby all members regularly. The core node periodically floods a query message with TTL set to acquire the group membership information and lifetime of its neighboring nodes. The core node will select the node that has the minimum total hop count of routes toward group members among neighboring nodes as the new core node.

Multicast Protocol for Ad Hoc Networks with Swarm Intelligence (MANSI)

MANSI relies on only one core node to build and maintain the mesh and applies swarm intelligence to tackle metrics like load balancing and energy conservation. Swarm intelligence refers to complex behaviors that arise from very simple individual behaviors and interactions. Although each individual has little intelligence and simply follows basic rules using local information obtained from the environment, globally optimized behaviors emerge when they work collectively as a group. MANSI utilizes this characteristic to lower the total cost in the multicast session.

The sender that first starts sending data takes the role of the core node and informs all nodes in the network of its existence. Reply messages transmitted by interested nodes construct the mesh. Each forwarding node is associated with a height which is identical to the highest ID of the members that use it to connect to the core node. After the mesh creation, MANSI adopts the swarm intelligence metaphor to allow nodes to learn better connections that yield lower forwarding cost. Each member P except the core node periodically deploys a small packet, called FORWARD ANT, which opportunistically explores better paths toward the core.

A FORWARD ANT stops and turns into a BACKWARD ANT when it encounters a forwarding node whose height is higher than the ID of P. A BACKWARD ANT will travel back to P via the reverse path. When the BACKWARD ANT arrives at each intermediate node, it estimates the cost of having the current node to join the forwarding set via the forwarding node it previously found. The estimated

cost, as well as a pheromone amount, is updated on the node's local data structure. The pheromone amounts are then used by subsequent FORWARD ANTs that arrive at this node to make a decision which node they will travel to next.

MANSI also incorporates a mobility-adaptive mechanism. Each node keeps track of the normalized link failure frequency (nlff) which reflects the dynamic condition of the surrounding area. If the nlff exceeds the threshold, the node will add another entry for the second best next hop into its join messages. Then the additional path to the core node increases the reliability of MANSI.

Neighbor Supporting Ad Hoc Multicast Routing Protocol (NSMP)

NSMP utilizes the node locality concept to lower the overhead of mesh maintenance. For initial path establishment or network partition repair, NSMP occasionally floods control messages through the network. For routine path maintenance, NSMP uses local path recovery which is restricted only to mesh nodes and neighbor nodes for a group.

The initial mesh creation is the same with that in MANSI. Those nodes (except mesh nodes) that detect reply messages become neighbor nodes, and neighbor nodes do not forward multicast packets. After the mesh creation phase (see Figure. 11), all senders transmit LOCAL_REQ messages to maintain the mesh at regular interval. Only mesh nodes and neighbor nodes forward the LOCAL_REQ messages. In order to balance the routing efficiency and path robustness, a receiver receiving several LOCAL_REQ messages replies a message to the sender via the path with largest weighted path length.

Since only mesh nodes and neighbor nodes accept LOCAL_REQ messages, the network partition may not be repaired. Hence, a group leader is elected among senders and floods request messages through the network periodically. Network partition can be recovered by the flooding of request messages. When a node P wishes to join a group as a receiver, it waits for a LOCAL_REQ message. If no LOCAL_REQ message is received, P locally broadcasts a MEM_REQ message.

The Core-Assisted Mesh Protocol (CAMP)

CAMP is a receiver-initiated protocol. It assumes that an underlying unicast routing protocol provides correct distances to known destinations. CAMP establishes a mesh composed of shortest paths from senders to receivers. One or multiple core nodes can be defined for each mesh, and core nodes need not be part of the mesh, and nodes can join a group even if all associated core nodes are unreachable.

It is assumed that each node can reach at least one core node of the multicast group which it wants to join. If a joining node P has any neighbor that is a mesh node, then P simply tells its neighbors that it is a new member of the group. Otherwise, P selects its next hop to the nearest core node as the relay of the join message. Any mesh node receiving the join message transmits an ACK message back to P. Then P connects to the mesh. If none of the core nodes of the group is reachable, P broadcasts the join message using an expanded ring search.

For ensuring the shortest paths, each node periodically looks up its routing table to check whether the neighbor that relays the packet is on the shortest path to the sender. The number of packets coming from the reverse path for a sender indicates whether the node is on the shortest path. A special message will be issued to search a mesh node and the shortest path can be re-established. At last, to ensure that two or more meshes eventually merge, all active core nodes periodically send messages to each other and force nodes along the path that are not members to join the mesh.

III. Present Status of Multicast Routing Protocols

Multicasting is a mechanism in which a source can send the same communication to multiple destinations. In multicast routing a multicast tree is to be found out to a group of destination nodes along which the information will be disseminated to different nodes in parallel. Multicast routing is more efficient as compared to unicast because in this data is forwarded to many intended destination in one go rather than sending individually. At the same time it is not as expensive as broadcasting in which the data is flooded to all the nodes in the network. It is extremely suitable for a bandwidth constrained network like MANET.

Table I: Comparison of Multicast Routing Protocols

Multicast Protocols	Multicast Topology	Initiali-zation	Independent On Routing Protocol	Dependency On Specific Routing Protocol	Maintenance Approach	Loop Free	Flooding of Control Packets	Periodic Control Messaging
ABAM	Source-Tree	Source	Yes	No	Hard State	Yes	Yes	No
BEMRP	Source-Tree	Receiver	Yes	No	Hard State	Yes	Yes	No
DDM	Source-Tree	Receiver	No	No	Soft State	Yes	Yes	Yes
MCEDAR	Source-Tree Mesh	Source or Receiver	No	Yes (CEDAR)	Hard State	Yes	Yes	No
MZRP	Source-Tree	Source	Yes	No	Hard State	Yes	Yes	Yes
WBM	Source-Tree	Receiver	Yes	No	Hard State	Yes	Yes	No
PLBM	Source-Tree	Receiver	Yes	No	Hard State	Yes	No	Yes
MAODV	Source-Tree	Receiver	Yes	No	Hard State	Yes	Yes	Yes
ADAPTIVE SHARED	Combination of Shared And Source tree	Receiver	Yes	No	Soft State	Yes	Yes	Yes
AMRIS	Shared-Tree	Source	Yes	No	Hard State	Yes	Yes	Yes
AMROUTE	Shared Tree Mesh	Source or Receiver	No	No	Hard State	No	Yes	Yes
ODMRP	Mesh	Source	Yes	No	Soft State	Yes	Yes	Yes
DCMP	Mesh	Source	Yes	No	Soft State	Yes	Yes	Yes
FGMP	Mesh	Receiver	Yes	No	Soft State	Yes	Yes	Yes
CAMP	Mesh	Source or Receiver	No	No	Hard State	Yes	No	No
NSMP	Mesh	Source	Yes	No	Soft State	Yes	Yes	Yes

Traditional multicast routing protocols for wireless network cannot be implemented as it is in mobile ad-hoc network which poses new problems and challenges for the design of an efficient algorithm for MANET.

Mobile Ad Hoc network mainly showed the following aspects:

Dynamic network topology structure: In mobile Ad Hoc network, the node has a arbitrary mobility, the network topology structure may change at any time, and this change mode and speed are difficult to predict.

Limited bandwidth transmission: Mobile Ad Hoc network applies wireless transmission technology as its communication means, it has a lower capacity relative to the wireless channel. Furthermore, affected by multiple factors of noise jamming, signal interference and etc, the actually available effective bandwidth for mobile terminals will be much smaller than the maximum bandwidth value in theory.

The limitation of mobile terminal: although the user terminals in mobile Ad Hoc network have characteristics of smart and portable, they use the fugitive energy like battery as their power and with a CPU of lower performance and smaller memory, especially each of the host computers doubles the router, hence, there are quite high requirements on routing protocols.

Distributed control: there is no central control point in mobile Ad Hoc network, all the user terminals are equal,

and the network routing protocols always apply the distributed control mode, so it has stronger robustness and survivability than center-structured network.

Multihop communication: as the restriction of wireless transceiver on signal transmission range, the mobile Ad Hoc network is required to support multihop communication, which also brings problems of hidden terminals, exposed terminals, equity and etc.

Security: as the application of wireless signal channel, wired power, distributed control and etc, it is vulnerable to be threatened by security, such as eavesdropping, spoofing, service rejecting and etc attacking means.

Till date so many multicast routing protocols have been proposed and they have their own advantages and disadvantages to adapt to different environments. Therefore the hope for a standard multicast routing protocol which will be suitable for all network scenarios is highly unrealistic.

At the same time, it is very difficult to confirm multicast routing algorithms or protocols adapted to specific application fields for mobile Ad Hoc network, because the application of Ad Hoc network requires a combination and integration of the fixed network with the mobile environment. So there still needs a deeper research of multicast application in the mobile Ad Hoc network environment.

IV. Comparison Of Multicast Routing Protocols

The design goal of any multicast routing protocol to transmit information to all intended nodes in an optimum way and incur minimum redundancy in the process.

All the protocols try to deal with many problems like nodes mobility, looping, routing imperfections, whether on demand construction, routing update, the control over packet transmission methods (net-wide flooding broadcast or broadcast subjected to member nodes) etc.

In all tree based multicast routing protocols a unique path is obtained between any pair of nodes which saves the bandwidth required for initializing muticast tree as compared to bandwidth requirement of any other structure. The disadvantage of these protocols is the survivability of communication system in case of link/node failure. For example if any nodes moves out of transmission range dividing tree into two or more sub-tree which makes the communication difficult among all the nodes in the tree. In addition the overhead involved in maintaining the multicast tree is relatively larger as compared to other protocols.

Resource requirement for mesh based multicast routing protocols is much larger as compared to tree based protocols. It also suffers from routing loop problems and special measures are taken to avoid such problems which incur extra overhead on the overall communication system.

The biggest advantage of such protocols are their robustness, if one link fails it will not affect the entire communication system. Therefore such protocols are suitable for harsh environments where topology of the network is changing very rapidly.

Hybrid routing protocol is a combination of both the tree and mesh and is suitable for an environment with moderate mobility. It is as efficient as tree based protocols and at the same time it survives the frequent breaks in the network due to high mobility of nodes.

A comparison of all multicast routing protocols discussed above has been summarized in Table1 at the end.

V. Conclusion

Mobile Ad hoc network faces variety of challenges like Dynamic network topology structure, Limited bandwidth transmission, The limitation of mobile terminal, Distributed control, Multihop communication and Security therefore routing is more difficult in such challenging environment as compare to other networks.

Multicast routing is a mode of communication in which data is sent to group of users by using single address. On one hand, the users of mobile Ad Hoc Network need to form collaborative working groups and on the other hand, this is also an important means of fully using the broadcast performances of wireless communication and effectively using the limited wireless channel resources.

This paper summarizes and comparatively analyzes the routing mechanisms of various existing multicast routing protocols according to the characteristics of mobile Ad Hoc network.

References

1. T. Nadeem, and S. Parthasarathy, "Mobility Control for Throughput Maximization in Ad hoc Networks," Wireless Communication and Mobile Computing, Vol. 6, pp. 951 967, 2006.
2. CHEN-CHE HUANG AND SHOU-CHIH LO, "A Comprehensive Survey of Multicast Routing Protocols for Mobile Ad Hoc Networks"
3. T. Ozaki, J.B. Kim, and T. Suda, "Bandwidth efficient Multicast Routing for Multi hop Ad hoc Networks," in Proceedings of IEEE INFOCOM, Vol. 2, pp. 1182 1191, 2001.
4. X. Zhang, L. Jacob, "MZRP: An Extension of the Zone Routing Protocol for Multicasting in MANETs," Journal of Information Science and Engineering, Vol. 20, pp. 535 551, 2004.
5. P. Sinha, R. Sivakumar, and V. Bharghavan, "MCEDAR: Multicast Core Extraction Distributed Ad hoc Routing," IEEE Wireless Commun. and Net.Conf. (WCNC), pp. 13131317, 1999.
6. L. S. Ji and M.S. Corson, "Differential Destination Multicast a MANET Multicast Routing for Multihop Ad hoc Network, in Proceedings of IEEE INFOCOM, Vol. 2, pp. 11921201, 2001.
7. C. W. Wu, Y. C. Tay, C. K. Toh, "Ad hoc Multicast Routing Protocol Utilizing Increasing IdNumberS (AMRIS) Functional Sspecification," Internet-Draft, draft-ietf-manet-amris-spec-00.txt, 1998.
8. J. Xie, R. Talpade, T. McAuley, and M. Liu, "AMRoute: Ad hoc Multicast Routing Protocol," ACM Mobile Networks and Applications (MONET) Journal, Vol. 7, No.6, pp. 429439, 2002.
9. E. M. Royer and C. E. Perkins, "Multicast Operation of the Ad-hoc On-demand Distance Vector Routing Protocol", in Proc. ACM MOBICOM, pp. 207-218, Aug. 1999.
10. C. E. Perkins and E. M. Royer, "Ad-hoc On-demand Distance Vector Routing", in Proc. IEEE WMCSA, pp. 90-100, Feb. 1999.
11. L.-S. Ji and M. S. Corson, "Explicit Multicasting for Ad Hoc Networks", *Mobile Networks and Applications*", Vol. 8, No. 5, pp. 535-549, Oct. 2003.
12. C. W. Wu and Y. C. Tay, "AMRIS: A Multicast Protocol for Ad Hoc Networks", *in Proc. IEEE MILCOM*, Vol. 1, pp. 25-29, Nov. 1999.
13. J. G. Jetcheva and D. B. Johnson, "Adaptive Demand-driven Multicast Routing in Multi-hop Wireless Ad Hoc Networks", *in Proc. ACM MOBIHOC*, pp. 33-44, Oct. 2001.
14. P. Sinha, R. Sivakumar, and V. Bharghavan, "CEDAR: A Core Extraction Distributed Ad Hoc Routing Algorithm", *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 8, pp. 1454-1466, Aug. 1999.

Relevance of Cloud Computing in Academic Libraries

Dr. Prerna Mahajan*
Dr. Dipti Gulati**

Abstract

Cloud computing is one of the most recent technology models for IT services which is being adopted by several organizations and individuals.Cloud computing allows them to avoid locally hosting and operating multiple servers over an organization's network and constantly dealing with hardware failure, software installation, upgrades, backup &various compatibility issues which also enables them to save costs. Cloud Computing emerged as a significant advantage to the libraries and is offering various opportunities for libraries to connect their services with Cloud computing. This paper presents an overview of cloud computing and its possible applications that can be clubbed with library services in a web-based environment.

Keywords: Cloud Computing, Academic Libraries

Introduction

Cloud computing is the latest technology model for IT services, which a large number of organizations and individuals are adopting. Cloud computing transforms, the way systems are built and services delivered, providing libraries with an opportunity to extend their impact. Cloud computing is internet-based computing, in which virtual shared servers provide software, infrastructure, platform devices and other resources and hosting to customers on a pay-as-you-use basis. Presently, most of the organizations and individuals use computers to work alone, inside a business or at home by investing on hardware, software and maintenance. This scenario is slowly altering due to the emergence of a new breed of Internet services, popularly known as Web 2.0, through which any individual can use the power of computers at a completely different location, what it is popularly called as '**in the cloud**' or '**Cloud Computing**'.

Dr. Prerna Mahajan*
Head of the Department
Institute of Information Technology and Management
Dr. Dipti Gulati**
Librarian
Institute of Information Technology and Management

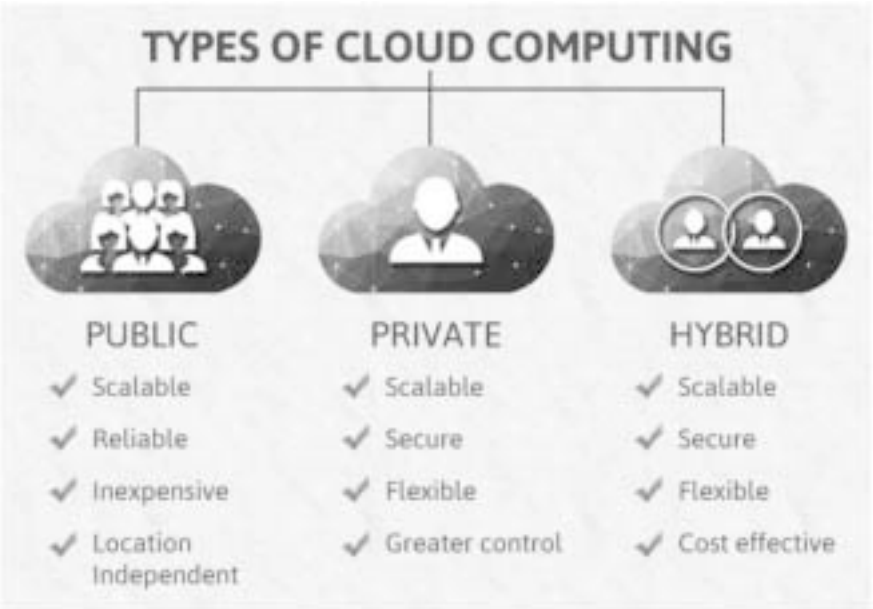
There are various synonyms for Cloud Computing such as, 'On-Demand Computing', 'Software as a Service', 'Information Utilities', 'The Internet as a Platform' besides numerous others.

According to the US National Institute of Standards Technology (NIST), "Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management efforts or service provider interaction". [1]

Cloud computing, often referred to as simply "the cloud," is the delivery of on-demand computing resources—everything from applications to data centers—over the internet on a pay-for-use basis.

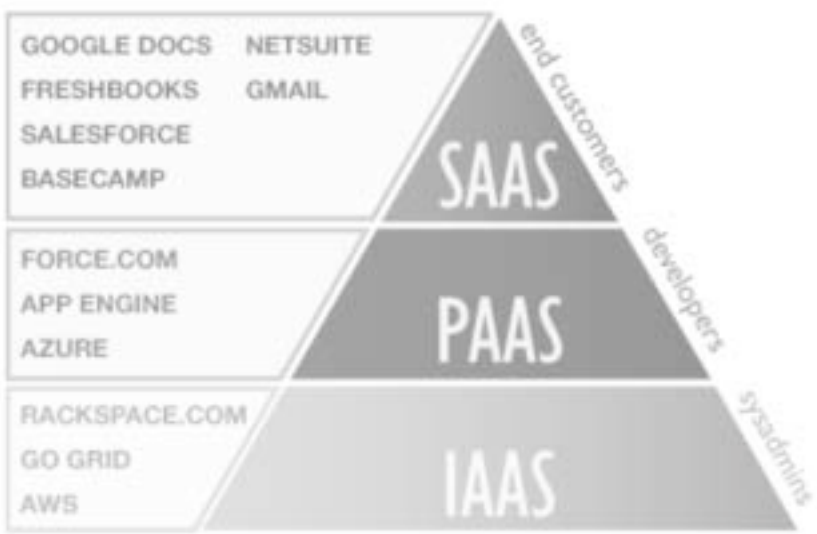
- Elastic resources—Scale up or down quickly and easily to meet demand
- Metered service so you only pay for what you use
- Self service—All the IT resources you need with self-service access.[2]

Cloud computing refers to the use of web for computing needs which could include using software applications, storing data, accessing computing power, or using a platform to build applications. There is a vast array of utilities ranging from e-mail, to word processing or photo sharing or video sharing where a person can use



http://convergenceservices.in/blog

products that live in the cloud, which are secure, backed-up and accessible from any Internet connection. The best live example of this is Gmail, which is increasingly being used by organizations and individuals to run their e-mail services. Google Apps being free for educational institutions is widely used for running a variety of applications, especially the email services, which were earlier being using on their own computer servers. This has proved to be cost effective organizations since they pay-per-use for applications and services and saves precious time for the computer staff, which they can invest on running other services without worrying about upgrading, backup, compatibility, and maintenance of servers, which is taken care of by Google. Libraries use computers for running services, such as, Integrated Library Management Software (ILMS), website or portal, digital library or institutional repository. These are either maintained by parent organization's computer staff or library staff, which involves huge investments on hardware, software, and helps staffs to maintain the



http://www.globaldots.com/cloud-computing-types-of-cloud/

services and undertake the backups and upgrades, when new version of the software gets released.

Library professionals in most of the cases are not being adequately trained in maintaining servers and often find it difficult to undertake some of these activities without the support of IT staff from within the organization or through external sources. In the present day, Cloud Computing has become the latest buzzword in the field of libraries, which is blessing in disguise to operate various ICT services without any problem since third-party services will manage servers and undertake upgrades and take back-up of data. Currently, some of the libraries have adopted the use of cloud computing services as an emerging technology to operate their services despite the fact that there are certain areas of concern in using cloud services such as privacy, security, etc.

Types of Cloud Computing

There are four types of Cloud Computing:

1. Private/Internal Cloud: Cloud operated internally for a single enterprise.
2. Public/External Cloud: Applications, Storage and other resource materials that are made available to the general public by the service providers.
3. Community Cloud: A Public Cloud tailored to a particular community.
4. Hybrid Cloud: A Combination of the internal and external cloud. This type of hybrid cloudin the Community clod and Hybrid Cloud are used interchangeably.

Cloud Computing Models

Cloud Computing Providers offer their services which can be grouped into three categories:

1. **Software as a Service (SaaS):** In this model, a complete application is offered to the customer, as a service on demand. A single request of the service runs on the cloud & multiple end users are serviced. Today SaaS is offered by the companies that are: Google, Salesforce, Microsoft and Zoho.
2. **Platform as a Service (PaaS):** In this model, a layer of software or development environment is condensed and offered as a service, upon which other higher levels of service can be built. The customer has the freedom to build his own applications, which run on the provider's infrastructure. To meet manageability and scalability requirements of the applications, PaaS providers offer a predefined combination of OS and application servers, such as LAMP Platform (Linux, Apache, MySql and PHP), restricted J2EE, Ruby, Google's App Engine, Force.com, which are some of the popular PaaS examples.
3. **Infrastructure as a Service (IaaS):** IaaS provides basic storage and computing capabilities as standardized services over the network. Servers, storage systems, networking equipment, data center space are pooled and made available to manage workloads. The customer would typically deploy his own software on the infrastructure. Some of the common examples are Amazon, GoGrid, 3 Tera, et al.

Application of Cloud Computing in Libraries

Libraries are shifting their services with the attachment of cloud and networking with the facilities to access these services anywhere and anytime.

In the libraries, the following possible areas were identified where cloud computing services and applications may be applied:

1. **Building Digital Library/Repositories:** In the present situation, every library requires a digital library to offer their resources, information and services at an efficient level to ensure access via the network. Therefore, every library has a digital library that is developed through the use of any digital library software.
2. **Searching Library Data:** OCLC is one of the best examples for utilizing cloud computing for sharing libraries data for years together. OCLC World Cat service is one of the well-accepted services for searching library data that now is available on the cloud. OCLC is offering various services pertaining to circulation, cataloguing, acquisition and other library related services on the cloud platform through the web share management system. A Web share management

system facilitates in the development of an open and collaborative platform in which each a library can share their resources, services, ideas and problems with the library community on the clouds. On the other hand, the main objective of web-scale services is to provide cloud based platforms, resources and services with cost-benefit and effectiveness to share the data and building the broaden collaboration in the community.

3. **Website Hosting:** Website hosting is one of the earliest adoptions of cloud computing as numerous organizations including libraries prefer to host their websites on third party service providers rather than hosting and maintaining their own servers Google Sites, which serve as an example of a service for hosting websites externally of the library's servers and allowing for multiple editors to access the site from varied locations.

4. **Building Community Power:** The Cloud Computing technology offers tremendous opportunities for libraries to build networks among the library and information science professionals as well as other interested people including information seekers by using social networking tools. One of the most well-known social networking services, such as, Twitter and Facebook play a dominating role in building community power. This cooperative effort of libraries will create time saving efficiencies and a wider recognition, cooperative intelligence for better decision-making and provides the platform for innovation and sharing the intellectual conversations, ideas and knowledge.

5. **Library Automation:** For library automation purpose, Polaris offers variant cloud- based services, such as, acquisitions, cataloguing, process system, digital contents and provision for inclusion of cutting edge technologies used in libraries and also supports various standards such as MARC21, XML, Z39.50, Unicode and so on which directly related to library and information science area. Apart from this, nowadays a majority of the software vendors such as Ex-Libris, OSS Labs are also offering this service on the cloud and third party services providing hosting of this service (SaaS approach) on the cloud to save libraries from investing in hardware for this purpose. Besides cost-benefit, the libraries will be free from taking maintenance that is software updates, backup and other facilities.

Advantages and Disadvantages of Cloud Computing in Libraries

In the present situation of Indian Libraries in India, cloud computing in libraries is in the development phase. Libraries are attempting to offer their users cloud-based services however in reality they are not fully successful mainly due to lack of good service providers and technical skills of LIS professionals in the field of library management using advanced technology. Yet some of the services such as digital libraries, web documentation and using Web2.0 technologies are operating on a successful mode. Some of the excellent examples of successful cloud computing libraries include Dura cloud, OCLC services and Google-based cloud services. In the current state, countless commercial as well as open sources venders (i.e. OSS) are clubbing the cloud computing technology into their services and products. However, cloud computing technology is not totally accepted in the Indian libraries although they are trying to develop themselves in this area.

Conclusion

Cloud Computing represents an exciting opportunity to bring on-demand applications to Digital Library in an environment of reduced risk and enhanced reliability. However, it is important to understand that existing applications cannot just be unleashed on the cloud as they are in existence. A careful attention to the design detail will help in ensuring a successful deployment. Certainly cloud computing can bring about strategic, transformation and even revolutionary benefits fundamental to digital libraries. As regards to organizations providing digital libraries, with significant investment in traditional software and hardware infrastructure, migration to the cloud will highlight considerable technology transition; for less-constrained organizations or those with infrastructure nearing end-of-life, adaptation of cloud computing technology may be more immediate.

No doubt, libraries are shifting towards cloud computing technology in the present times and taking advantages of these services, especially in building digital libraries, social networking and information communication with manifold flexibilities yet some issues related to security, privacy, trustworthiness and legal issues are still not completely resolved. Therefore, it is high time for libraries to think seriously before clubbing libraries services with cloud-based technologies and provide reliable and rapid services to their users. Another responsibility of LIS professionals in this virtual era is to make cloud based services a reliable medium to disseminate library services to their target users with ease of use and trustworthiness.

References

1. Aravind Doss, and Rajeev Nanda. (2015). "*Cloud Computing: A Practitioner's Guide.*" TMH. New Delhi. P-265.
2. https://www.ibm.com/cloud-computing
3. Anna Kaushik and Ashok Kumar. (2013). "*Application of Cloud Computing in Libraries.*" International Journal of Information Dissemination and Technology. 3 (4): 270-273.
4. Jadith Mavodza. "*Impact of Cloud Computing on the Future of Academic Libraries and Services.*" Proceedings at the 34th Annual Conference of the International Association of Scientific and Technological University Libraries (IATUL), Cape Town, South Africa.
5. Anthony T Velte. and Others. (2015). "Cloud Computing: A Practical Approach". TMH: New Delhi. P- 1-23.
6. Aravind Doss, and Rajeev Nanda. (2015). "*Cloud Computing: A Practitioner's Guide.*" TMH. New Delhi. P-265-268.

A brief survey on metaheuritic based techniques for optimization problems

Kumar Dilip*
Suruchi Kaushik**

Abstract

This paper aims to provide a brief review of few popular metaheuristic techniques for solving different optimization problems. In many non-trivial real life optimization problems finding an optimal solution is a very complex and computationally expensive task. Application of the classical optimization techniques is not suitable for such problems due to its inherent complex and large search space. In order to solve such optimization problems, metaheuristic based techniques have been applied and popularized in recent years. These techniques are increasingly getting the recognition as effective tools for solving various complex optimization problems in reasonable amount of computation time. In this brief survey of metaheuristic techniques we discuss few existing as well as ongoing developments in this area.

Keywords: Optimization problems; metaheuristics; Gentic algorithm; Ant Colony Optimization

I. Introduction

Application of metaheuristic based techniques for solving real life complex decision making problems is gaining popularity as the underlying search space of such problems are complex and huge in size [2,22]. Although, the heuristic based methods have been considered as a viable option for solving the complex optimization problems as they are likely to provide good solutions in reasonable amount of time. However the limitation with the heuristic based technique is the focus on the specific feature of the underlying problem, which makes the design of approach very difficult. In order to address this issue the application of metaheuristic based methods is considered as a feasible option. They are not problem specific and can be effectively adapted for the different types of optimization problems. Alternatively, the metaheuristic techniques provide a generic algorithmic approach to solve various optimization problems by making comparatively few adjustments according to problem specification. In general three common features can be identified in most of the metaheuristic techniques among others. First, majority of them are inspired by several working mechanisms of nature which include biology and physics. Second, they consider many random variables to perform the flexible stochastic search of the large search space. And third, they also involve the various parameters and proper tuning of them can greatly affect the overall performance of the techniques for the considered problem. The effectiveness of the metaheuristic technique for problem at hand significantly lies on two major concepts, known as intensification or exploitation and diversification or exploration. The exploration tries to identify the potential search area containing good solutions while exploitation aims to intensify the search in some promising area of search space. The optimal balance between these two mechanisms during search process may lead towards comparatively better solutions [2, 22].

The application of metaheuristic techniques is considered well suited for those optimization problems where no acceptable problem-specific algorithms are available for solving them. The application area of metaheruistic techniques include, finance, marketing, services, industries, engineering, multi-criteria decision making among others. These techniques may provide good or acceptable solutions to various complex optimization problems in this area with effective computation time.

Kumar Dilip*
Department of IT
IITM
Suruchi Kaushik**
Department of IT
IITM

In recent years , popular metaheuristic techniques such as Evolutionary algorithm, Genetic algorithm, Ant Colony Optimization, Particle Swarm Optimization, Bee colony optimization, Simulated Annealing, Tabu Search etc. have been widely used for different optimization problems[11,12, 13, 16, 17, 21, 24, 25, 26]. All of the above techniques have certain underlying working principle and various strategic constructs that may enable them to solve the problems efficiently. However, in recent few years a new kind of metahueristic which is unlike the above approaches, do not belong to a specific metaheuristic category but combines the approaches form the different areas like computer science, biology, artificial intelligence and operation research etc. These new class of metaheuristic techniques are normally referred as Hybrid metaheuristc. In order to improve the performance, concept of quantum computing has also been applied to solve the optimization problems. With the intent of further improving the performance of the approaches various quantum inspired metaheuristic techniques have been proposed in literatures [14].

The lists of metaheuristic techniques are extensive and it is difficult to summarize them in a brief survey, this paper also not intended to do so. Rather, this paper attempt to give a brief introductory overview of few popular metaheuristic techniques. In the next section classification of the metaheuristic based techniques has been described.

II. Classification of metaheuritstic techniqeus

Many criteria can be found for the classification of various metaheuristic techniques. However the more common classification of metaheuristic techniques, based on the use of single solution and population of solutions can be found in literature. The popular single solution based techniques also known as the trajectory methods include, Simulated Annealing, Tabu Search, Variable Neighborhood Search, Guided Local Search, Iterated local search [27,28]. The single solution based approaches start with single initial solution and gradually move off from this solution depicting a trajectory movement in large search space [27, 28].

Unlike single solution based metaheuristic techniques the population based metaheuristic techniques begin with a population of solutions and in every algorithmic iteration attempt to move towards the better solutions. In recent years the population based metaheuristic techniques have been gaining comparatively more popularity and more new population based techniques are getting reported in literature [21, 22, 23]. Keeping this in mind this paper majorly focus on the population based techniques. However the details of the single solution based or trajectory based metaheuristic techniques can be found in the literature [21, 22, 23]. In the next section we describe two popular population based metahuristic techniques.

III. Population based metaheuristic techniques

The majority of population based methods either belongs to class of Evolutionary algorithms or Swarm Intelligence based methods. The inherent mechanism of evolutionary algorithm is mainly based on the Darwin's theory of the survival of the fittest. The population of solutions improves iteratively generation after generation. Fitter solutions are selected to reproduce the better solutions for the next generation. However, in Swarm intelligence based techniques, instead of a single agent, the collective intelligence of the group is exploited to find the better solutions iteratively.

Evolutionary algorithms refer to a class of metaheuristic techniques whose underlying working mechanism is based on the Darwin's theory of evolution. According to this theory the fitter living beings which can better adapt in the changing environment can survive and can be selected to reproduce the better offspring. This generic class of techniques includes evolutionary programming, Genetic algorithms, Genetic programming, evolutionary strategies etc.[15,18,19,20,29]. Though these techniques differ in their algorithmic approach, yet their core underlying working is similar. The evolutionary algorithms are mainly characterized by three important aspects, first the solution or individual representation, second the evolution function and third population dynamics throughout the algorithmic runs. All of the evolutionary techniques in every generation or algorithmic iteration attempt to select the better solutions in terms of its objective function values. These solutions further apply the mechanism of recombination and mutation operator to produce the

Procedure Evolutionry Algorithm

```
Begin Procedure
Initialize the population of the individuals or solutions,
Evaluate the fitness of the each individulas,
While stopping criteria not met, do
    Select the fitter individual as parents
    Recombine the pair of fitter solutions to produce offspring
    Perform the mutation on the offspring solutions
    Evaluate the new individuls or solutions
    Select the fitter solutions for the next generation
End While
End Procedure
Return solution.
```

Figure 1: A generic view of Evolutionary Algorithm

better solutions in the next generations. Next a generic evolutionary approach has been described in order to depict the common algorithmic steps in the above evolutionary algorithms.

In the above procedure each iteration indicates a generation in which population of individuals or candidate solutions are evaluated to check its fitness according to given objective function of the problem at hand. Among those individuals the set of fitter individuals are selected by applying some suitable selection mechanism. The pairs of fitter solutions are selected to perform the recombination to produce the better offspring solutions. Further the mutation is performed on the offspring with the intent of promoting the diversity in the solutions. These newly created solutions are evaluated for the given objective function to check their suitability to use it for the next generation. The above procedure will continue iteratively till the termination condition is satisfied. The possible termination condition can be predetermined number of generation or the condition when there is no further improvement in solutions. There may also be other possible criteria for the termination of the algorithmic runs.

Genetic Algorithm (GA)

The idea of Genetic algorithm were first introduced by John Holland in 1970's. This evolutionary search Technique has been widely applied for different types of real world optimization problems. As an evolutionary technique, the concepts of Genetic algorithms are based on the Darwin's evolutionary theory in which fitter indivdulas are likely to survive and having the higher probability of production offsprings for the next genration. This very idea has been adapted in the algorithmic framework of genetic algorithms. The candiadate slutions or population of individuals iteratively evolve towards the search space of fitter or better solutions in each algorithmic iteration. In order to apply the GA for problem solving, the algorithmic requirement is to decide the repersentation of the solution or the chromosome. A binary or alphabetic string of fixed length is common representation of candidate solution in GA implementation. Next rquirement is to choose from the various selection strategy in order to select the fitter solutions, most popular selection and use of various possible crossover and mutation operators. A candidate solution is represented by a chromosome and a number of chromosomes constitute the entire population of the current generation. A population in current generation evolves to next generation through above mentioned three main operators i.e. selection, crossover and mutation. All these operators play a crucial part in the performace of the Genetic algorithm for the considered problem and their proper tuning is essential aspect of the GA implementiation. In most of the cases the focus is on the crossover as a variation operator. The crossover operator is usually applied on the pair of the selected chromosome after performing selection strategy. The various crossover operators can be found in the literature and their application may

depend upon the considered problem and or also on the solution representation. With the help of crossover operator two or more solutions may exchange their genetic materials or some part of the solutions and create new individuals. The cross over rate of the population indicates the total number of chromosomes or solutions that would undergo the crossover or recombination. Each chromosome in the population has a fitness value determined by the objective function. This fitness value is used by selection operator to evaluate the desirability of the chromosome for next generation. Generally, fitter solutions are preferred by the selection operator but some less fitter chromosomes can also be considered in order to maintain the population diversity. Crossover operator is applied on the selected chromosomes to recombine them and generate new chromosome which might have better fitness. Mutation operator is applied to maintain the population diversity throughout the optimization process by introducing random modifications in the population.The Evoluationary algorithms have been applied for the optimization problems of the diverse area. It has been succesfully applied for the different combinatorial optimization problems and constrained optimization problems[7]. In recent years, it is also getting popularity in the area of multi-criteria optimization problem. Finding the trade-off solutions for the multi-objective optimization problem is a complex task. Evoluationary algorithms based techniques like NSGA-II has been successfully applied for several multi-objective optimization problem [1,3,8,9,10].

In recent years the quantum inspired Genetic algorithm is also getting a lot of attention. It applies the pricipal of quantum computing combined with evolutionary algorithm [14]. Insetead of binary, numeric or symbolic repersentation, Quantum inspired algorithm applies Q-bit repersentation and Q-gate operator is used as a variation operator.

Next we describe the swarm intelligence based technique, Ant colony optimization or ACO.

Ant Colony Optimization (ACO)

Ant colony optimization is a metaheuristic wich is inspried by the behaviour of the real ants. This approach was first applied for solving Travelling Salesman problem [5]. In majority of the cases, where ACO is applied the problem subjected to is represented with a graph. ACO is a population based metaheuristic. Various ants of real world, in search of their food, work in a group and they find the shortest path from nest to the food source. This very behaviour of real ants has inspired the ant colony optimization, in which a group of simple agents work in co-operation in order to achieve the complex task. The real world ants attempt to find the quality food sources nearest to their colony. In this pursuit they deposit some chemicals on the search path also known as pheromones. The paths with good food sources and lesser distance from nest is likely to get more amount of pheromones. Paths with higher pheromone density are highly likely to be selected by following ants. Such behaviour of ants gradually leads towards the emergence of the shortest path from nest to good food source. Alternatively, it can be observed that the indirect communication or communication through enviroment, by using pheromone trails and without any central control among ants, they are likely to find the shortest path from their colony to food source. In addition, artficial ants of Ant Colony Optimization have some extra characteristics which real ants do not have. These characteristics include presence of memory in artificial ants of ACO, which helps in constructing the feasible candidate solutions and awareness about its environment for better decsion making during the solutions construction. In ACO, ants probabilistically construct solutions using two important information known as pheromone information and heuristic information. The pheromone information $\tau(ij)$ repersents the amount of pheromone on edge or solution component (i,j) and $\eta(ij)$ repersents the preference of selection of node j from node i, during solution construction. Both of these values are reperented using numeric values. Both of these values influence the process of search towards higher pheromone values and heuristic information values. In addition, the pheromone information or denstiy on the path are updated at every algorithmic iteration. The pheromone information repersents the past search experience while heuristic information is problem specfic which remains unchanged throughout the algorithmic run of ACO. The solution in each iteration is probabilistically constructed using the following formula:

$$p(ij) = \begin{cases} \dfrac{[\tau(ij)]^{\alpha}[\eta(ij)]^{\beta}}{\sum_{l \in N_i}[\tau(il)]^{\alpha}[\eta(il)]^{\beta}}, & \text{if } l \in N_i \\ 0, & \text{otherwise} \end{cases}$$

P(ij) repersents the probability of selection of node *j* after node *i* in partially consturcted solution, *l* indicates the available nodes for the solution construction or the nodes which are not already part of partially constructed solution. Here α and β indicate the relative importance for pheromone information and heuristic information respectively.

After the completion of solution construction, a mechanism of evaporation is applied with the intent of forgetting the unattractive choices and no path become too dominating as it may lead towards the premature convergence. The path update at every iteration performed using the following formula:

$$\tau(ij) \leftarrow (1-\rho)\tau(ij) + \rho.\tau(0)$$

In the above formula, ρ indicates the pheromone decay coefficient, $\tau(0)$ indicate some intial pheromone value deposited on the edge (*ij*).

In addition, daemon actions such as local search can be applied as an optional action to further improve the quality of solution. The first ant colony based optimization technique was proposed in [6] to solve the single objective optimization problems. After the initial work of ant system, many variants of ant based optimization techniques have been proposed in literature for solving various combinatorial optimization problems such as Travelling salesman problem, vehicle routing problem, production scheduling, quadratic assignment problems, among others[4,5,6]. An abstract view of the ACO is as follows:

Procedure ACO

Initialize pheromone matrix τ,
Initialize heuristic factor η,
While stopping criteria not met do
Perform *ProbailisticSolutionsConstuction( )*
Perform *LocalSearchProcess( )* // optional action
Perform *PheromoneUpdateProcess()*
End While
End Procedure
Return best solution.

Figure 2. An ACO procedure [4,5,6]

An ant based system consists of multiple stages as shown in figure 2. In the first step, evaluation function and the value of pheromone information (τ) are initialized. In the next step, at each algorithmic iteration, each ant in a colony of ants incremently constructs the solution by probabilistically selecting the feasible components or nodes from the available nodes. As an optional action, local serach can be performed for further improvement of the quality of solution. Once each ant completes the process of the solution constuction, the process of pheromone update using evaporation mechanism is performed. The best solution/solutions in terms of the value of the given objective function is chosen to update the pheromone

information. The algorithmic iteration of solution construction and pheromone update ends when it meets some predefined condition and the best solution is returned. This could be some predefined number of generation or the condition of stagnation when there is no further imporvment in solution is found.

The ACO has been widely and succesfully applied for the various problems which include Travelling Salesman problem, vehicle routing, Sequential ordering, Quadratic Assignment, Graph coloring, Course timetabling, Project sheduling, Total weighted tardiness, Open shop, Set covering, Multiple knapsack, Maximum clique, Constraint satisfaction, Classification rules, Bayesian networks, Protein folding among others [4]. In recent years it has been also gaining popularity for solving various multi-objective optimization problems.

Conclusion

In this survey we have briefly described the metaheuristic based techniques for solving various optimization problems. Considering the distinction between the metaheuristic techniques based single solutions approach and population based approaches, we described introductory idea of two popular and widely used population based approaches including Genetic algorithm and Ant colony optimization.

References

1. Asllani, A., & Lari, A. (2007). 'Using genetic algorithm for dynamic and multiple criteria web-site optimizations', European journal of operational research, Vol. 176, No. 3, pp. 1767-1777
2. Basseur, M., Talbi, E., Nebro, A. & Alba, E. (2006). 'Metaheuristics for Multiobjective Combinatorial Optimization Problems: Review and recent issues', INRIA Report, September 2006, pp. 1-39
3. Coello-Coello, C. A., Lamont, G. B. & van Veldhuizen, D. A. (2007). 'Evolutionary Algorithm for solving multi-objective problems, Genetic and Evolutionary Computation Series', Second Edition, Springer.
4. Dorigo, M. & stutzle, T. (2004). Ant colony optimization, Cambridge: MIT Press, 2004
5. Dorigo, M. & Gambardella, L.M.,(1997) 'Ant colonies for the traveling salesman problem', BioSystems, vol. 43, no. 2, pp. 73–81, 1997.
6. Dorigo, M., Maniezzo,V. & Colorni, A., (1996) 'Ant System: Optimization by a colony of cooperating agents,' IEEE Transactions on Systems, Man, and Cybernetics—Part B, vol. 26, no. 1, pp. 29–41, 1996.
7. Kazarlis, S.A., Bakirtzis, A.G. & Petridis, V (1996). 'A genetic algorithm solution to the unit commitment problem', IEEE Transactions on Power System, Volume 11, Number 1, pp. 82-92
8. Deb, K., Pratap, A., Agarwal, S & Meyarivan, T. (2002). 'A fast and elitist multiobjective Genetic Algorithm: NSGA-II', IEEE Transaction on Evolutionary Computation, Vol. 6, No. 2. pp. 182-197
9. Deb, K. (2010). Multi-objective optimization using Evolutionary algorithms. Wiley India.
10. Doerner, K. F., Gutjahr, W. J., Hartl, R. F., Strauss, C. and Stummer, C (2004). "Pareto ant colony optimization: A metaheuristic approach to multiobjective portfolio selection," Annals of Operations Research, vol. 131, pp. 79–99,2004.
11. T'kindt, V., Monmarch´e, N., Tercinet, F. & La¨ugt, D (2002). "An ant colony optimization algorithm to solve a 2-machine bicriteria flowshop scheduling problem," European Journal of Operational Research, vol. 142, no. 2, pp. 250–257, 2002
12. Wang L., Niu, Q. & Fei, M.(2007) 'A Novel Ant Colony Optimization Algorithm', Springer Verlag Berlin Heidelberg. LNCS 4688, pp. 277– 286, 2007
13. Goldberg, D. E. (1989). Genetic Algorithm in Search, Optimization and Machine Learning, Pearson Education, India

14. Han, K.–H. & Kim, J.–H., (2000)'Genetic quantum algorithm and its application to combinatorial optimization problem,' in Proc. Congress on Evolutionary Computation, vol. 2, pp. 1354-1360, La Jolla, CA,2000.

15. X. Yao, Y. Liu, Fast evolutionary programming, in: Evolutionary Programming, 1996, pp. 451–460.

16. F. Vandenbergh, A. Engelbrecht, A study of particle swarm optimization particle trajectories, Information Sciences 176 (2006) 937–971.

17. S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.

18. J.R. Koza, Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems), first ed., The MIT Press, 1992.

19. T. Bäck, H.P. Schwefel, An overview of evolutionary algorithms for parameter optimization, Evolutionary Computation 1 (1993) 1–23.

20. S. Baluja, Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning, Technical Report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

21. F. Glover, Tabu search for nonlinear and parametric optimization (with links to genetic algorithms), Discrete Applied Mathematics 49 (1994) 231– 255.

22. M. Birattari, L. Paquete, T. Stützle, K. Varrentrapp, Classification of Metaheuristics and Design of Experiments for the Analysis of Components, Technical Report AIDA-01-05, FG Intellektik, FB Informatik, Technische Universität Darmstadt, Darmstadt, Germany, 2001.

23. E.G. Talbi, Metaheuristics: From Design to Implementation, first ed., Wiley-Blackwell, 2009.

24. S. Jung, Queen-bee evolution for genetic algorithms, Electronics Letters 39 (2003) 575–576.

25. D. Karaboga, An Idea Based on Honey Bee Swarm for Numerical Optimization, Technical Report TR06, Erciyes University, 2005.

26. D. Karaboga, B. Akay, A survey: algorithms simulating bee swarm intelligence, Artificial Intelligence Review 31 (2009) 61–85.

27. N. Mladenovic, A variable neighborhood algorithm – a new metaheuristic for combinatorial optimization, in: Abstracts of Papers Presented at Optimization Days, Montréal, Canada, 1995, p. 112.

28. N. Mladenovic, P. Hansen, Variable neighborhood search, Computers and Operations Research 24 (1997) 1097–1100.

29. X. Yao, Y. Liu, G. Lin, Evolutionary programming made faster, IEEE Transactions on Evolutionary Computation 3 (1999) 82–102.

Cross-Language Information Retrieval on Indian Languages: A Review

Nitin Verma*
Suket Arora**
Preeti Verma***

Abstract

Cross Language Information Retrieval on Indian Languages (CLIROIL) can be used to improve the ability of users to search and retrieve documents in different languages. The aim of CLIR is to provide the benefit to the user in finding and assessing information without being limited by language barriers. We can use Simple measures to get high - accuracy in cross-language retrieval in which translation is one of them. Translation is one of the technique that makes use of software that translates text from one language to another language. Different type of translation techniques (dictionary based translation, machine translation, transitive translation, dual translation) can be used to achieve Cross Language Information Retrieval. IR deals with presentation, storage, space, retrieval, and access of a multiple document collection. This paper describes the work done in CLIR and translation techniques for CLIR. This paper translates the work done.

Keywords: CLIROIL, Translation, Dictionary-based, Machine translation, Transitive translation.

I. Introduction

Cross Language Information Retrieval On Hindi Language allows the users to read and search pages in the language different from the other language of being searched. Cross language information retrieval is a kind of information retrieval in which the language of the query is different from the language of the documents retrieved as in a search result. In Cross Language Information Retrieval system a user is not limited to his own native language, different set of languages are there, so the user can make his query in his native language but the system returns set of documents in another different languages. Different foreign languages have been used like English, French, Spanish, Chinese. But Indian languages always have Cross Language Information Retrieval On Hindi Language allows the users to read and search pages in the language different from the other language of being searched. Cross language information retrieval is a kind of information retrieval in which the language of the query is different from the language of the documents retrieved as in a search result. In Cross Language Information Retrieval system a user is not limited to his own native language, different set of languages are there, so the user can make his query in his native language but the system returns set of documents in another different languages. Different foreign languages have been used like English, French, Spanish, Chinese. But Indian languages always have system simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help from translators for using other languages documents. CLIR system simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help from translator for using other languages documents. CLIR. System simplifies the search process for multiple users and enables those who know only one language to provide queries in their language and then get help

Nitin Verma*
Assistant Professor, Computer Science Dept., Hindu College, Amritsar

Suket Arora**
Assistant Professor, Dept. of Computer Applications, Amritsar College of Engineering & Technology, Amritsar

Preeti Verma***
Assistant Professor, Dept. of Computer Applications, Amritsar College of Engineering & Technology, Amritsar

from translators for using other languages documents. Due to the "standardization" of terms, stemming sometimes contributes in increasing the retrieval effectiveness. This is, however, not always the case. Current search engines usually do not use aggressive stemming, while in the area of research, stemming is still generally used as a standard pre-processing.

II. Translation

A full document translation can also be applied offline to create translation of an entire document. The translations provide the basis for constructing an index for information retrieval and also offer the user the possibility to access the content in his native language. Multiple information search becomes important due to large amount of online information available in different languages. We can also use an online translation through sources like i.e. Google, Wikipedia which confirms the accuracy of the search. Usually machine translation system supports the translation. Searching strategies are continuously improving their techniques to provide more relevant, accurate and proper information for a given query. A common problem with translation is word accuracy. This problem can be solved by using different techniques .Various techniques are used to reduce the grammatical mistakes. The Search can also be filtered by providing the unrestricted domains. Machine Translation is not always available as a realistic option for every pair of languages. Widely translation system supports the translation between language pairs which involve the languages likely as English, German or Spanish, and Chinese. In translating the document, firstly we select a single query language and then translate every single document into that language then single retrieval is carried out. This technique provides more context but current systems don't damage the context widely. But one must have to determine in which language each document should be translated; translated documents in all the languages should be stored.

III. Translation Techniques

Translation techniques in CLIR are categorized into two types:

- Direct translation
- Indirect translation

A. Direct Translation

The direct is of three types. Now we will explain them:

- Corpus Based Translation
- Dictionary Based Translation
- Machine Based Translation

1) Corpus Based Translation

Parallel corpora are commonly used in cross-language information retrieval to translate queries. The basic technique involves a side-by-side analysis of the corpus producing a set of translation probabilities for each term in a given query[1]. Large collections of parallel texts are referred to as parallel corpora. Parallel corpora can be acquired from a variety of sources.

2) Dictionary Based Translation

A dictionary-based approach for the translation is very easy but it is having two limitations such as ambiguity and lack of coverage[1].

3) Machine Translation

Machine Translation is not only performs the substitution of words from one language to other; but it also involves finding phrases and its counterparts in target language to produce good quality translation.

B. Indirect Translation

Indirect translation relies upon the use of an intermediary which is placed between the source query and the target document collection. In the case of transitive translation, the query will be translated into an intermediate to enable comparison with the target document collection. The Indirect translation is two types:

- Transitive translation
- Dual translation

1) Transitive Translation

Transitive translation relies upon the use of a pivot language which acts as an intermediary between the source query and the target document collection[1].

2) Dual Translation

Dual translation systems attempt to solve the query document mismatch problem by translating the query representation and the document representations into some "third space" prior to comparison. This "third space" can be another human language, an abstract

language or a conceptual inter-lingual. This general category also includes translation techniques that induce a semantic correspondence between the query and the documents in a cross-language dual space defined by the documents.

IV. Approaches of clir

There are different approaches for CLIR. Following are approaches:

A. Query Translation

Multilingual information search becomes important due to increasing the amount of online information available in non-English languages and multiple language document collections. This can be achieved by Query translation. Query translation using CLIR became the widely used technique to access documents of the different languages from the language of query. For translating the query, we can use an online translation i.e. Google Translate, train a Statistical Machine Translation system using parallel corpora, employ Machine Readable Dictionaries to translate query terms or use of large scale multilingual information sources like Wikipedia . Google Translate query translation approach. Translation can be applied to the query terms online. Online query translation can be achieved by using one of the Google Translate API which will convert the query into the other languages. Online query translation will help the user to translate his query in the other languages. Online query translation will help the user to translate his query in the other languages [3].

B. Interlingual Translation

The Inter-lingual technique is useful if there is no resource for a direct translation but it has lower performance than the direct translation. The Inter-lingual technique is useful if there is no resource for a direct translation but it has lower performance than the direct translation [4].

C. Document Translation

In Document translation we select a single query language and then translate every document into that language then perform monolingual retrieval. Typically machine translation systems supports the translation between language pairs which involve languages, such as English, German or Spanish, and English.

D. Some Advance Approaches

1) Universal words

They confirm the vocabulary of the language. To be able to express any concept occurring in a natural language, the UNL proposes the use of English words modified by a series of semantic restrictions that eliminate the innate ambiguity of the vocabulary in natural languages. If there isn't any English word suitable to express the concept, the UNL allows the use of words from other languages. In this way, the language gets an expressive richness from the natural languages but without their ambiguity.

2) Relations

These are a group of 41 relations that define the semantic relations among concepts. They include argumentative (agent, object, goal), circumstantial (purpose, time, place), logic (conjunction, and disjunction) relations, etc.

V. Knowledge Representation

By knowledge bases in our context we understand the set of concepts belonging to a specific domain and the relations between these concepts that also belong to this domain. But when we turn to ontologies, the richness of a domain becomes relegated to a mere enumeration of concepts and a taxonomic organization of them. That is, there is danger of identifying ontologies as mere theasauri.[8]

VI. Challenges In CLIR

- Dictionaries only include the most commonly used proper nouns and technical terms used such as major cities and countries. Their translation is crucial for a good cross-language IR system. A common method used to handle untranslatable keywords is to include the non-translated word in the target language query. A phrase cannot be translated by translating each of the word in the phrases.
- Named entities extraction and translation are vital in the field of natural language processing for research on machine translation, cross language IR, bilingual lexicon construction, and so on. There are three types of Named entities; entity names such as organizations, persons and

locations, temporal expressions such as dates and times, and number expressions such as monetary values and percentages.

- Using the dictionary-based translation is a traditional approach in cross-lingual IR systems but significant performance degradation is observed when queries contain words or phrases that do not appear in the dictionary. This is called the Out-of-Vocabulary. This is to be expected even in the best of dictionaries. Translation Disambiguation, which is rooted from homonymy and polysemy[6]. Homonymy refers to a word which has at least two entirely different meanings, for example the word "left" can either mean opposite of right or the past tense of leave. Input queries by user usually short and even the query expansion cannot help to recover the missing words because of the lacking information.[7]
- A common problem with query translation is word inflection used in the query. This problem can be solved by stemming and lemmatization. Lemmatization is where every word is simplified to its uninflected form or lemma; while stemming is where different grammatical forms of a word are reduced to a common shortest form which is called a stem, by removing the ending in word. For example, the stemming rules for word "see" might return just "s" by stemming and "see" or "saw" by lemmatization[4].

VII. Applications of CLIR

- This CLIR System can be helpful for immigration department. For eg. Immegration department interact with thousands of the Indian native Language speakers which are not able to understand English Languages .
- This System can be used for multilingual population regions so that the peoples having different native languages retrieve documents in their native languages.
- This system can also be used for intelligence departments.
- The CLIR will be beneficial for students for their research work regarding historical places.

VIII. Conclusion

CLIROIL provides us a new technique for searching documents through different kinds of languages across the whole world .By using the different type of translation techniques CLIROIL make it possible to provide the better search results in the other language to the language which is queried. So it will be beneficial for wide population regions. Survey proves that query translation is much better than document translation. It is more convenient way to translate the query than the whole documents. Document translation which uses machine translation is computationally quite expensive and the size of document collection is large. However, it might be practical in the future when the computer technology would be much improved.

References

1. Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, Helen Ashman," Translation Techniques in Cross-Language Information Retrieval.
2. J. Cardeñosa, C Gallardo, Adriana Toni," Multilingual Cross Language Information Retrieval A new approach".
3. UNL Center. UNL specifications v 2005. http://www.undl.org/unlsys/unl/unl2005-e2006/
4. D. Manning, C., P. Raghavan, and H. Schütze, "*An Introduction toInformation Retrieval",* 2009.
5. Nurul Amelina, Nasharuddin, Muhamad Taufik Abdullah, "Crosslingual Information Retrieval",Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1.
6. Abusalah, M., J. Tait, M. Oakes, "Literature Review of Cross Language Information Retrieval",2005
7. Nurul Amelina, Nasharuddin, Muhamad Taufik Abdullah,"Crosslingual Information Retrieval",Electronic Journal of Computer Science and Information Technology,Vol. 2,No. 1,
8. Bateman, J.A; Henschel, R. and Rinaldi, F. "The Generalized Upper Model 2.0." 1995. http:// http:// www.fb10.unibrem en.de/anglistik/langpro/webspace/jb/gum/index.htm

Enhancing the Efficiency of Web Data Mining using Cloud Computing

Tripti Lamba*
Leena Chopra**

Abstract

Data Mining is the process of discovering actionable information from raw data, which helps to enhance the capability of existing business process. Due to the unrestricted use of Internet by individuals ubiquitously, limitless data has to be stored and maintained on servers. World Wide Web is a group of massive amount of information resources, interconnected files on Internet. Mining the valuable information from this huge source is the main area of concern. In cloud computing web mining techniques and applications are major areas to focus on. Another name for cloud Computing is a distributed computing over the Network. Cloud computing doesn't require to deploy the application on local computer as it directly delivered the hosted services over the internet. The objective of the paper is to study the Map-Reduce programming model and the Hadoop development platform of cloud computing and to ensure efficiency of Web mining using these parallel mining algorithms.

Keywords: Data Mining, Web mining, Cloud Computing, map-reduce

I. Introduction

A) Web Mining

Extensive version of data mining can be termed as web mining. On web data is stored in a heterogeneous manner in a semi-structured or unstructured form due to which mining on web is difficult as compared to traditional data mining. Web data mining is used to extract useful information or facts from Web Usage logs[2], Web Hyperlinks, Web Page contents. Different types of web Mining are:

- Web structure Mining
- Web Content Mining
- Web Usage Mining [4]

The process of extracting the information on Web is called Web content mining. In Web Mining, data collection is a substantial task especially for Web Structure and Web content mining, and involves crawling a large number of Web pages[3]. The Internet has today changed computing to distributed computing or cloud computing. All the major Social Media sites: Twitter, Facebook, Linked In, and Google+ contains abundance of information are today on cloud platform. For instance Tweets happen every millisecond on Twitter, they happen at the "speed of thought". This data is available for consumption all the time. The data on Twitter ranges from small tweets to long conversational dialogues to interest graphs etc. Now which data mining technique to apply, how to find association or correlation or how to cluster the data based on their similarity, so as to gain efficiency in the platform of cloud computing is the research area.

Tripti Lamba*
Research Scholar
Jagan Nath University, Jaipur, India
Leena Chopra**
Research Scholar
Amity Univesity, Noida, India

Problems associated with Web Mining

1. **Scalability:** The database is huge and it contains large dataset so mining interesting rules adds on to uninterested rules that are huge. There is no efficient algorithm for extracting useful pattern from the huge database.
2. **Type of Data**: The data on Web is heterogeneous[5]. Web cleaning is the most important process and is very difficult for semi structured data and unstructured data. According to researchers 70% of the time is spent on data pre-processing.

3. **Efficiency:** Mining rules from semi structure and unstructured as in the semantic web is a great challenge. Lot of time and memory consumption leads to decreased efficiency.
4. **Security:** The data on web is accessed publicly. There is no data that is hidden, so this is another challenge in Web Mining.

B) Cloud Computing

The computer resources these days are consumed as utility by various companies the same manner one consumes electricity or a rented house. There is no need to fabricate and retain computing infrastructures in-house. There are three types of cloud private, public and hybrid. Cloud services are mainly categorized into three types: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service(IaaS)[8]. There are various benefits of Cloud, some of which are mentioned below:

- **Self-service provisioning:** It all depends on the end users, which type of services they yearn for. Users can revolve around multiple computing assets for almost any type of workload on-demand.
- **Elasticity:** Companies can scale up as computing needs increase and then scale down again as demands decrease.
- **Pay per use:** There is a flexibility of using the services and computing resources as per the need of demand of the user. This facility permits users to pay only for the resources and workloads they utilize.

Cloud computing is most impressive technology because it is cost efficient and flexible. Cloud Mining's Software as Service (SaaS) is used for implementing Web Mining, as it reduces the cost and increases the security. Compared to all the other web mining techniques, Web usage mining is immeasurably used and have known productive outcomes[7].

C) Web Mining and Cloud Computing

One of the mostly used technologies in Web Mining is Web Usage Mining[1]. Web Usage mining using Cloud Computing is majorly adopted these days due to its reduced cost efficiency and flexibility[6]. However, in spite of improved movement and attention, there are considerable, continual concerns about cloud computing that ultimately compromise the vision of cloud computing as a new IT procurement model. Fundamentally Cloud Mining is novel approach to faced search interface for your data. The major challenge which is a security of web mining is been offered by SaaS (Software-as-a Service) and used for dropping the cost which is termed as cloud mining technique. It's been targeted to change the existing framework of web mining to generate an influential framework by Hadoop and map Reduce communities for projecting analytics. [9]

In the next section we have discussed how to use Map/ Reduce Model in Cloud Computing and what are the various benefits of using this model.

II. Cloud Computing and Map/ Reduce Model

The term cloud is a representation designed for the Internet, an intellection of the Internet's fundamental infrastructure that helps to spot the point at which accountability moves from the user to an external provider. Cloud Computing is one of the most captivating areas where lots of services are being utilized. The main objective of Cloud computing is to fully utilize the resources dispersed at various places[10]. Map/ Reduce model which is a programming model, proposed by Google is used for processing voluminous data sets. Map/Reduce Model processes around 20 petabytes of data in a single day. This model is gaining more popularity in cloud computing these days[11][12]. Map/ Reduce model is used for parallel and disseminated processing of huge data sets on clusters[13]. Some of the applications of Map/Reduce are:

At Google:

- Index building for Google Search
- Article clustering for Google News
- Statistical machine translation

At Yahoo!:

- Index building for Yahoo! Search
- Spam detection for Yahoo! Mail

At Facebook:

- Ad optimization
- Spam detection

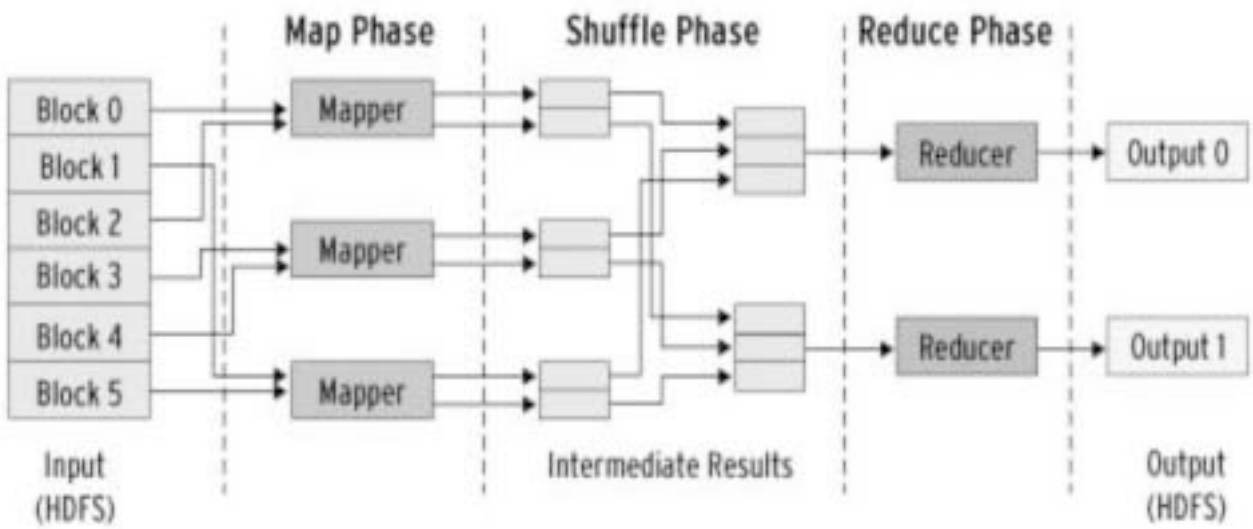


Fig. 1 Map/ Reduce System Framework[14]

A) Advantages of Map/Reduce Framework:

The main advantage of the MapReduce framework is its fault tolerance, where periodic reports from each node in the cluster are expected when work is completed. A task is transferred from one node to another. If the master node notices that a node has been silent for a longer interval than expected, the main node performs the reassignment process to the frozen/ delayed task. Some of the advantages [15] of Map/ Reduce Framework are mentioned below:

Scalability and Distributed Processing: Hadoop platform that utilizes Map/Reduce framework is extremely scalable. It has the capability to accumulate and distribute large data sets across ample of servers which operates in parallel which leads to reduced cost.

Flexibility: It operates on Structured and Unstructured data from variety of sources like email, e-commerce, social media, etc.

Fast: This framework works on Distributed architecture so huge amount of data ranging from Terabytes to petabytes. It takes minutes to process terabytes of data, and hours for petabytes of data.

Security and Authentication: Security is the major area of concern in almost every field. MapReduce works with HDFS and HBase security which allows only access to only authenticated users.

B) Map/ Reduce System Framework

The basic architecture of Map/Reduce is mentioned in Fig. 1[14] Map/ Reduce involve two basic steps:

- Map: performs filtering and sorting and
- Reduce :performs a summary operation

The input and output are in the form of key-value pairs. After the input data is partitioned into splits of appropriate size, the map procedure takes a series of key-value pairs and generates processed key-value pairs, which are passed to a particular reducer by a certain partition function; later after the data sorting and shuffling, the reduce procedure integrates the results. The scalability achieved using MapReduce to implement data processing across a large volume of CPUs with low implementation costs, whether on a single server or multiple machines, is a smart proposition.

III. Conclusion

Cloud Computing is definitely one of the widely used technologies as it is cost efficient and flexible. Web Usage Mining uses Cloud Computing Service SaaS (Software as a Service) to increase the security and reduce the cost. In this paper we have discussed the basic Map/Reduce model and its advantages. The future work will focus on new ways to improve the current model so as to aim at more accurate and faster approach for Web Usage mining, based on Cloud Computing.

References

1. M. U. Ahmed and A. Mahmood, "Web usage mining:," International Journal of Technology Diffusion, vol. 3, no. 3, pp. 1–12, Jul. 2012.
2. S. K. Pani, et.al L "Web Usage Mining: A Survey On Pattern Extraction From Web Logs", International Journal Of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.
3. Singh, Brijendra, and Hemant Kumar Singh. "Web data mining research: a survey." In Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, pp. 1-10. IEEE, 2010.
4. J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.
5. Li, J., Xu, C., Tan, S.-B, "A Web data mining system design and research". Computer Technology and Development 19: pp. 55-58, 2009
6. Robert Grossman , Yunhong Gu, "Data mining using high performance data clouds: experimental studies using sector and sphere", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2008
7. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining," ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, p. 12, Jan. 2000.
8. Khanna, Leena, and Anant Jaiswal. "Cloud Computing: Security Issues And Description Of Encryption Based Algorithms To Overcome Them." International Journal of Advanced Research in Computer Science and Software Engineering 3 (2013): 279-283.
9. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using modelbased clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pages 280{284, Boston, Massachusetts, 2000.
10. Zhu, W., & Lee, C. (2014). A new approach to web data mining based on cloud computing. Journal of Computing Science and Engineering, 8(4), 181–186. doi:10.5626/jcse.2014.8.4.181
11. "MapReduce." Wikipedia. N.p.: Wikimedia Foundation, 11 Jan. 2017. Web. 2 Jan. 2017.
12. Divestopedia, and Securities Institute. What is MapReduce? - definition from Techopedia. Techopedia.com, 2017. Web. 2 Jan. 2017.
13. Posted, and Margaret Rouse. What is MapReduce? - definition from WhatIs.com. SearchCloud Computing, 25 June 2014. Web. 2 Jan. 2017.
14. Hornung, T., Przyjaciel-Zablocki, M., & Schätzle, A. (2017). Giant data: MapReduce and Hadoop » ADMIN magazine. Retrieved January 10, 2017, from http://www.admin-magazine.com/HPC/Articles/MapReduce-and-Hadoop
15. Lee, K.-H., Lee, Y.-J., Choi, H., Chung, Y. D., & Moon, B. (2012). Parallel data processing with MapReduce. ACM SIGMOD Record, 40(4), 11. doi:10.1145/2094114.2094118

Role of Cloud computing in the Era of cyber security

Shilpa Taneja*
Vivek Vikram Singh**
Dr. Jyoti Arora***

Introduction

Cloud computing is taking the IT landscape further away from the organization. There are numerous benefits of cloud based system where software is managed and upgraded. Cost of hardware is very low as it requires only internet connection and browser, so other hardware devices become unnecessary. Cloud computing in simplification is considered as a form of outsourcing. With this the major issue is lying with most important asset for any organization i.e. information. Most of the IT organizations are losing control of their technology. As the cloud computing is emerging so as the cyber security trends of today are evolving at high speed pace. Prediction and detection of attack in cyber security is the shifting of incident response which is a continuous process. It generates the requirement of a security architecture that integrates prediction, prevention, detection and response. Cloud computing in cyber security provides the advantages of a public utility system in aspect of economic, flexibility and convince; but simultaneously raises the issue on security and loss of control. This paper presents the user centric measure of cyber security and provides the comparative study on different methodology used for cyber security.

Cloud computing in cyber security

Cloud computing provides high level of security and uptime than typical network. It is the simplest form of outsourcing. There are numerous benefits of cloud based system. Cost of hardware is lowers down and on the offside software is managed and upgraded. It saves cost and time as it controls the buying and upgrading of servers and other hardware. It diminishes the requirement of large IT staff. It provides faster time to market and increased employee productivity. Cloud computing provide the next generation of IT resources through a platform which is scalable and easy to manage the local area network. The legal system is running behind to adopt cloud computing. As most of the cloud vendors donot take responsibility for data loss, downtime or loss of revenue caused by cyber-attacks there is a need of taking preventive as well as corrective measures for solving the problem. According to foster, the cloud computing market will have a tremendous growth of $191 billion by 2020 which is $91 in 2015.

Risks to cloud computing

The study has revealed the 9 cloud risks. It follows high profile breaches of cloud platform evernote, adobe creative cloud, slack and lastpass. The lastpass breach is problematic as it stores all of user's website and cloud service password. It is protected with password especially those belonging to administrator with extensive permission for a company's critical infrastructure, a critical criminal could launch a devasting attack.

1. Loss of intellectual property

Cyber criminals are benefited by gaining the access on sensitive data. Skyhigh in its report says that21% of the uploaded files share services contains responsive data. A few services can even pose risk if the terms and conditions claim ownership of data uploaded to them.

2. Compliance violations and regulatory actions

Most of the companies these days follow some regulatory control of their information being it is about health information or student record. It becomes requirement for the companies to know about the location of their data and about its protection. It is also required to know about the person who will access it.

Shilpa Taneja*
Assistant Professor, IITM
Vivek Vikram Singh**
Assistant Professor, IITM
Dr. Jyoti Arora***
Assistant Professor, IITM

3. Loss of control over end user actions

Employees can harm the company by downloading a report of all customer contacts, upload the data to a personal cloud storage service and then access that information once he left the company and joins some competitor. It can be misused when companies are in dark about the working moment of their employees. It is one of the more common insider threats today.

4. Malware infections that unleash a targeted attack

Cloud services are the vector of data exfiltration. Study reveals that a novel data exfiltration technique is that where attackers encoded sensitive data into video files and uploaded them to social media. There are numerous malware that exfiltrates sensitive data via a private social media accounting the case of the Dyre malware variant, cyber criminals used file sharing services to deliver the malware to targets using phishing attacks.

5. Contractual breaches with stake holders

Contracts among business parties often restrict how data is used and who is authorized to access it. When employees move restricted data into the cloud without authorization, the business contracts may be violated and legal action could ensue. The cloud service maintains the right to share all data uploaded to the service with third parties in its terms and conditions, thereby breaching a confidentiality agreement the company made with a business partner.

6. Diminished trust of customer

Data breaches results in diminished trust of customers. The biggest breach reported was that where cyber criminals stole over 40 million customer credit and debit card numbers from different Target. The breach led customers to stay away from Target stores, and led to a loss of business for the company, which ultimately impacted the company's revenue.

7. Data breach requiring disclosure and notification to victims

If sensitive or regulated data is put in the cloud and a breach occurs, the company may be required to disclose the breach and send notifications to potential victims. Certain regulations like the EU Data Protection Directive require these disclosures. Following legally-mandated breach disclosures, regulators can levy fines against a company and it's not uncommon for consumers whose data was compromised to file lawsuits.

8. Increased customer churn

If customers even suspect that their data is not fully protected by enterprise-grade security controls, they may take their business elsewhere to a company they can trust. A growing chorus of critics is instructing consumers to avoid cloud companies who do not protect customer privacy.

9. Revenue losses

According to the Ponemon BYOC study, 64% of respondents confirmed that their companies can't confirm if their employees are using their own cloud in the workplace. In order to reduce the risks of unmanaged cloud usage, companies first need visibility into the cloud services in use by their employees. They need to understand what data is being uploaded to which cloud services and by whom. With this information, IT teams can begin to enforce corporate data security, compliance, and governance policies to protect corporate data in the cloud. The cloud is here to stay, and companies must balance the risks of cloud services with the clear benefits they bring.

In this era of digitization, data security is paramount to every business. In past, on-premise servers were the business technology model, but now there are more choices. For the last several years, a debate has flowed through businesses. How will cloud computing affect them? Should they adopt a public cloud approach, opt for private cloud, or stick with their on-premise servers? The use of cloud computing is steadily rising. In fact, a recent study has shown that cloud services are set to reach over $130 billion by 2017. Before making any decisions, it's important to think about how this shift towards cloud computing will affect cyber security for your business.

Measures or models of cloud computing in cyber security

Boehm et al. poised that all dilemmas that arise in software engineering are of an economic nature rather

than a technical nature, and that all decisions ought to be modeled in economic terms: maximizing benefit; minimizing cost and risk. Their work is perfectly compatible with the philosophy of value-based software engineering, as it models system security not by an arbitrary abstract scale but rather by an economic function (MFC), quantified in monetary terms (dollars per hour), in such a way as to enable rational decision making.

Brunette and Mogull (2009) discuss the promise and perils of cloud computing, and single out security as one of the main concerns of this new computing paradigm. They have cataloged and classified the types of security threat that arise in cloud computing. Their work can be used to complement and provides a comprehensive catalog of security threats that are classified according to their type.

Black et al. (2009) discussed about categorization of metrics and measures and among different type of metrics. These metrics can be used as standard by organization to compare between current situations and expected one. This provides the organization facility to raise the level in order to meet the goal.

Jonsson and Pirzadeh (2011) proposed a framework to measure security by regrouping the security and dependability attributes on the basis of already existing conceptual model applicable on application areas varying from small to large scale organization. They discussed how different matrices are related to each other. They categorize the security metric into protective and behavior metrics. Choice of measures affect the results and accuracy of a metric.

Carlin and Curran (2011) founded that using cloud computing companies can decrease the budget by 18%. The findings comprise mainly three services Software-as-a-service (SaaS), Platform-as-a-service (PaaS) and Infrastructure-as-a-service (IaaS). Three kinds of model public private and hybrid, encryption is not a way to fully protect the data.

Chow et al. (2009) discusses the three types of security concern raised in cloud computing- provider-related vulnerabilities, which represent traditional security concerns; availability, which arises in any shared system, and most especially in cloud computing; and third party data control, which arises in cloud computing because user data is managed by the cloud provider and may potentially be exposed to malicious third parties. They also discuss strategies that maybe used to mitigate these security concerns.

Center for Internet Security (2009)used mean time to incident discovery, incident rate, mean time between security incidents, mean time to incident recovery, vulnerability scan coverage, percentage of systems without known severe vulnerabilities, mean time to mitigate vulnerabilities, number of known vulnerability instances, patch policy compliance, mean time to patch and proposed a set of MTTF-like metrics to capture the concept of cyber security.

Benefits of Cyber security in Cloud Computing

Cyber security has numerous benefits in cloud based applications like improvement in gathering and threat model, enhanced collaboration, reduction of lag time between detection and remediation. With the increase in cyber-attacks in era of cloud computing organization need to take precautions and adequate measures to deal with threats. The four pillars of cloud based cyber security comprise updated Technologies, extremely protected platforms, skilled manpower and high bandwidth connectivity. Learning collection can support real time integrated security information. Usage of cyber security ensures that security while maintaining sensitive data. The concept of out-of-band channels can be used to deal with cyber-attacks. 41% of business employ infrastructure-as-a-service (IaaS) for mission-critical workloads. Cloud-based cyber security solution developed by PwC and Google can provide advanced detection, analysis, collective learning, high performance, scalability in analytic processes to enable an advanced security operations capability (ASOC).This will create honeypots and dummies for maintaining connection to end point for analysis and learning.

Conclusion

This paper discusses about numerous benefits of cloud based system and various risks related to it. We also discussed the various models which talks about how to maximize the benefits, minimizing cost and risks. On the basis of classification of metrics and measures

of cloud computing we can facilitate organization to raise the efficiency and to meet their goals. Various strategies maybe used to mitigate these security concerns. At last we can say that usage of cyber security ensures security while maintaining sensitive data as well.

References

1. Rabia, L., Jouini, M., Aissa, A., Mili, A., 2013. A cybersecurity model in cloud computing environments. Journal of King Saud University –Computer and Information Sciences.
2. Boehme, R., Nowey, T., 2008. Economic security metrics. In: Irene, E.,Felix, F., Ralf, R. (Eds.), Dependability Metrics, 4909, pp. 176–187.
3. Brunette, G., Mogull, R., 2009. Security guidance for critical areas offocus in cloud computing V 1.2. Cloud Security Alliance.
4. Black, P.E., Scarfone, K., Souppaya, M., 2009. Cyber Security Metricsand Measures. Wiley Handbook of Science and Technology forHomeland Security.
5. Jonsson, E., Pirzadeh, L., 2011. A framework for security metricsbased on operational system attributes. In: International Workshopon Security Measurements and Metrics – MetriSec2011,Bannf, Alberta, Canada.
6. Carlin, S., Curran, K., 2011. Cloud computing security. InternationalJournal of Ambient Computing and Intelligence.
7. Chow, R., Golle, P., Jakobsson, M., Shi, E., Staddon, J., Masuok, R.,Molina, J., 2009. Controlling data in the cloud: outsourcingcomputation without outsourcing control. In: ACM Workshop onCloud computing Security (CCSW).
8. The Center for Internet Security, The CIS Security Metrics v1.0.0, 2009. <https://www.cisecurity.org/tools2/metrics/CIS_Security_Metrics_v1.0.0.pdf>.

Cryptography and its Desirable Properties in terms of different algorithm

Mukta Sharma*
Dr. Jyoti Batra Arora**

Abstract

The proliferation of Internet has revolutionized the world. The world has become a smaller place to communicate. Especially in India, after demonetization Indian government is encouraging both customer and buyer to transact online (go cashless). Electronic payment is a new trend to transact online as any e-commerce environment needs a payment system. Payment system requires an intricate design which ensures payment security, transaction privacy, system integrity, customer's authentication, and purchaser's promise to pay and supplier promise to sell a high-quality product. There are several e-payments systems like paying via Plastic money (credit/debit/smart card), e-wallet, e-cash, UPI, Net banking, Aadhaar Card, etc. Electronic payment is made online without face to face interaction, which leads to electronic frauds. Therefore, the emphasis is given on security methods opted by banks especially on cryptography.

This paper begins with the primary security threats, followed by the prevention plan. It highlights the cryptography and discusses the desirable property to check the strength of encryption algorithm.

Keywords: Avalanche, Cryptography, Decryption, Encryption, Cipher Text, DES, Plain Text, Symmetric Cryptography

I. Introduction

With the technological advancement, everyone is using the Internet on their smart phones, laptops, desktops, iPads, etc. Users are transacting funds online. E-banking is growing phenomenally well. There are numerous advantages of using online banking from both customers and bankers' perspective such as cost-effective, paperless, immediate transfer of funds, geographical convenience, 24*7, etc. Several issues in internet banking are security, trust, authentication, Non-repudiation, privacy and availability. Since the inception of e-banking security is and always will remain a matter of great concern. After the development of e-banking, the bank needs to ensure payment security, transactions privacy, system integrity, customer authentication as it is a payment system online.

Every coin has two facets with the internet having numerous advantages it has significant security threats.

Mukta Sharma*
Research Scholar, TMU

Dr. Jyoti Batra Arora**
Assistant Professor, IITM

Customers are reluctant to share their demography especially financial details online because of the security concerns. The need for the safety means to prevent unwanted access to confidential information. Cybercriminals steal sensitive data and misuse it for their benefits.

II. Security Threats

Electronic transactions have been facing various obstacles with context to security. Crimes like hacking, cracking, phishing; DOS, etc. are among few attacks or threats for the safety. Following attacks breach the security:

a) *Cracking / Hacking*- It defined as the unauthorized access to someone else information.

b) *Denial of Service attack*- DoS floods the computer with more requests than it can handle causing the web server to crash. Denying authorized users the service offered by the resource. Distributed Denial of Service (DDoS) attack wherein the perpetrators are many and are geographically widespread. Controlling such attacks is tough. The attack is initiated by sending excessive demands to the

victim's computer(s), exceeding the limit that the victim's servers can support and making the server's crash.

c) *E-mail spoofing*- A spoofed e-mail is one, which misrepresents its origin. It shows its origin to be different from which it originates.

d) *Phishing*- It is another criminally fraudulent process, in which a fake website resembling the original site is designed. Phishing is an attempt to acquire sensitive information such as usernames, passwords and credit card details, by masquerading as a trustworthy entity in an electronic communication.

e) *Salami Attack*- is an attack which is difficult to detect and trace, also known as penny shaving. The fraudulent practice of stealing money repeatedly in small quantities, usually by taking advantage of rounding to the nearest cent (or other monetary units) in financial transactions.

f) *Virus / Worm Attacks* – Malicious Programs are dangerous may it be Viruses, worms, logic bombs, trap doors, Trojan Horse, etc. As they are programs written to infect and harm the data by altering or deleting the information, or by making a backdoor entry for unauthorized person.

g) *Forgery*- Counterfeit currency notes, postage, and revenue stamps, mark sheets, etc. can be forged using sophisticated computers, printers, and scanners.

III. Security Measures

Security has become a necessity, and need to keep data safe, achieve it and many techniques are available. By using these techniques, one can ensure the confidentiality, authentication, privacy and integrity of their information. Information can be of any type; may it be in the form of text, image, audio or video. The need for security means to prevent unwanted access to confidential information, this can be attained by the following ways:-

a) *SSL*- Secure Socket Layer is a protocol developed by Netscape. It was designed so that sensitive data can be transmitted safely via the Internet. SSL creates a secure connection between a client and a server, over which any amount of data can be sent securely. All browsers support SSL, and many Web sites use the protocol to obtain confidential user information, such as credit card numbers.

b) *HTTPS*- Hyper Text Transfer Protocol combined with SSL to ensure security. S-HTTP is designed to transmit individual messages securely. SSL and S- HTTP, can be seen as complementary rather than competing technologies. Both protocols have been approved by the Internet Engineering Task Force (IETF) as a standard.

c) *Firewall*- Firewalls can be implemented in both hardware and software, or a combination of both to prevent unauthorized access. Firewalls are frequently used to prevent unauthorized Internet users from accessing private networks connected to the Internet, especially intranets. All messages are entering or leaving the intranet pass through the firewall, which examines each message and blocks those messages that do not meet the specified security criteria.

d) *SET*- Secure Electronic Transaction is a standard developed jointly by Visa International, MasterCard, and other companies. The SET protocol uses digital certificates to protect credit card transactions that are conducted over the Internet. The SET standard is a significant step towards securing Internet transactions, paving the way for more merchants, financial institutions, and consumers to participate in electronic commerce.

e) *PGP*- Pretty Good Privacy provides confidentiality by encrypting messages to be transmitted or data files to be stored using an encryption algorithm. PGP uses the "public key" encryption approach - messages are encrypted using the publicly available key, but can only be deciphered by the intended recipient via the private key.

f) *Anti-Virus*- To secure PC, laptop, smartphone from any malicious attack the user must install a good anti- virus and always update the anti-virus software fortnightly for better security.

g) *Steganography*- It is the process of hiding a secret message with an ordinary message. The original

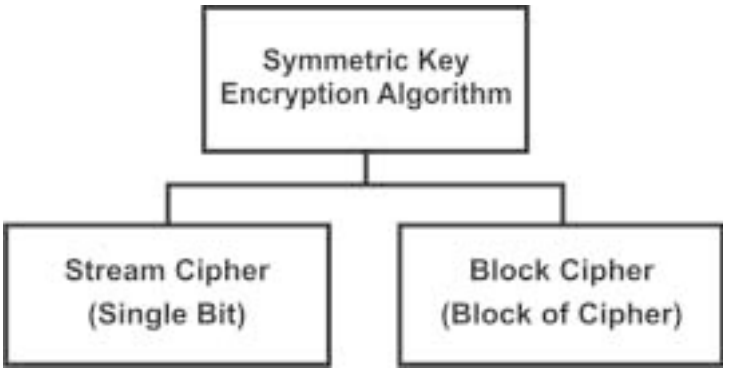


Figure 1: Symmetric Key Encryption Algorithm

user will view the standard message and will fail to identify that the message contains a hidden or encrypted message. The secret message can be extracted by only the authentic users who are aware of the hidden message beneath the ordinary file. Steganography is now gaining popularity among the masses because of ease of use and abundant tools available.

h) *Cryptography*- It is the "scrambling" of data done using some mathematical calculations and only authentic user with a key and algorithm can "unscramble" it. It allows secure transmission of private information over insecure channels.

IV. Cryptography

Cryptology is the study of reading, writing, and breaking of codes. It comprises of cryptography (secret writing) and cryptanalysis (breaking code). Cryptography is an art of mangling information into apparent incomprehensibility in a way permitting a secret method of unscrambling [11]. Human has a requirement to share private information with only intended recipients. Cryptography gives a solution to this need.

Cryptographic algorithms play a significant role in the field of network security. To perform cryptography, one requires the secure algorithm which helps the conversion efficiently, securely if carried out with a key. Encryption is the way to transform a message so that only the sender and recipient can read, see or understand it. The mechanism is based on the use of mathematical procedures to scramble data so that it is tough for anyone else to recover the original message.

There are two basic types of cryptosystems such as symmetric cryptosystems and asymmetric cryptosystems. Symmetric cryptography is a concept in which both sender and receiver shares the same key for encryption and decryption process. In contrast to symmetric cryptography, asymmetric cryptography uses a pair of keys for encryption and decryption transformations. The public key is used to encrypt data, and the private key is used to decrypt the message.

1) Symmetric Key Encryption Algorithms

Symmetric Key is also known as a private key or conventional key; shares the unique key for transmitting the data safely. The symmetric key was the only way of enciphering before the 1970s. Symmetric Key Encryption can be performed using Block Cipher or Stream Cipher.

Stream Cipher takes one bit or one byte as an input, process it and then convert it into 1bit or 1-byte cipher-text. Like RC4 is a stream cipher used in every mobile phone.

Block Cipher works with a single block or chunks of data or message instead of a single stream, character, or byte. Block ciphers mean that the encryption of any plaintext bit in a given block depends on every other plaintext bit in the same block. Like DES, 3DES have a block size of 64 bits (8bytes), and AES has a block size of 128 bits (16 bytes).

2) Need for Cryptography

It has given a platform which can ensure not only confidentiality but also integrity, availability, and non-repudiation of messages/ information. Symmetric Key

encryption algorithm focuses on privacy & confidentiality of data.

3) Symmetric Key Block Cipher Algorithm

The paper focuses on Symmetric Key block ciphers. DES, 3DES, AES, IDEA, Blowfish are among most used and popular algorithms of Block ciphers.

a) *DES*- DES is based on Feistel network. It takes 64 bit Plain Text as an input and 64 bit Cipher Text comes as an output. Initially a 64 bit Key is sent which is later converted to 56 bits (by removing every 8th bit). Later using 16 iterations with permutation, expansion, substitution, transpositions and basic mathematical functions encryption is performed and decryption is the reverse process of encryption.

b) *3DES* – Triple DES is an enhancement of Data Encryption Standard. To make it more secure the algorithm execute three times with three different keys and 16*3=48 rounds; and a key length of 168 bits (56*3) [22]. The 3DES encryption algorithm works in a sequence Encrypt-Decrypt-Encrypt (EDE). The decryption process is just reverse of Encryption process (Decrypt- Encrypt-Decrypt). 3DES is more complicated and designed to protect data against different attacks. 3DES has the advantage of reliability and a longer key length that eliminates many attacks like brute force. 3DES higher security was approved by the U.S. Government. Triple DES has one big limitation; it is much slower than other block encryption methods.

c) *IDEA*-International Data Encryption Algorithm is another symmetric key block cipher algorithm developed at ETH in Zurich, Switzerland. It is based on substitution-permutation structure. It is a block cipher that uses a 64 bit plain text, divided equally into 16 bits each (16*4=64); with 8 and s half rounds and a Key Length of 128-bits. For each round 6 sub keys are required 4 before the round and 2 within the round (8*6= 48 sub keys+ 4 sub keys are used after last or eighth round that makes total 52 sub- keys). IDEA does not use S-boxes. IDEA uses the same algorithm in a reverse order for decryption [2] [21].

d) *AES*- AES is also a symmetric key algorithm based on the substitution–permutation Network [4][7][23].

AES use a 128-bit block as plain text, which is organized as 4*4 bytes array also called as State and is processed in several rounds. It has variable Key length 128, 192 or 256-bit keys. Rounds are variable 10, 12, or 14 depends on the key length (Default # of Rounds = key length/32 + 6). For 128 bit key, number of rounds are 10; 192 bit key, 12 rounds and for 256 bit key, 14 rounds. It only contains a single S- box (which takes 8bits input, and give 8 bits output) which consecutively work 16 time. Originally the cipher text block was also variable, but later it was fixed to 128 bits.

The Encryption and decryption process consists of 4 different transformations applied consecutively over the data block bits, in a fixed number of iterations, called rounds. The decryption process is direct inverse of the encryption process. Hence the last round values of both the data and key are first round inputs for the decryption process and follows in decreasing order. AES is extremely fast and compact cipher. For implementers its symmetric and parallel structure provides great and an effective resistance against cryptanalytic attacks. The larger block size prevents birthday attacks and large key size prevents brute force attacks

e) BlowFish- It is a symmetric block cipher and works on of 64-bit block size. Key length is variable from 32 bits to 448 bits. It has16 rounds and is based on Feistel network. It has a simple structure and it's easy to implement. It encrypts data on 32 bit microprocessors at a rate of 18 clock cycles per byte so much faster than AES, DES, and IDEA. Since the key size is large it is complex to break the code in the blowfish algorithm. It is vulnerable to all the attacks except the weak key class attack. It is unpatented and royalty-free. It requires less than 5K of memory to run Blowfish [6] [18].

IV. Comparative Analysis

	DES	3DES	IDEA	AES	Blowfish
Avalanche effect	Resists	Resists	Resists	Resists	Resists
Completeness	Yes	Yes	Yes	Yes	Yes
Statistical Independence	Yes	Yes	Yes	Yes	Yes

Modern Encryption				
Algorithm	**Plain Text**	**Cipher Text**	**Avalanche Effect**	**Reference**
AES	11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11	79 f8 cc 24 01 82 dd 7f 2d 89 f7 e7 78 b7 ee 30	43.75% (56)	[9]
	11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 10	9d 4c 1d b4 6a 93 27 b5 20 64 37 d1 3d 9d 2a		
	11 22 33 66 55 44 55 44 77 88 99 66 44 45 36 12	4a a9 16 11 e2 8a 9f 67 35 30 1f 80 16 c5 b7 cd	51.53% (66)	
	11 22 33 66 55 44 55 44 77 88 99 66 44 45 36 11	D7 00 43 2d 51 78 f7 65 50 03 03 75 b1 e4 2d a0		
	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	C6 a1 3b 37 87 8f 5b 82 6f 4f 81 62 a1 c8 79	44.53% (57)	
	10 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	0d 19 33 06 27 42 fe 01 9c fe 06 e1 a8 1a a0 01		
Blow fish	ADF278565E262AD1F5DEC94A0BB2527		42% (27) 46.09% (59)	[10] [8]
	ADF278565E262AD1F5DEC94A0BF25B27		99.6126% different pixel	[19]
DES	0100010001001001010100110100000101 0100110101010001000101010100010	0101011110100101000001 0011011011101100010101 1101100111000010101 1	54.36% (35)	[16]
	0100010010101001001001100000110101 0011010101000100010101010010	1111101101010100010010 0100101111111011101000 0110100111010111011 1		

Classical Encryption				
Caesar Cipher	ABCD	DEFG	1.56% (1)	[15]
	ABED	DEHG		
Viginere Cipher	DISASTER	IIMZSGGV	3.1% (2)	[15][16]
	DISCSTER	IIMBSGGV		
Play fair Cipher	DISASTER	ELPNOYDP	10.9% (7)	[15][16]
	DISCSTER	ELOGOYDP	6.75% (4)	

Table 1 : Comparative Analysis

V. Algorithm Security

The two essential properties to check the complexity of any algorithm is time and space. According to Kerckhoff, the cryptanalyst knows the complete process of encryption and decryption except for the value of the secret key. It implies that the security of a secret-key cipher system rests entirely on the secret key [17]. Therefore, for better security in symmetric encryption one should keep the following criteria's in mind:

- Key should be exchanged very safely because if the key is known the entire algorithm is compromised.
- A secure encryption algorithm is robust & resilient against a potential breach using combinations of cipher texts & key [14] [20].

1) Desirable Properties of Block Cipher

The strength of a block cipher can be tested through these properties like Avalanche, Completeness and Statistical Independence.

- Avalanche Effect- It is an excellent property of cryptographic algorithm also stated as Butterfly effect. It means that by changing only one bit (small change) of the plain text or the key should produce a radical shift in the final output. If the final output is modified or flipped with 50% of bits, then it is said to be strict Avalanche effect. SAC is harder to perform an analysis on cipher text when trying to come up with an attack [5] [8] [17]. It's easy to impose conditions on Boolean functions so that they satisfy certain avalanche criteria, but constructing them is a harder task. Avalanche can be categorized as follows:
  - The strict avalanche criteria (SAC) guarantee that exactly half of the output bits change when one input bit changes [17].
  - The bit independence criterion (BIC) states that output bits j and k should change independently when any single input bit i is inverted, for all i, j and k[17].

  Avalanche Effect= Number of flipped bits in ciphered text/ Number of bits in ciphered text.
- Completeness -According to encryption, this is a necessary property. Completeness means that each bit of the cipher text/ output block needs to depend on each bit of the plaintext [15]. Change in one bit of the input (plaintext) will bring change in every bit of the output (Ciphertext). It has an average of 50% probability of changing.

  Let us imagine an eight-byte plain text, and there is a change in the last byte, it would only have affected the 8th byte of the Ciphertext. An attacker can very easily guess 256 different plaintext-Ciphertext pairs. Finding out 256 plaintext-Ciphertext pairs is not hard at all in the internet world, and standard protocols have standard headers and commands (e.g. "get," "put," "mail from:," etc.) which the attacker can safely guess.

  If the cipher has this property, the attacker need to collect 264 (~1020) plaintext-Ciphertext pairs to crack the cipher in this way.
- Statistical independence that input and output should appear to be statistically independent.

VI. Conclusion

Cryptography is a good way to protect data from getting breached. Symmetric cryptography ensures confidentiality of data. Asymmetric cryptography takes care of authenticity, integrity, non-repudiation of data. As can be seen in the above table of comparative analysis, where all the algorithms are built on these three desired properties. The percentages may vary but they all fulfil the basic criteria of an encryption algorithm. While building the understanding about the encryption algorithm and designing a new algorithm anybody can establish the significant role of thee building blocks.

These three important properties decide the strength and resistance of the algorithm.

References

1. Daemen, J., Govaerts, R. and Vandewalle, J. (1998). *Weak Keys for IDEA.* Springer-Verlag.
2. Engelfriet, A. (2012). *The DES encryption algorithm.* Available at www.iusmentis.com/technology/encryption/des.

3. Forouzan, B.A., &Mukhopadhyay, D. (2010). *Cryptography and Network Security*. Tata McGraw-Hill, New Delhi, India

4. Gatliff, B. (2003). *Encrypting data with the Blowfish algorithm.* Available at http://www.design-reuse.com/articles/5922/ encrypting-data-with-the-blowfish-algorithm.

5. Kak, A. (2015). *Computer and Network Security- AES: The Advanced Encryption Standard.*Retrieved from https://engineering.purdue.edu/kak/compsec/NewLectures/Lecture8.pdf

6. Koukou, Y.M., Othman, S.H., Nkiama, M. M. S. H. (2016). *Comparative Study of AES, Blowfish, CAST-128 and DES Encryption Algorithm.* IOSR Journal of Engineering, 06(06), pp. 1-7.

7. Kumar, A., Tiwari, N. (2012).*Effective Implementation and Avalanche Effect of AES.* International Journal of Security, Privacy and Trust Management (IJSPTM).

8. Mahindrakar, M.S. (2014). *Evaluation of Blowfish Algorithm based on Avalanche Effect*. International Journal of Innovations in Engineering and Technology, 1(4), pp. 99-103.

9. Menezes, A., Van, P., Orschot, O. and Vanstone, S. (1996). *Handbook of Applied Cryptography*, CRC Press.

10. Mollin, R.A. (2006). *An Introduction to Cryptography*. Second Edition, CRC Press

11. National Bureau of Standards (1997). *Data Encryption Standard.* FIPS Publication 46.

12. Paar, C., Pelzl, J. (2010). *Understanding Cryptography: A Textbook for Students and Practitioners'*. Springer, XVIII, 372.

13. Ramanujam, S., &Karuppiah, M. (2011). *Designing an algorithm with high Avalanche Effect*. International Journal of Computer Science and Network Security. 11(1).

14. Saeed, F., & Rashid, M. (2010). *Integrating Classical Encryption with Modern Technique*. International Journal of Computer Science and Network Security, 10(5).

15. Schneier B. (1994). *Applied Cryptography*. John Wiley& Sons Publication, New York.

16. Schneier, B. (1994).*Description of a New Variable-Length Key, 64-Bit Block Cipher (Blowfish), Fast Software Encryption*, Cambridge Security Workshop Proceedings, Springer-Verlag, 1994, Available at http://www.schneier.com/paper-blowfish-fse.html

17. Shailaja, S. & Krishnamurthy, G.N. (2014). *Comparison of Blowfish and Cast-128 Algorithms Using Encryption Quality, Key Sensitivity and Correlation Coefficient Analysis*. American Journal of Engineering Research, 7(3), pp. 161-166.

18. Stallings, W. (2011). *Cryptography and Network Security: Principles and Practice*. Pearson Education, Prentice Hall: USA

19. Thaduri, M., Yoo, S. and Gaede, R. (2004). *An Efficient Implementation of IDEA encryption algorithm using VHDL.* Elsevier

20. Tropical Software, Triple DES Encryption, Available at http://www.tropsoft.com/strongenc/des3.htm,

21. Wagner, R. N. *The Laws of Cryptography*. Retrieved From http://www.cs.utsa.edu/~wagner/laws/

A Review: RSA and AES Algorithm

Ashutosh Gupta*
Sheetal Kaushik**

Abstract

ARPANET to today's Internet, the amount of data and information increased to several thousand times. The amount of security problems are also increased with this development. In this paper we aim to review the working of two algorithms, RSA and AES to secure our data over the internet and communication channels. One of these algorithms is symmetric which is developed in early days of modern cryptography and other one is asymmetric, which is advance and still trustworthy.

Keywords: Asymmetric, symmetric, RSA, AES, Cryptography, Encryption.

I. Introduction

Cryptography Practice of the enciphering and deciphering of messages in secret code in order to render them unintelligible to all but the intended receiver. Cryptography may also refer to the art of cryptanalysis, by which cryptographic codes are broken [1].Information is the most important thing for a company or a nation to be secure after human resource. While most of the information now a days are in Digital form, they are equally in that much unsecured Environment[2].So, techniques like cryptography help in making the environment and the path of information travelling more secure and trustworthy. A good encryption algorithm must provide confidentiality, integrity, non- repudiation, and Authentication [3].

Cryptography can be further divided in two major types: Secret-Key Cryptography and public key cryptography.Secreate key encryption uses same key for encryption and decription.This type of encryption easier and faster but equally less secure. While on the other hand Public key encryption is more secure and most preferable now days. In this encryption key for encryption and decryption both are different but logically and mathematically they are linked [1][4] [5].

Ashutosh Gupta*
BCA-II Year
Institute of Information Technology and Management
Sheetal Kaushik**
IT Department
Institute of Information Technology and Management

A. Data Encryption

This is the process of scrambling, stored or transmitted information so that it is meaningless until it is unscrambled by the intended recipient. This is also known as Ciphering of data. With increasing data and technology advancement, the significance of data encryption is also increasing not only for highly diplomatic and military uses but also from life of ordinary men's to the high value money and information transfer of big multinationals[6].

The history of the cryptography can be traced back into hieroglyphs of early Egyptian civilization (c.1900 B.C.).Ciphering is always considered as the essence of diplomatic and military secrecy. There is several other example of cryptography even in the era of Holy Bible which replete with examples of ciphering [7].

Now a day's Encryption standards are increased so high that Several Government even talking about banning of strong encryption over certain level. The reason behind is the time consumption and work involved even in simple day to day federal cases. For example, the United Kingdom could pass a law that bans encryption stronger than 64-bit keys, knowing its intelligence agency has the resources to crack any form of legal encryption in the country [5].

The early cryptography is done with the standard algorithm of 64 bit key known as DES or Data Encryption Algorithm given by FIPS (Federal Information Processing Standard) [3], [8].

DES algorithm is further replaced by Rijndael algorithm and named as Advance encryption algorithm or AES [8], [9].AES has more flexible key strength that may be help in future manipulation for betterment of it.

RSA was named on their inventor names in 1977, Ron Rivest, Adi Shamir and Len Adleman[10].This algorithm is asymmetric and still in use. RSA algorithms have dual benefit as it used for data encryption as well as digital signatures.

II. AES

Now a Days Security is Equally essential as Speed of data communication and Advance Encryption standard has best suited for it as it provide speed as well as increase security with hardware. Because of its dual base which consists of hardware as well as software this System is more advance and secure than basic DES [8].

AES also advance in the sense of its structure as it uses key in bytes instead of bits whereas in DES number of rounds for encryption of data is not fixed, it depends on the size of the plain text it has to encrypt. If size of text is 128 bit it will treated as 16 Bytes and these 16 Bytes then arranged in form of 4x4 matrixes. In AES 10 rounds of encryption is performed for 128 bit key, 12 rounds for 192 bit keys, and 14 rounds for 256 bit keys. Following Algorithm Encrypt the data [11].

Step 1:- Input a plaintext of 128 bits of block cipher which will be negotiated as 16 bytes.

Step 2: - Add Round Key: - each byte is integrated with a block of the round key using bitwise XOR.

Step 3:- Byte Substitution: - the 16 input bytes are substituted by examining S- box. The result will be a 4x4 matrix.

Step 4:- Shift row: - Every row of 4x4 matrices will be shifted to left. Entry which will be left placed on the right side of row.

Step 5:- Mix Columns: - Every column of four bytes will be altered by applying a distinctive mathematical function (Galois Field).

Step 6:- Add Round Key: - The 16 bytes of matrix will be contemplated as 128 bits and will be XORed to 128 bits of the round key.

Step 7:- This 128 bits will be taken as 16 bytes and similar rounds will be performed.

Step 8:- At the 10th round which will be last round a ciphered text will be produced.

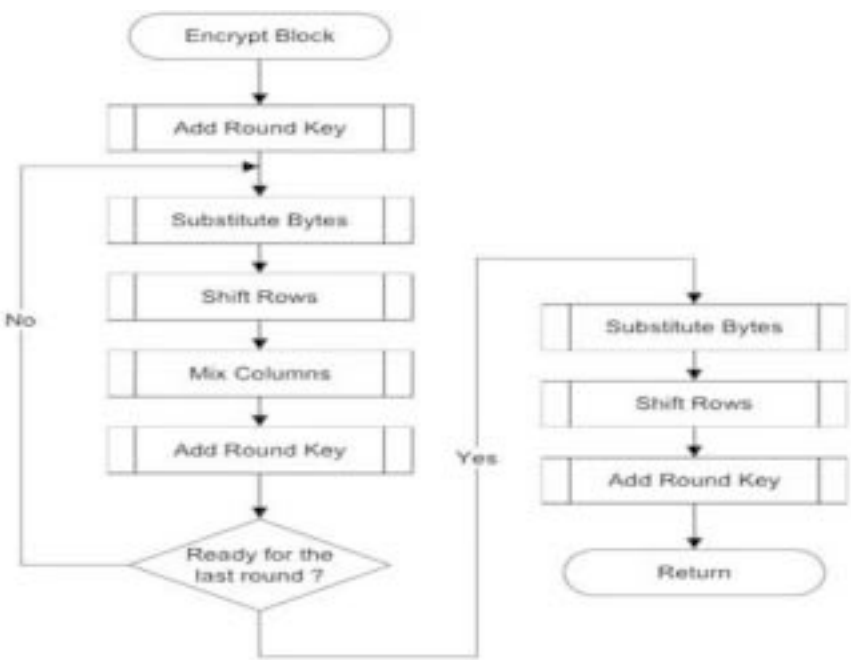


Fig.1 Flow Chart of AES Encryption.

III. RSA

RSA is a public key algorithm, means it uses two different keys one of which must be kept private know as private key and other is public key which is not essentially needed to be secret. Public Key from these two keys is usually used for encryption and private key is used for decryption [14].The RSA Encryption method is Explained Below:

Equations

Step 1: Select Two Large Prime number (Such that Number does not exceed printable ASCII Character).

Select Two Large Prime number p and q

Step 2: Generate the RSA modulus (The answer of multiplication will be considered the Key Length)

n=p*q(Public Key)

Step 3: Generate Random Key using Euler function.

e= (p-1) *(q-1)

Step 4: Form the public key

(n, e) form RSA public Key

Step 5: Generate the private key (Number d is the inverse of e modulo (p - 1) (q – 1).This means that d is the number less than (p - 1) (q - 1) such that when multiplied by e, it is equal to 1 modulo (p - 1) (q - 1))

ed = 1 mod (p H 1)(q H 1)

RSA security system depends on two different functions.RSA is one of the most secure Cryptography algorithm, whose difficulty is actually based on practical factoring of very large prime numbers [15][16].

IV. Comparison

In the below table the comparison is done between RSA and AES on the base of the keysize,block size, speed , key used in encryption and decryption, type of algorithm, round of encryption and decryption.[17]

FACTOR	AES	RSA
DEVELOPED	2000	1978
KEY SIZE	128,192,256 bits	>1024 bits
BLOCK SIZE	128 Bits	Minimum 512 bits
ENCRYPTION AND DECRYPTION	SAME	DIFFERENT
ALGORITHM	SYMMETRIC	ASYMMETRIC
SPEED	FASTER	SLOWER
ROUNDS	10/12/14	1

V. Conclusion

Encryption of Data plays very vital role in today's time. Our research work served the famous AES and RSA algorithm. Based on research work used in this survey, we can conclude that RSA takes more time for encryption compared to AES. We also concluded that the RSA is more secured than AES, because of its longer key size and different keys for encryption and decryption.

Our future work will be focused on the study of other algorithm including Hyper Image Encryption Algorithm. Our focus will also be on the path of transferring the private key of Asymmetric Encryption.

References

1. www.britannica.com/topic/cryptography.
2. ENISA's Opinion Paper on Encryption December 2016.
3. https://www.tutorialspoint.com/cryptography/data_encryption_standard. htm.
4. https://www.tutorialspoint.com/cryptography/cryptosystems.htm.

5. http://www2.itif.org/2016-unlocking-encryption.pdf.
6. http://www.infoplease.com/encyclopedia/science/data-encryption.html.
7. http://www.infoplease.com/encyclopedia/society/cryptography.html.
8. http://www.ijarcce.com/upload/2016/march-16/IJARCCE%20227.pdf0.
9. https://www.britannica.com/topic/AES#ref1095337.
10. http://www.di-mgt.com.au/rsa_alg.html.
11. https://www.irjet.net/archives/V3/i10/IRJET-V3I10126.pdf.
12. ahttps://en + b =.wikipedia c. .org/wiki/Advanced(1) (1) _Encryption_Standard.
13. https://www.tutorialspoint.com/cryptography/advanced_encryption_stan dard.html.
14. A Novel Approach to Enhance the Security Dimension of RSA Algorithm Using Bijective Function.
15. http://paper.ijcsns.org/07_book/201608/20160809.pdf.
16. Research and Implementation of RSA Algorithm for Encryption and Decryption.
17. https://globaljournals.org/GJCST_Volume13/4-A-Study-of-Encryption-Algorithms.pdf

Evolution of new version of internet protocol (IPv6) : Replacement of IPv4

Nargish Gupta*
Sumit Gupta**
Munna Pandey***

Abstract

Taking into consideration today's scenario internet is becoming a vital part of modern life. The basic functioning of Internet is based on Internet Protocol (IP). As we were using IPv4 but it has resulted in an unwanted growth issue. The reason behind its detonation is the brisk use of network addresses which leads to the decrement in the performance for routing. So in the coming years the unease of the internet will not decrease and the increase cannot be imagined with so much advancement in the technology. So to achieve this evolution in Internet there is a need for transition from IPv4 to IPv6. IPv4 address spaces has finally drained and IANA (Internet Assigned Numbers Authority) is left with no choice as to move towards the transition from IPv4 to IPv6. This paper reevaluates the main issue and the complications in IPv4- IPv6 transition and proposes the principles of tunneling and translation techniques. In this we surveys the mainstream tunneling and translation mechanisms, it new mechanism, techniques, pros and cons and appropriateness.

Keywords: Internet Protocol, IPv4, IPv6, Routing.

I. Introduction

Since the very early stage of the Internet IPv4 [1] has been used as the network layer protocol. No one has thought at the designing time of the protocol that the span of IPv4 Internet can be so bigger [2]. It was actually unexpected. The set of obstacles which are currently coming in IPv4 Internet is the exhaustion, routing scalability, and broken end-to-end property. IANA (Internet Assigned Numbers Authority) had been depleted with IPv4 address pool in Feb 2011, so as per the status we will soon be exhaust their IPv4 address space [3]. On the other hand, the technology is growing as fastest as possible especially the number of mobile users and it will continue. Thus resulting in the excessive demand for new IP address allocation which is difficult to gratify with IPv4. ChinaTelecom is among the biggest telecom ISPs (Internet Service Providers), as per them by the end of 2012, they will use up all the IPv4 addresses. Besides, the prefix de-aggregation caused by address block subdivision, multihoming and traffic engineering has caused a burst in Global IPv4 RIB (Routing Information Base) and FIB (Forwarding Information Base). Scalability problem is the biggest issue with which Internet is suffering. The basic end-to-end property all over the Internet has been broken down with the ample use of NAT.

Nargish Gupta*
IITM Janakpuri, New Delhi
Sumit Gupta**
LNCT Bhopal
Munna Pandey***
IITM Jankpuri, New Delhi

II. Challenges of IPv4

Since the advancement of technology our life style is become easier but there are various things under consideration. Now the new technology immersed which is internet of things means thing will communicate with each other. Due to this every device needs a unique address to identify uniquely this leads to various challenges on existing IP protocol i.e. IPv4 listed below:

- ***IP Address Depletion:***

In IPv4 limited number of unique public address are available (i.e. 4 billion) and IP enabled device are increases day by day. So every device needs a unique

IP address which immerses the some extra IP address especially for always on devices. IPv4 are not able to fulfill the IP demands.

- ***Internet Routing Table Expansion:***

Routing table is used by routers to make best path so network and entities connected to internet increases so does the number of network routes. These IPv4 routes consume a great deal of memory and processor resources on internet routers. Which will increases the complexity of the network as well as takes lots of space.

- ***Lack of end to end Connectivity:***

Due to better use of IP address IANA introduce public and private addressing. By using private address multiple devices are able to connect through the internet by single IP address. But it needs translation between public address to private ip address as well as private to public IP address. Network Address Translation (NAT) is an technology commonly implemented within IPv4 network NAT provide a way for multiple devices to share a single public IP address. This is an overhead which leads to increase complexity of the network and increases the possibility of error [4].

III. Improvement that IPv6 Provides

In early 1990's the internet engineering task force(IETF) grew concerned about the issues with IPv4 and began to look for replacement this activity leads to development of IP version 6. IPv6 overcome the limitation of IPv4 some are listed below:

- ***Internet address space:***

It increases address space 128 bit long instead of 32bit which is in IPv4. Due to increases the size it has more number of addresses which is sufficient to present as well as future scenario. IPv6 can allot 340 undecillion addresses to unique devices which is sufficient to handle present traffic.

- ***Improved Packet Handling:***

Ipv6 packet has eliminated the un required field which is not required from IPv4 and include required fields which is not present in the IPv4 header. IPv6 simplified with fewer fields this improve packet handling by intermediate routers and also provides support for extensions and options for increased scalability.

- ***Eliminates need of NAT:***

As mention earlier IP version4 does not have sufficient Ip addresses. So this problem is solved by Public and Private addresses. But use of private addresses required NATing which is an overhead. In IPv6 NATing concept is eliminated because of large number of IPv6 addresses.

- ***Integrated Security:***

IPv4 is the first IP version which is mostly focuses on the how we can transfer data from two or more devices. This requirement was successfully accomplished by IPv4. But as an technology increases chance to theft also increases. Ipv4 does not provide any security fields. By keeping in a mind IPv6 has integrated security. It provides authentication and privacy capabilities.

IV. Internet Protocol Version 6 (IPv6)

On Monday Jan 31 2011 IANA allocated the last two /8 IPv4 address block to Regional internet registries (RIR) so IANA implement IPv6. The packet format of IPv6 kept simple by adding fewer fields. All Fields of IPv6 are described in the packet format in figure 1.

<table>
<tr><td>Version
(4bit)</td><td>Traffic Class
(8 bit)</td><td colspan="2">Flow Control
(20 bit)</td></tr>
<tr><td colspan="2">Payload Length
(16 bit)</td><td>Next Header
(8 bit)</td><td>Hop Limit
(8 bit)</td></tr>
<tr><td colspan="4">Source IP Address
(128 bit)</td></tr>
<tr><td colspan="4">Destination IP Address
(128 bit)</td></tr>
</table>

Figure 1: Packet format of IPv6

Table 1: Comparison of Internet Protocol version 4 and Internet Protocol version 6

Characteristic Factor	IPv4	IPv6
Header Length	It is of 32 bit long	It is of 128 bit long
IP Security	It does not have any security	It provides integrated authentication and privacy capabilities
Address Resolution and Address Auto Configuration	It has ICMPv4 which does not includes address resolution and address auto configuration	ICMPv6 which includes address resolution and address auto configuration
NATing	Here we need Network Address Translator(NAT)	Due to large number of address space no need of NATing
Header	!2 basic header field not including option and padding field	Simplified with 8 fields this improve packet handling
Octets	20 (Up to 60 bytes if option field used)	40 (Large because of the length of source and destination)

Version: Version is same as IPv4 which is used to identify the version of the packet. It is of 4 bit long field. For IPv6 always set version field to 0110 and 0100 for IPv4.

Traffic Class: This field is same as type of service field in IPv4. It is of 8 bit long field used for real time application. It can be used to inform router and switches to maintain same path for the packet flow so that packet are not reordered.

Payload Length: Payload length field is 16 bit long field. It is equivalent to total length field in IPv4. Define entire packet size including header and optional extensions [5].

Next header: Next Header field is 8 bit long field which is similar to time to live field of IPv4. These values are decremented by one by each router that forwards the packets when value reaches zero packet is discarded and ICMPv6 message is forwarded to sending host indicate that packet did not reach to destination.

Source Address: It is of 128 bit long. This is used to specify the address of the sender who tries to send the message.

Destination Address: It is of 128 bit long. This address is used to specify the destination address that to sender wants to sends the message.

IPv6 packet might also contain extension header (EH) which provides optional network layer information. EH are optional and are placed between IPv6 header and payload. EH are used for fragmentation, for security, to support mobility and more [6].

V. IPv4 and IPv6 Coexistence

There is not a single date to move IPv6. Both Ipv4 and Ipv6 will coexist. The transition is expected to take years. IETF (Internet engineering task force) has created various protocols and tools to help network administrator migrate their network to IPv6. These migration techniques are divided into three categories:

Dual Stack: It allows Ipv4 and IPv6 to coexist on the same network. Dual stack devices run both IPv4 and IPv6 protocol stack simultaneously.

Tunneling: It is method to transporting IPv6 packet over an IPv4 network. IPv6 packet is encapsulated inside an IPv4 packet similar to other type of data.

Translation: NAT64 allows IPv6 enabled device to communicate with IPv4 enabled device using a translation technique similar to NAT for IPv4

VI. Comparision and Analysis

IPv6 provides 340 undecillion addresses roughly equal to every grain of sand on earth. Some field are renamed same. Some field from IPv4 is not used. Some field changed name and position. In addition new field has been added to IPv6 which is not IPv4 [7]. The detailed comparison between Internets Protocol version 4 and

Version 6 are shown in Table 1. In Table 1 first column shows the various characteristic factor bases on these two are differ. While second column is for IPv4 and third column is for IPv6 [8].

VII. Conclusion

IPv6 and IPv4 both are the Internet Protocols which we are currently used. Definitely IPv6 is the best among two because it comes after the IPv4 so it eliminate the drawbacks of IPv6. IPv4 is the popular protocol which we use since long time due to this both protocol keeps their equal importance. In this paper we can clearly see that the IPv6 is better replacement of IPv4 which will take time to overcome the IPv4.

References

1. W. Stalling, Data and Computer Communication, 5th Edition,upper saddle river, NJ: Prentice Hall, 2012.
2. M. Mackay and C. Edwards, "A Managed IPv6 Transitioning Architecture for Large Network Deployments," IEEE Internet Computing, vol. 13, no. 4, pp. 42 –51, july-aug. 2009.
3. S. Bradner and A. Mankin, IPng: internet protocol next generation reading, MA: Addision-Wesley, 2011.
4. R. Gillign and R. allon ,"IPv6 Transition mechanism overview" Connexions, oct 2002.
5. E. Britton, J. Tavs and R. Bournas, "TCP/IP: The next generation", IBM sys, J.No. 3, 1995.
6. C. Huitema, IPv6 the new internet protocol, Upper saddle river, NJ. Prentice Hall, 1996
7. R. Hinden,"IP next generation overview" connexions, Mar 1995.
8. Fernandez, P. Lopez, M. A. Zamora, and A. F. Skarmeta, "Lightweight MIPv6 with IPSec support (Online First, DOI: 10.3233/MIS-130171)," Mobile Information Systems, http://iospress.metapress.
9. G. Huston, "IPv4 Address Report," Tech. Rep., Sep. 2010. [Online]. Available: http://www.potaroo.net/tools/ipv4
10. S. Deering and R. Hinden, "Internet Protocol, Version 6 (IPv6) Speci- fication," 1998, IETF RFC 2460.
11. S. Thomson, T. Narten, and T. Jinmei, "IPv6 Stateless Address Autoconfiguration," 2007, IETF RFC 4862
12. R. Hinden and S. Deering, "IP Version 6 Addressing Architecture," 2006, IETF RFC 4291

Social Engineering – Threats & Prevention

Amanpreet Kaur Sara*
Nidhi Srivastava**

Abstract

The term "social engineering" (SE) has gained wide acceptance in the Information Technology (IT) and Information Systems (IS) communities as a social/psychological process by which an individual (called attacker) can gain information from an individual (called victim) about a sensitive subject. This information can be used immediately to by-pass the existing Identification-Authentication-Authorization (IAA) process or as part of a further SE event. Social engineering methods are numerous and people using it are extremely ingenious and adaptable. Nonetheless, the field is new but the tactics of the attackers remain same. Therefore, this paper provides an overview of the current scenario in social engineering and the security issues associated with it.

Keywords: Cyber security; risks; hacking; social engineering

I. Introduction

A typical misunderstanding regarding cyber-attacks/ hacks is that a very high end tools and technologies are used to retrieve sensitive information from someone's account, machines or mobile phones. This is essentially false. Hackers have discovered very old and simple method to steal your data by just conversing with you and misguiding you.[1] In this paper we will figure out how these sorts of human assaults (called social engineering assaults) work and what you can do to ensure yourself.

II. Types of Social Engineering Attacks

Here are some of the techniques that are commonly used to retrieve sensitive information.

A. Phishing

Phishing is the main type of social engg assaults that are commonly conveyed as an chat, email, web promotion or site that has been intended to imitate a real system and organisation. Phishing messages are created to convey a feeling of earnestness or dread with the objective of catching an end client's sensitive information. A phishing message may originate from a bank, the govt or a noteworthy organizations. The conversation or content of the call may vary. Some request that the customer to verify their login details, and incorporate a taunted up login page finish with logos and marking to look honest to goodness. Some claim the customer is the winner of a great prize or draw and demand access to a bank account in which to send the rewards. Some request altruistic gifts after a natural calamity or disaster.[2]

Amanpreet Kaur Sara*
IT Department,
Institute of Information Technology and Management

Nidhi Srivastava**
IT Department,
Institute of Information Technology and Management

B. Baiting

Baiting, like phishing, includes offering something very attractive to a customer at the cost of their login details or private information. The "Bait" is available in both forms digital and physical. Digital say for example some music or movie file download. While downloading you get the infected files and caught into trap. Physical say for example some flash drive with a name "Annual Appraisal Report" is intentionally left on someone's desk. As its name is so attractive anybody who will come and see it will definitely insert this drive to the system and he/she will be trapped. [2, 3]

C. Quid Pro Quo

This type of Assault happens when assailants ask for private or sensitive data from somebody in return for something attractive or some kind of pay. Say for eg a customer may get a telephone call from the assailants

who, acted like a technology expert, offers free IT help or innovation enhancements in return for login accreditations. [1,4] Another regular case is a assailants, acted like a specialist, requests access to the organization's system as a major aspect of an analysis or experiment in return for Rs.1000/- . On the off chance that an offer seems to be very genuine. Then is defiantly it is a quid pro quo.

D. Pretexting

In pretexting preplanned situation is created (pretext) to trap a targeted customer in order to reveal some sensitive information. In these type of situations customer perform actions that are expected by a hacker and he caught into the trap and reveal his/her sensitive information. [4] An elaborate lie, it most often involves some prior research or setup and the use of this information for impersonation (e.g., date of birth, Social Security number, last bill amount) to establish legitimacy in the mind of the target. [5]

E. Piggybacking

Other name for piggybacking is tailing. When a unauthorized person physically follows an authorized person into an organization's private area or system. Say for example sometimes a person request another person to hold the gate as he has forgotten his access card. Another example is to borrow someone's laptop or system for some times and installing malicious software by entering into his restricted information zone.

F. Hoaxing

Hoaxing is an endeavor to trap the people into thinking something false is genuine. It likewise may prompt to sudden choices being taken because of fear of an unfortunate incident.

III. Preventions

By educating self, user can prevent itself from the problem of social engineering to large extent. Extremely common and easy way is not to give the password to anyone and by taking regular backup of the data. There has to be strict action. Application of authentication system like smart cards or biometrics is a key. By doing this, you can prevent a high percentage of social engineering attempts. There has to be good policies for successful defense against the social engineering and all personnel should ensure to follow them. It is not about typical software system for Social engineering attacks but the people which in themselves are quite fickle. There are certain counter measures which we can help in reduction of these attacks.[18]

Below mentioned are the prevention techniques for individual defense.

A. We should always be vigilant of any email which asks for personal financial information or warns of termination of online accounts instantly.

B. If an email is not digitally signed, you cannot ensure if the same isn't forged or spoofed. It is highly recommendable to check the full headers as anyone can mail by any mail.

C. Generally fraudulent person would ask for information such as usernames, passwords, credit card numbers, social security numbers, etc. This kind of information is not asked normally by even the authorized company representative. Hence one should be careful.

D. You may find Phisher emails are generally not personalized you may find something like this "Dear Customer". This is majorly because of the fact that these are intended to trap innocent people by sending mass mailers. Authorized mails will have personalized beginning. However one should be vigilant as phisher could send specific email intending to trap an individual. It could well then be like our case study.

E. One should very careful while contacting financial institutions. It has to be thoroughly checked while entering your critical information like bank card, hard-copy correspondence, or monthly account statement. Always keep in mind that the e-mails/ links could look very authentic however it could be spurious.

F. One should always ensure that one is using a secure website while submitting credit card or other sensitive information via your Web browser.

G. You should log on and change the password on regular basis.[15]

H. Every bank, credit and debit card statements should be properly checked and one should ensure that all transactions are legitimate

I. You should not assume that website is legitimate just by looking at the appearance of the same.

J. One should avoid filling forms in email messages or pop-up windows that ask for personal financial information. These are generally used by spammers as well as phisher for attack in future.[10]

IV. Conclusion

In today's world, perhaps we could have most secured and sophisticated network or clear policies however we humans are highly unpredictable due to sheer curiosity and never ending greed without concern for the consequences. We could very well face our own version of a Trojan tragedy [11]. Biggest irony of social engineering attacks is that humans are not only the biggest problem and security risk, but also the best tool to defend against these attacks. Organizations should definitely fight social engineering attacks by forming policies and framework that has clear sets of roles and responsibilities for all users and not just security personnel. Also organization should make sure that, these policies and procedures are executed by users properly and without doubt regular training needs to be imparted given such incidents' regular occurrence.

References

1. "Ouch" The monthly security newsletter for computer users issue(November 2014)
2. "Mosin Hasan, Nilesh Prajapati and Safvan Vohara" on "CASE STUDY ON SOCIAL ENGINEERING TECHNIQUES FOR PERSUASION" in International journal on applications of graph theory in wireless ad hoc networks and sensor networks (GRAPH-HOC) Vol.2, No.2, June 2010
3. "Christopher Hadnagy " -A book on "Social Engineering -The Art of Human Hacking "Published by Wiley Publishing, Inc. in 2011
4. The story of HP pretexting scandal with discussion is available at Davani, Faraz (14 August 2011). "HP Pretexting Scandal by Faraz Davani". Scribed. Retrieved 15 August 2011.
5. "Pretexting: Your Personal Information Revealed", Federal Trade Commission
6. "Tim Thornburgh" on "Social Engineering: The Dark Art" published in ACM digital library Proceeding New York in infoSecCD '04 Proceedings of the 1st annual conference on Information security curriculum development page 133-135.
7. "Valericã GREAVU-ªERBAN, Oana ªERBAN" on " Social Engineering a General Approach" in Informatica Economicã vol. 18, no. 2/2014
8. Malware : Threat to the Economy, Survey Study by Mosin Hasan, National Conference IT and Business Intelligence (ITBI - 08)
9. White paper: Avoiding Social Engineering and Phishing Attacks,Cyber Security Tip ST04-014, by Mindi McDowell,Carnegie Mellon University, June 2007.
10. Book of 'People Hacking' by Harl
11. FCAC Cautions Consumers About New "Vishing" Scam, Financial Consumer Agency of Canada, July 25, 2006.
12. Schulman, Jay. Voice-over-IP Scams Set to Grow, VoIP News, July 21, 2006.
13. Spying Linux: Consequences, Technique and Prevention by Mosin Hasan, IEEE International Advance Computing Conference (IACC'09)

14. Redmon,- audit and policy Social Engineering manipulating source , Author: Jared Kee,SANS institute.

15. White paper 'Management Update: How Businesses Can Defend against Social Engineering Attacks' published on March 16, 2005 by Gartner.

16. White paper, Social Engineering:An attack vector most intricate to tackle by Ashish Thapar.

17. The Origin of Social Engineering Bt Heip Dand MacAFEE Security Journal, Fall 2008.

18. Psychology: A Precious Security Tool by Yves Lafrance,SANS Institute,2004.

19. SOCIAL ENGINEERING: A MEANS TO VIOLATE A COMPUTER SYSTEM, By Malcolm Allen, SANS Institute, 2007

20. Inside Spyware – Techniques, Remedies and Cure by Mosin hasan Emerging Trends in Computer Technology National Conference