ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Applying Data Mining Techniques in Cyber Crimes

[1]Mr. G .Sivaselvan, [2]Dr.V.Vennila, [3]R.Senbagavalli, [4]E.Shanmugapriya, [5]K.Umadevi, [6]S.Suganthi
[1,2]Assistant Professor, Department of CSE, K.S.R. College of Engineering,
[3-6] UG Students, Department of Computer Science and Engineering, K.S.R. college of Engineering

Email id: [1]trnpsiva@gmail.com,vennilview@gmail.com,[3]senbagavalli.rcse@ksrce.ac.in, [4]shanmugapriya.ecse@ksrce.ac.in, [5]umadevi.kcse@ksrce.ac.in,[6]suganthi.scse@ksrce.ac.in

***Abstract*— Globally the internet is been accessed by enormous people within their restricted domains. When the client and server exchange messages among each other, there is an activity that can be observed in log files. Log files give a detailed description of the activities that occur in a network that shows the IP address, login and logout durations, the user's behaviour etc. There are several types of attacks occurring from the internet. Our focus of research in this paper is Denial of Service (DoS) attacks with the help of pattern recognition techniques in data mining. Through which the Denial of Service attack is identified. Denial of service is a very dangerous attack that jeopardizes the IT resources of an organization by overloading with imitation messages or multiple requests from unauthorized users.**

Index Terms— ***Denial of Service, Log File, Cyber Crimes, Data mining, outliers, Association rules.***

I. INTRODUCTION

In our research paper we are focusing the detection of the Denial of Service (DoS) attacks using Data Mining techniques. This will jeopardize the network and IT resources by artificially increasing the network traffic and load on the server by sending imitation requests.

In network architecture, the networking strategies should be made prone to identify intruders that attack the system (or cause a denial of service attack). DoS attacks are some of the oldest Internet threats and continue to be the top risk to networks around the world. DoS events making it difficult to detect them. DoS attack remains a serious problem that increasingly affects company resources, in most cases a DoS attack caused the services to be completely unavailable impacting their business directly which is a potential financial loss to businesses.

According to, the number of DoS attacks durations by the end of the year 2015 became shorter and more discreet. The figure 1 below shows the ups and downs of DoS attacks.

A. Cyber Crimes

Cyber refers to something that can be done on internet. Crime refers to something that is done illegally or without authorization. All those crimes that are done on the internet in order to gain access to secured information or authorization rights is termed as ''Cyber Crime''. Globally the cyber-crime hindrance is spread across abundantly.

B. Cyber Security

Cyber Security is that branch of Computer Technology that deals with security in cyberspace. Cyberspace refers to the description of policies regarding the networks and computer systems. The policies laid out in the Cyber security are for the reason of avoiding the malicious activity or unauthorized access to secured information. Since the emergence of high structured networks [1], there arises a concern about how intelligently these networks are secured. These issues are major concerns in the internet era. Cyber security is concerned with protecting IT resources like server; network etc. from performing illegal activities or fraudulent acts. Data mining is also applicable to problem solving or network intrusions. Therefore in this paper we focus the applications of data mining for cyber security applications.

II. RELATED WORKS

The previous work that the performance attained is highly correlated to the distance-preserving properties of the anonymization format used. The parameters, of the blocking scheme, are optimally selected so that we achieve the highest possible accuracy in the least possible running time. We also introduce an SMC-based protocol in order to compare the formulated record pair's homomorphically , without running the risk of breaching the privacy of the underlying records.

The above match criteria may be applied to compare the attributes of two or more tuples to determine whether they are co-referent or otherwise related. The manner in which the techniques are applied to the target relations may vary producing different running times and coverage of the constituent tuples.

Once similarity techniques above have been established, the manner in which those techniques are applied must be chosen. A number of applications are SMC Application: Using this application, each tuple in a relation is compared to every other tuple. When a tuple matches, it is changed (for example, if it is determined to form an entity and is merged), this changed tuple must again be compared to every other tuples.

Data Mining emphasizes the extraction of data from databases and various patterns can be concluded for deriving association rules. Although Data Mining is eventually gaining a wider scope in different areas, its research has made remarkable significance in Cyber Crimes.

Data mining, which is defined as the process mining or extracting data into productive information. Based on this data, significant patterns are formed.

A. Association Rules

The term association means, "connectivity" or "together" or frequently that appears. Therefore association rules are related to those conditions where the values in a data set are frequently appearing and this appearance will show relationship or "connectivity" among the values. For this reason, in order to show the relationship, we assign a support or threshold value and henceforth the association rules are generated based on the algorithm adopted, like "Apriori" or "K-means" etc.

B. Cluster Analysis

The term "cluster" refers to "groups" or categorizing the data into separate labels. These labelled classes hold similar type of data. Henceforth it implies that data in different class labels differ from each other in terms of their features. When this data is analysed in different classes or labels then using different analysis techniques, the data is extracted. This process is called as "Cluster Analysis". In fig. 3, the cluster analysis is shown along with outliers.

III. PROPOSED SYSTEM

Our proposed work could save many computation cycles and thus allow accurate information provided to the right people at the right time. Two considerations when forming a data warehouse are data cleansing (including entity resolution) and with schema integration (including record linkage).

Uncleansed and fragmented data requires time to decipher and may lead to increased costs for an organization, so data cleansing and schema integration can save a great many (human) computation cycles and can lead to higher organizational efficiency.

In this work based on our previous methodologies proposed or developed for entity resolution and record linkage. This survey provides a foundation for solving many problems in data record linkage analysis. For instance, little or no research has been directed at the problem of maintenance of cleansed and linked relations.

Our proposed work used an BFS (Breadth First Search) algorithm is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent entities. Record linkage identifies matching record pairs in two separate data files.

The record linkage results in a classification of pairs of records as links and non-links. Pairs of records which represent identical observational units are called match. Our fuzzy BFS which work based on two modules, exact match, and distance match.

Cyber-crimes their types clustering, outliers and pattern recognition. We have applied the famous data mining techniques called as pattern recognition on the log file. We set a threshold value. If the number of similar request are received at the server, which is greater than the threshold value assume this as an attack and the administrator is been informed. By this approach we can identify the denial of service attack easily as in DOS attack or the hacker sends same multiple request in order to mitigate the server performance.

IV. PROPOSED DESIGN

A. Input Design

The system design is an interactive process through which requirements are translated into a blue print or a representation of software that can be accessed for quality before code generation begins.

Input design is a part of overall system design, which requires careful attention. Input of data as designed as user-friendly and easier. Input design is a process of converting the user- oriented description of the input to the computer based information system into programmer- oriented specification. The objective of the input design is to create an input layout that is easy to follow and prevent operator errors algorithm is used to find all pairs of shortest record linkage Profiles, i.e. P. Each record linkage Profile pi consists of sequence of vertices from source to destination. The graph traversal module produces set of shortest record linkage Profiles between all pair of source and destination as intermediate results. All the shortest record linkage Profiles are computed using well-known BFS algorithm. Secondly, the overlapped regions of shortest record linkage Profiles are identified through pattern mining approach. We limit the number of BFS execution up to K. Therefore, instead all pair shortest record linkage Profiles, $K * N$ number of shortest record linkage Profiles are computed where $K << N$. These sample shortest record linkage Profiles are further utilized for identifying the overlapped regions. The social networks are usually dense,

follows the power law distribution, so even small number of shortest record linkage Profiles can lead us to better or acceptable analysis.

B. Output Design

The output design refers to the results and information that are generated by the system for many end users. Efficient and intelligent output design improves the system relationships with the user and help in decision making. We randomly generate 100 shortest record linkage Profile queries with constraint record linkage Profiles for comparing the average time cost. We use the single directional BFS approach using the index table for searching shortest record linkage Profile. The proposed method (BFS) requires about 40K rows and BFS requires about 1,300K rows. We also randomly generate 100 shortest record linkage Profile queries with constraint record linkage Profile for comparing the average time cost. In Figure 1(b), LSH consumes 0.75 seconds and BFS consumes 0.72 seconds. These methods show similar time cost. From the experimental results, LSH shows higher space efficiency than BFS with similar execution time.

Our proposed work could save many computation cycles
And thus allow accurate information provided to the right people at the right time. Two considerations when forming a data warehouse are data cleansing (including entity resolution) and with schema integration (including record linkage).

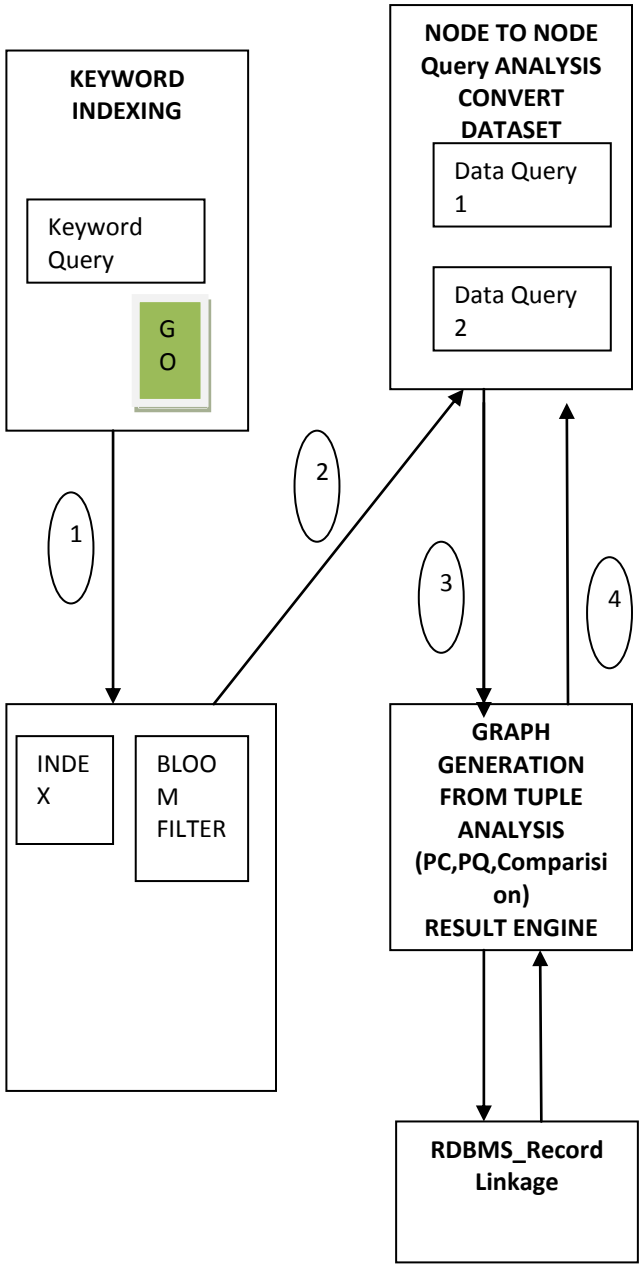
Uncleansed and fragmented data requires time to decipher and may lead to increased costs for an organization, so data cleansing and schema integration can save a great many (human) computation cycles and can lead to higher organizational efficiency.

In this work based on our previous methodologies proposed or developed for entity resolution and record linkage. This survey provides a foundation for solving many problems in data record linkage analysis. For instance, little or no research has been directed at the problem of maintenance of cleansed and linked relations.

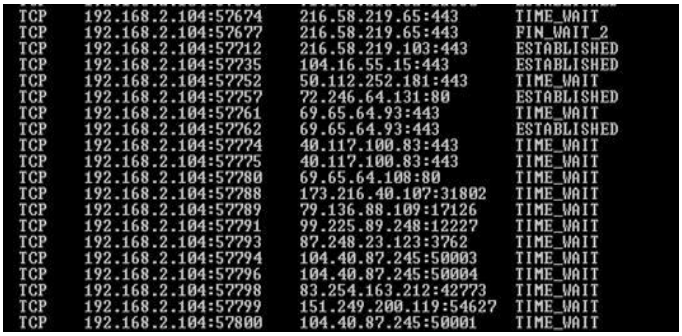
Our proposed work used an BFS (Breadth First Search) algorithm is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent entities. Record linkage identifies matching record pairs in two separate data files.

The record linkage results in a classification of pairs of records as links and non-links. Pairs of records which represent identical observational units are called match.

Our fuzzy BFS which work based on two modules, exact match, and distance match.



C.Result Analysis



TCP	192.168.2.104:57674	216.58.219.65:443	TIME_WAIT
TCP	192.168.2.104:57677	216.58.219.65:443	FIN_WAIT_2
TCP	192.168.2.104:57712	216.58.219.103:443	ESTABLISHED
TCP	192.168.2.104:57735	104.16.55.15:443	ESTABLISHED
TCP	192.168.2.104:57752	50.112.252.181:443	TIME_WAIT
TCP	192.168.2.104:57757	72.246.64.131:80	ESTABLISHED
TCP	192.168.2.104:57761	69.65.64.93:443	TIME_WAIT
TCP	192.168.2.104:57762	69.65.64.93:443	ESTABLISHED
TCP	192.168.2.104:57774	40.117.100.83:443	TIME_WAIT
TCP	192.168.2.104:57775	40.117.100.83:443	TIME_WAIT
TCP	192.168.2.104:57780	69.65.64.108:80	TIME_WAIT
TCP	192.168.2.104:57788	173.216.40.107:31802	TIME_WAIT
TCP	192.168.2.104:57789	79.136.88.109:17126	TIME_WAIT
TCP	192.168.2.104:57791	99.225.89.248:12227	TIME_WAIT
TCP	192.168.2.104:57793	87.248.23.123:3762	TIME_WAIT
TCP	192.168.2.104:57794	104.40.87.245:50003	TIME_WAIT
TCP	192.168.2.104:57796	104.40.87.245:50004	TIME_WAIT
TCP	192.168.2.104:57798	83.254.163.212:42773	TIME_WAIT
TCP	192.168.2.104:57799	151.249.200.119:54627	TIME_WAIT
TCP	192.168.2.104:57800	104.40.87.245:50001	TIME_WAIT

V. CONCLUSION

In this work, we empirically analyze the shortest record linkage Profiles to anticipate the behaviour of BFS algorithm on real life networks. A set of shortest record linkage Profiles are evaluated using pattern mining approach.

We have found that the nodes with very high degree are retained in majority shortest record linkage Profiles. However, nodes with average degree are not considered by the traversal algorithm.

The statistical analysis also shows the similar behaviour in terms of network properties, including clustering coefficient, average shortest record linkage Profile, and betweens centrality, on various types of networks.

The influence of edge weights and directional information on shortest record linkage Profile traversal is still an interesting area of research. It achieves over 30% mean squared error reduction over BFS-LSH in estimating angular similarity, when the Super-Bit depth N is close to the data dimension d.

Moreover, BFS-LSH performs best among several widely used data-independent LSH methods in approximate nearest neighbor retrieval experiments.

We have applied the data mining techniques for identifying the Denial of Service attack. This type of attack is very dangerous as it jeopardizes the IT resources. It makes the server busy by imitation messages and repeated queries. The server is congested by traffic packets, in order to mitigate the server performance. In this research paper, we have discussed about Cyber security, cyber-crimes their types, clustering, outliers and pattern recognition. We have applied the famous data mining technique called as pattern recognition on the log file. We set a threshold value. If the number of similar requests are received at the server, which is greater than the threshold value, we assume this as an attack and the administrator is been informed. By this approach we can identify the denial of service attack easily as in DoS attack, the attacker or the hacker sends same multiple requests in order to mitigate the server performance.

VI. REFERENCES

[1] Know Your Enemy: Learning about Security Threats, 2nd Edition.ISBN: 0321166469. The Honeypot Project 2004.

[2] M.Khan, S.K.Pradhan, M.A.Khaleel, "Outlier Detection for Business Intelligence using data mining techniques", International journal of Computer Applications (0975 -8887), Volume 106- No. 2, November 2014.

[3] Masud, M.M, Gao,J.Khan,"Peer to Peer Botnet Detection for Cyber Security: A Data Mining Approach". In proceedings: Cyber-security and information Intelligence research workshop. Oakridge national Laboratory, Oakridge May 2008.

[4] Internet Security Threat Report, Volume 21, April 2016, Symantec Crime Report.

[5] Ibrahim Salim, T.A.Razzack,"A study on IDS for Preventing denial of service attack using outlier's techniques", 2nd IEEE international conference on Engineering and technology, March 2016.

[6] S.S Rao, SANS Institute InfoSec Reading Room.,"Denial of service Attack and mitigation techniques: Real time implementation with detailed analysis", 2011.

[7] Data Mining: Concepts and Techniques, Third Edition, Jiawei Han and Micheline Kamber, ISBN-13, 9780123814791.

[8] Mining of Massive Data Sets, AnandRajaraman, Jure Leskovec, Jeffrey D. Ullman,2014

[9] A. Klein, F. Ishikawa, and S. Honiden. Efficient heuristic approach with improved time complexity for QoS-aware service composition. In ICWS, pages 436–443. IEEE, 2011.

[10] Tripathy, M.Khan, M.R.Patra, H.Fatima, P.Swain, "Dynamic web service composition with QoS clustering" IEEE, International Conference on Web services, 2014.

[1]G.Sivaselvan is an Assistant Professor in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. He received his Master of Engineering degree in Computer Science and Engineering in 2011 from Anna University, Chennai, India. He is a Research scholar in Anna University, Chennai. He has published more than 25 papers in referred journals and conference proceedings. His research interest includes Wireless Sensor Networks, Cloud computing and Computer networks. He is a professional member of ISTE.

[2]V.Vennila is an Assistant Professor in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. She received her Ph.D. Degree in Data Mining in 2017 from Anna University, Chennai, India. She has published more than 26 papers in referred journals and conference proceedings. Her research interest includes Data Mining, Big Data, Cloud Computing, Databases and Artificial Intelligence. She is a professional member of ISTE.

[3]R.Senbagavalli is a final year student in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. Currently she is doing her final year project regarding cyber crimes in data mining.

[4]E.Shanmugapriya is a final year student in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. Currently she is doing her final year project regarding cyber crimes in data mining.

[5]K.Umadevi is a final year student in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. Currently she is doing her final year project regarding cyber crimes in data mining.

[6]S.Suganthi is a final year student in the Department of Computer Science and Engineering, K.S.R. College of Engineering (Autonomous), India. Currently she is doing her final year project regarding cyber crimes in data mining.