Survey on Prediction of Weblogs for improving the performance in Web Usage Mining

Pradip Suresh Mane [1], Dr. Ashok Kumar Jetawat[2] , Dr. Pravin Nikumbh[3]
[1] Pacific Academy Of Higher Education And Research University, Udaipur, India
[2] Pacific Academy Of Higher Education And Research University, Udaipur, India
[3] Lokmanya Tilak college of Engineering , Kopar Khairane, Maharashtra, India
pradipmane510@gmail.com

Abstract: Today, the massive data that is topographically appropriated has inferred a requirement for circulated registering to diminish the time devoured to process the information. In such a focused domain, specialist are restless to consider are they giving best organization in the market, paying little respect to whether people are acquiring their thing, are they finding application fascinating and neighborly to use, or in the field of setting aside extra cash they need to consider what number of customers are foreseeing our bank scheme. In comparable way, they additionally need to think about issues that have been happened, how to determine them, how to make sites or web application fascinating, which items individuals are not acquiring and all things considered how to enhance publicizing techniques to draw in client, what will be the future marketing plans.

Aim is that Web applications will constrain the sorts of questions that can be produced to a protected subset of every single conceivable query, paying little heed to what input users give. Be that as it may, deficient information approval can empower aggressors to increase full access to such databases.

Keywords: **MapReduce, Log Files, Web security,**

1. Introduction:

Log records contain rundown of activities that have been happened at whatever point somebody gets to your web based application or website. The log files are present in servers. Digital data like Access log files are generated by web servers which contain the server requests. This data is semi structure format and Querying using relational database is not possible. Using Web usage mining data mining will process Weblog for extract sequential patterns, association patterns, and current trends in web access. in marketing campaign analysis such type of Analyzed and explored records is used Analyzing and exploring of such records is used.

Advanced information created by the web servers are access log records which contain the majority of the solicitations made to the server. The majority of the information are semi organized organization by which questioning through social database is beyond the realm of imagination. Information Mining can process Weblog records to discover consecutive examples, affiliation examples, and current patterns in web access utilizing Web use mining. Breaking down and investigating of such records is utilized in promoting effort examination

Web application security is an important problem in today's internet. A major reason of this status is that most of programmers don't have substantial knowledge about secure coding, so they leave applications with vulnerabilities. The method to solve this issue is to use source code static analysis to find these errors, but report known these tools many false positives to correcting the application that make hard the task. Normal users only visit the pages and never wants check that the connected system is a server or not. The behavior of machine is mostly captured through Web server logs not the behavior of end user. Caching suspicious end user significant help is providing by Log files that provide security, troubleshooting and pro-active system administration. Source for IP7 addresses will be intersection of log files coming out from different machines can be intruders. Create a system which will identifying these IP7 addresses by using such information in real-time, and then over the LAN8 these list will be distributed to other machines ,but the firewall access lists is exempted , or the list will be send to network firewall.

2. Related Work

Ibéria (2016) explore an approach while the programmer will keeping in the loop, it automatically protecting web applications. The web application source code consists in analyzing by this approach and input validation vulnerabilities to be searched and in the same code insert fixes, will correct these flaws. To understand where the vulnerabilities were found by keeping programmer will be in the loop, and how these vulnerabilities were corrected. to the by eliminating vulnerabilities form the web applications, this process provides security to the application and also by the mistakes programmers will learn indirectly[2].

Meena (2016) described an approach for web applications to finding and correcting vulnerabilities in it. is also used To identify false positives used Data mining with the top 3 machine learning classifiers, and using an induction rule classifier to justify their presence. After a thorough comparison of several alternates all classifiers were selected. It is important to note that it cannot provide entirely correct results by this combination of detection techniques. This problem is unwanted, and this undesirability resorting to data mining only generate probabilistic results ,but cannot circumvent. The tool by using a framework of collaborative testing corrects the code in which completed the test tasks through the using the ontology of software testing collaboration of various test services that are discovered, registered and invoked at runtime[4].

P. O. Prakash (2016) The aim of the research was to classify the data of success response and analyze the user navigation by using IncSpan and Partition Algorithm Frequent itemset (PAFI) algorithm. This research work explored to analyze the user prediction, based on the user preference present in various levels that was captured from weblogs. The system showed the interest on vehicle but failed to identify the buying prediction of user by using PAFI algorithm. [6].

Priyanka (2016) proposed a framework that can predict the user behaviour accurately. Markov model can to use to determine next state based on previous state but Low order Markov models don’t consider past in detail and therefore accuracy is very low, whereas high order Markov models has higher complexity. Association rules can also be used for prediction but it generates many rules, which gives result in to predictions for a user session inconsistently. Approach based on mixture of Markov models and association rules that result in good prediction accuracy. This paper used Markov model with association rule to predict the future page. Predicted page can be pre-fetched to improve performance [7].

Savitha (2014), elucidates the analysis of log files using Hadoop MapReduce framework which incorporates the major pre-processing task and session identification algorithm to handle vast amount of log data. From results it is concluded that processing a huge file in distributed fashion reduces the time and data transfer cost, without moving the data. The text files can be screwed in order to produce a statistical report for better understanding of users view. Also performing the same task for multiple files of large volume of data reduces the memory utilization, CPU load and other factors that results the process in easy head. The proposed work aims on processing the session accessed by the user in log files which is the main part of analysis. It focuses on the total time span exhausted by the user for each requested page. Based on the results of the time spent in each page of a website, path tracking modifications can be done on the structure of the site. This framework implements MapReduce model that incorporates batch processing on commodity hardware. The main principle of this work is to move computation on data rather than moving data over the computation[9].

Srinivasa (2013) shown the architecture for Map Reduce of Single cluster for situations which is suitable when data and computer resources are widely distributed. They did experiment by using the 4GB size of data and single node, 2 and 4 nodes results were compared. It is examined that as the number of nodes increment the performance also increases. Therefore the time required to process a large file will reduce by the Map Reduce, for reduced memory utilization, analysis, eliminate redundancies in the log files, and also distributed the CPU load, because of the model for parallel programming. From their experimental evaluations, as a observation from the lessons learned, when to apply the different MapReduce architectures provide the recommendations, as a function of various key parameters: network topology, data transfer costs, workloads and data partitioning[8].

E. A. Neeba (2017) a novel algorithm named ‘swarm-based cluster algorithm’ (SBCA) was proposed to complete the feature selection task in order to produce optimized feature based clusters for effective data and weblogs classification. The classifier performance of the proposed algorithm results in high accuracy, minimized the output errors and consumed less time. Hence, SBCA has the enhanced optimization ability in comparison with other attribute selection algorithms. The method was unable to handle high-dimensional data related to real-time applications[11].

S. Zhang (2018) presents a sentiment analysis method for Chinese micro-blog text based on the sentiment dictionary to support network regulators' work better. First, the sentiment dictionary was extended by extraction and construction of degree adverb dictionary, network word dictionary, negative word dictionary and other related dictionaries. Second, the sentiment value of a micro-blog text was obtained through the calculation of the weight. Finally, micro-blog texts on a topic was classified as positive, negative and neutral. The experimental results showed that sentiment analysis of Chinese micro-blog topic based on sentiment dictionary was effectively and accurately analyze micro-blog's sentiment, which was helpful for the public opinion supervisors to make appropriate decisions. The method analyzed the sentiment dictionary but failed to analyze the topic tracking and topic prediction[12].

V. Pushpa(2016) concentrated on data pre-processing of web usage mining, which was performed to eliminate the inconsistence, irrelevant data and extracted the interesting knowledge patterns from the weblogs, which assists to know about the user behaviour. This paper described the data preprocessing of web usage mining from which the weblogs were cleaned effectively using the web log explorer tool. This tool had the flexible system of filters, which gave information about visitors who accessed the specific web page. It was used to remove the irrelative entries and generate useful log report. The method focused only on pre-processing method but failed to concentrate on pattern discovery and pattern analysis of web usage mining[13].

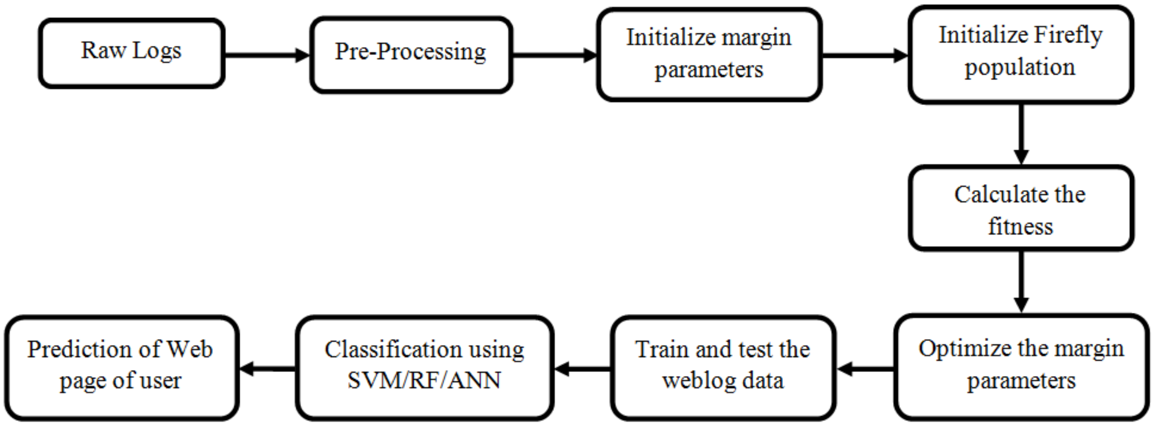
3. Research Gap

The exponential growth of data gave us an opportunity to know more about the hidden secrets by which one can make golden nuggets. Existing techniques were becoming inadequate to process such large datasets. As the growth of data increases over years, analysis and storage becomes difficult, this will increases the cost efficiency and processing time. Though various techniques and algorithms are used in distributed computing the problem remains still idle. Although for more than a decade a large research effort has been going on web application security, the more continues challenging problem is the security of web applications. An important part of that vulnerable source code written in unsafe languages like PHP derives problem. To find vulnerabilities from Source code, static analysis tools are used, but they lead to generate false positives, and require great effort for programmers to manually make the changes to fix the code.

4. Problem Definition

In the current situation more number of users attracted towards the internet, so there is growth of users accessing the internet are increases day by day and reduces the shopping time, so data size will increase. Identifying the interested user and not interested user is difficult, based on weblog user interest can be classified. The weblog consists of history of information while the user accessing the websites. The prediction of user behaviour is a difficult task.

Block Diagram


Raw Logs
Pre-Processing
Initialize margin parameters
Initialize Firefly population
Calculate the fitness
Optimize the margin parameters
Train and test the weblog data
Classification using SVM/RF/ANN
Prediction of Web page of user

4. Proposed System

This research mainly concentrates on processing the session accessed by the user in log files which is the main part of analysis. It focuses on the total time span exhausted by the user for each requested page. Based on the results of the time spent in each page of a website, path tracking modifications can be done on the structure of the site.The point of a log file is to keep track of what is happening with the web server.

Log files are also used to keep track of complex systems, so that when a problem does occur, it is easy to pinpoint and fix. But there are times when log files are too difficult to read or make sense of, and it is then that log file analysis is necessary. These log files have tons of useful information for service providers, analyzing these log files can give lots of insights that help understand website traffic patterns, user activity, there interest. Thus, through the log file analysis we can get the information about all the above questions as log is the record of people interaction with websites and applications. Log analysis can be done by various methods but what matters is response time.

The following steps describe the proposed method for predicting the web page prediction from weblogs.

- While user accessing the web server, the user transaction is generated in web server as log, it contains unstructured format.
- The weblog consists of various entries like IP address, date, time, request method, protocol, categories and number of bytes transmitted and status code.
- The above attributes helps to identify user navigation and the pattern can be classified after pre-processing.

- The prediction of user behaviour can be identified only through weblogs.
- The weblog contains unstructured format, so convert to raw weblog to processed weblog using data pre-processing, the data pre-processing includes data cleaning, user and session identification.
- Data cleaning: This task is usually site-specific and removing extraneous references to style files, graphics, or sound files that may not be important for the purpose of analysis.
- User identification: The remaining entries are grouped by individual users. Because no user authentication and cookie information is available in most server logs, we used the combination of IP, user agent, and referrer fields to identify unique users.
- User session identification: The activity record of each user is segmented into sessions, with each representing a single visit to a site. Without additional authentication information from users and without the mechanisms such as embedded session IDs, one must rely on heuristics for session identification.
- After the preprocessing of server log file, data mining techniques are applied for prediction.
- But, before applying to the data mining technique, margin parameters of Support Vector Machine (SVM)/Random forest (RF)/ Artificial Neural Network (ANN) should be initialized.
- Firefly algorithm is used to optimize these margin parameters by calculating their fitness function.
- After optimization, these parameters are trained then given as input to the data mining techniques such as SVM/RF/ANN to predict the web pages of users from weblogs.

Objective is to provide security; hence research focuses on detection and prevention of related vulnerabilities, identify the unauthorized access in weblogs, thus improve the performance of the server.

1. Analyze weblogs of all the browsers from the user end and combining all the user web logs into a single datasheets
2. To evaluate the value of weblogs on performance and compare their probabilities with a threshold.

3. Implement intelligent prediction of weblog access for improving performance.
4. Compare and study the usage of data that is gathered from datasets to produce the results for effective web logger behavior.

All the above objectives will provide the security to the system and enhanced the performance of the system. More important it helps to detect unauthorized access to the systems from log and also predict the vulnerability of the system.

6. Conclusion

This paper presents an approach for web application to find and correct vulnerabilities. By combination of static code analysis and data mining finding the approach to search vulnerability. Identify false positives identification is done using a machine learning classifier which will be selected after a comparison of many alternatives hy using Data Mining.

References:

1] Ashwani Garg,Shekhar Singh, January 2013,”A Review on Web Application Security Vulnerabilities”, IJARCSSE,Vol. 3, Issue 1, pp. 222-226.

2] Ibéria Medeiros, Nuno Neves and Miguel Correia, 2016, “Detecting and Removing Web Application Vulnerabilities with Static Analysis and Data Mining” IEEE Transactions, vol. 65,pp/ 54 – 69.

3] Ibéria Medeiros, Nuno Neves and Miguel Correia, 2014,“Automatic Detection and Correction of Web Application Vulnerabilities using Data Mining to Predict False Positives”, Proceedings of the 23rd international conference on WWW, pp. 63-74.

4] R. Meena, Dr. R. Kalpana, March- 2016,”Collaborative Framework for Testing Web Application Vulnerabilities Using STOWS”, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3, pp. 173-178

5] L. K. Shar, H. B. K. Tan, and L. C. Briand, 2013,“Mining SQL injection and cross site scripting vulnerabilities using hybrid program analysis,” in Proc. 35th Int. Conf. Software Engineering, pp. 642–651.

6] P. G. Om Prakash, Dr. A. Jaya, 2016,”Analyzing and Predicting User Behavior Pattern from Weblogs”, International Journal of Applied Engineering Research, Vol. 11, No. 9,pp 6278-6283.

7] Priyanka Makkar, Meenal Shingare, Dipali Kadam, March 2016,”Predicting User Access Pattern Using Markov Model and Association Rule”IJETAE,Vol 6, Issue 3, pp. 181-185.

8] P. Srinivasa Rao, K. Thammi Reddy and MHM. Krishna Prasad, 2013, “A Novel and Efficient Method for Protecting Internet Usage from Unauthorized Access Using Map Reduce”. I.J. Information Technology and Computer Science, 03, 49-55.

9] Savitha K, Vijaya MS,2014,”An Efficient Analysis of Web Server Log Files for Session Identification using Hadoop Mapreduce”,Proc. of Int. Conf. on Advances in Communication, Network, and Computing,pp.241-246.

10] Sayalee Narkhede and Tripti Baraskar, July 2013, “HMR Log Analyzer: Analyze Web Application Logs over Hadoop MapReduce”, International Journal of UbiComp (IJU) vol.4, No.3, pp.41-51.

11] E. A. Neeba, S. Koteeswaran, and N. Malarvizhi. "Swarm-based clustering algorithm for efficient web blog and data classification." *The Journal of Supercomputing* (2017): 1-14.

12] S. Zhang, Wei, Z., Wang, Y., & Liao, T. (2018). Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary. *Future Generation Computer Systems*, 81, 395-403.

13] V. Pushpa, and V. Vidyapriya. "An Efficient Preprocessing Method to Detect User Access Patterns from Weblogs." *International Journal of Computer Science and Mobile Computing* 5.9 (2016): 16-22.