A New Data Mining Technique for Web Log

Gajendra Singh
M.Tech CS Dept, SSSIST Sehore, RGPV Bhopal

Priyanka Dixit
M.Tech CS Dept, SSSIST Sehore, RGPV Bhopal

Abstract

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to digital businesses. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future. Prime concern of the proposed research is to analyze, and design an efficient and highly secured mechanism for web log mining. The proposed technique is concentrating and supporting to the basic log mining principal during analysis and design of the whole process. Expected results are showing the superiority of the proposed technique.

Keywords: *Server Log File, Data Mining, Web Mining, Web Log Mining, Association Rules, Apriori Algorithm.*

1. Introduction

Web log mining is another category in web mining. This type of web mining allows for the collection of Web access information for Web log this usage data provides the paths leading to accessed Web log record. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access [7]. Log files are files that list the actions that have been occurred. These log files reside in the web server. Computers that deliver the web pages are called as web servers. The Web server stores all of the files necessary to display the Web pages on the user's computer. All the individual web pages combines together to form the completeness of a Web site. Images/graphic files and any scripts that make dynamic elements of the site function. , The browser requests the data from the Web server, and using HTTP, the server delivers the data back to the browser that had requested the web page. The browser in turn converts s, or formats, the files into a user viewable page. This gets displayed in the browser. In the same way the server can send the files to many client computers at the same time, allowing multiple clients to view the same page simultaneously. The Log files in different web servers maintain different types of information. [11]The basic information present in the log file is

- User name: This identifies who had visited the web site. The identification of the user mostly would be the IP address that is assigned by the Internet Service provider (ISP). This may be a temporary address that has been assigned. Therefore here the unique identification of the user is lagging. In some web sites the user identification is made by getting the user profile and allows them to access the web site by using a user name and password. In this kind of access the user is being identified uniquely so that the revisit of the user can also be identified
- Visiting Path: The path taken by the user while visiting the web site. This may be by using the URL directly or by clicking on a link or trough a search engine.

- Path Traversed: This identifies the path taken by the user within the web site using the various links.
- Time stamp: The time spent by the user in each web page while surfing through the web site. This is identified as the session.
- Page last visited: The page that was visited by the user before he or she leaves the web site.
- Success rate: The success rate of the web site can be determined by the number of downloads made and the number copying activity under r gone by the user. If any purchase of things or software made, this would also add up the success rate.
- User Agent: This is nothing but the browser from where the user sends the request to the web server. It's just a string describing the type and version of browser software being used.
- URL: The resource accessed by the user. It may be an n HTML page, a CGI program, or a script.
- Request type: The method used for information transfer is noted. The methods like GET, POST A Web log is a file to which the Web server write s information each time a user requests a web site from that particular server.

A log file can be located in three different places [12]:

- Web Servers
- Web proxy Servers
- Client browsers

2. Literature survey

Due to increased usage of web as a medium of publication by entrepreneurs and federal agencies, a massive amount of data is stored in structured and unstructured forms over the web. This leads to page-cluttering, which makes it difficult to distinguish the content of interest from superfluous data such as billboards, and patent and other announcements. Web Usage Mining (WUM) is a type of data mining technique that identifies usage patterns of web data, so as to perceive and better serve the requirements of web applications. The working of WUM involves three steps – pre-processing, pattern discovery and analysis. In this identification of web usage patterns based on the user's interest / choice, thereby creating an intelligent semantics-based web usage mining technique has discussed [1]. In [2] apriori algorithm is an influential data mining algorithm which can mine the frequent sets of Boolean association rules. But its efficiency is not high and cannot do dynamic mining, for these reasons a new association rules algorithm which is suitable for dynamic database mining was proposed. Furthermore, the presented algorithm in [2] is applied to the web log mining. Compared with original algorithm, experiments show that the performance of the presented algorithm is improved to some extent. In [3] is an attempt to apply an efficient web mining algorithm for web log analysis. The results obtained from the web log analysis may be applied to a class of problems; from search engines in order to identify the context on the basis of association to web site design of a ecommerce web portal that demands security. The algorithm is compared with its other earlier incarnation called Improved apriori all Algorithm. It has been shown beyond any doubt through performance analysis that proposed efficient web mining algorithm; Web Miner has much better performance in terms of time and space complexity when compared. Even a tracing of the algorithm shows the inherent flow in the Tong and Pi-lian's algorithm that fails to give correct output. The presented algorithm, Efficient Web Miner or E-Web Miner can be traced for its valid results and can be verified by computational comparative performance analysis. The number of data base scanning's drastically gets reduced in E-Web Miner and the candidate sets are found to be much smaller in stage wise comparison with Improved Apriori All Algorithm of Tong and Pi-lian, E-Web Miner, thus, is successful to be applied in any web log analysis, including information centric network design. In [4] author has analyze web logs using data mining so as to present users with more personalized web content. They classify users based on their internet usage patterns and for each class, maintain a cache of web documents. Searching is performed based on term set analysis and direction cosine distance approach. Finally a more personalized and relevant result set is generated for the query. In [5] Web Mining Systems make use of the redundancy of data published on the Web to automatically extract formation from existing web documents. The crawler is an important module of a web search engine. The quality of acrawler directly affects the searching quality of such web search engines. Such a web crawler may interact with millions of hosts over a period of weeks or months, and thus issues of robustness, flexibility, and manageability are of major importance. Given some URLs, the crawler should retrieve the web pages of those URLs, parse the HTML files, add new URLs into its queue and go back to the first phase of this cycle. The crawler also can retrieve some other information from the HTML files as it is parsing them to get the new URLs. In [6] presented a framework and algorithm, TOPCRAWL for mining. The presented TOPCRAWL algorithm is a new crawling method which emphasis on topic relevancy and outperforms state-of-the-art approaches with respect to recall values achievable within a given period of time. This method also tries to offer the result in community format and it makes use of a new combination of ideas and techniques used to identify and exploit navigational structures of websites, such as hierarchies, lists or maps. This algorithm is simulated with web mining tool Deixto and the basic idea has been implemented using the JAVA and Results are given.

Analysis: Inference on user's next page appeal follows the mining of web log as an additional phase. Until now several discoveries have made successfully. Yet, there exists certain constraints in the analysis of usage logs and extraction of required information as follows.

• Limited types of available usage logs and lack of standard ways to infer the meaning of usage patterns.

• Uncertainty of usage patterns as a result of several contextual factors concurrently influencing the usage patterns.

• User experiences in the historical aspect influencing variation of usage patterns.

Considering these difficulties, it can be clearly understood that studying short-term usage alone cannot sufficiently perform the analysis a web usage pattern since it is supposed to be the time-consuming process. Moreover, the usage data should be extracted at browser side in a real web environment over a long time period without limitation on the specificity of web server.

3. Proposed work

An emerging challenge for data mining is the problem of mining richly structured datasets, where the objects are linked in some way. Many real-world datasets describe a variety of object types log via multiple types of relations. These log provide additional context that can be helpful for many data mining tasks. Logs among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Log mining is a newly emerging research area that is at the intersection of the work in log analysis [8, 9], hypertext and web mining [10]. Logs have more generically relationships, among data instances are ubiquitous. These logs often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object. In some cases, not all logs will be observed. Therefore, it may be interested in predicting the existence of logs between instances. In other domains, where the logs are evolving over time, our goal may be to predict whether a logs will use or not in the future, given the previously observed logs and enhance performance of system. By taking logs into account, more complex patterns arise as well. This leads to other challenges focused on discovering substructures, such as communities, groups, or common sub graphs.

The figure 1 based on web log mining block diagram for proposes technique, the basic interactive elements for web logs Miner. Rule base of web logs miner should consist of

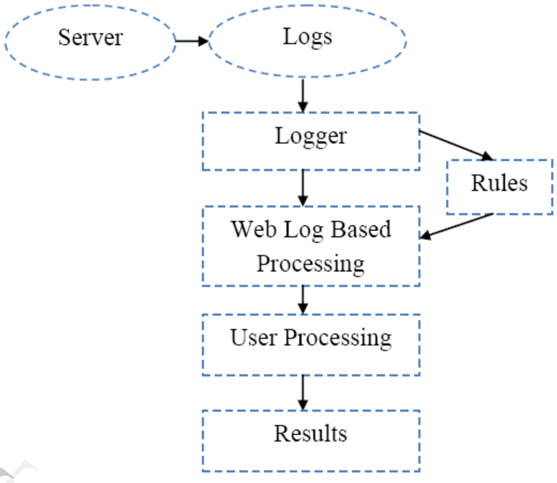

Figure 1: Block Diagram of Proposed Technique

- Database mechanism in data/knowledge base
- Rule base configuration of base line

The rule base generations secure web log Miner mines web logs in a secured way by defining a property set and rule base configuration mechanism. web logs Miner is the proposed log mining technique that removes the flaws of previous technique [1, 2, 3] and improves upon the complexity in term of time of the earlier presented algorithms. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines quickly result of candidate set where as the existing algorithm take too much time to deliver the result. The algorithm will be designed independent of existing apriori algorithm.

Characteristic: Proposed concept has many attribute in which some of them are as follow.

Reliability and Fault Tolerance: The Proposed concept will reliable and of enterprise quality.

Availability and Security: we can place in public to the proposed concept. Anyone can use for further research. The proposed technique will strong web algorithm.

Correctness and Consistency: The specification of proposed technique is correct and consistent.

Portability and Performance: The proposed concept will develop in suitable language. The requirements of the computers do not important for proposed technique. It may be important for big sized files

4. Result and Conclusions

For the experiments we will use suitable the data sources from standard database which is available on different web sites. Here we will use some performance factors like CPU Utilization, execution time, throughput, which will define on different data volume and minimum support. For this performance factor we will try to improve result of proposed model. Throughput of any technique can be calculated by using execution time. This indicates technique speed. The throughput of the technique is the ratio of the total data during execution and the total execution time. The purpose of the proposed model is to improve efficiency. To achieve this we are combining the some basic rules with effective filtering approach to improve performance of the proposed technique. The expected experimental results are showing the superiority of the proposed technique in terms of the execution time, and throughput. During experiments, the proposed technique will executes different set of larger data volume. Three performance factor will participate for calculating by the proposed technique which is already defined above. Expected results shown in table 1, 2, & 3 the proposed technique will execute approximately hundred times and every time, same log data will respectively execute by **"Proposed technique"**. Size of the selected data will be same in every time.
Execution Time: - "The Proposed technique" Execution time of various logs data comparisons shown in table 1.

Table 1: Expected Execution Time of Proposed Concept

S.NO	Data Volume	Proposed Algorithm
1	1000	Low
2	2000	Low
3	3000	Low

Throughput: Throughput can be calculated by using execution time. It denotes the speed of execution. The throughput of the execution scheme is calculated as in equation (1).

Throughput of Execution = Total Size of Log Data/ Total Execution time (1).

Table 2: Expected Throughput of Proposed Concept

CPU Consumption: CPU utilization of 1000 to 3000 data logs comparisons shown in table 3.

Table 3: Expected CPU Utilization of Proposed Concept

S.NO	Data Volume	Proposed Algorithm
		CPU uses in % (Approx)
1	1000	High
2	2000	High
3	3000	High

S.NO	Data Volume	Proposed Algorithm
		Throughput (Approx)
1	1000	High
2	2000	High
3	3000	High

Summary: From the table 1, we are expecting: In the same data size, with minimum support reduce gradually, the proposed concept grow slowly. What's more, the time spending of improved model always much less than previous? From the table 1, we are expecting: In the same minimum support, with the increase of data quantity, the time cost of previous technique can be increases but the proposed model will not. Moreover, the former is the nearly 10 times of the latter under each date volume.

Conclusion: The Paper gives a detailed look about the web log mining & log files, its contents, its types, its location etc., the various mechanisms that perform each step in mining the log file is being discussed along with their disadvantages. The performance factor that can be increased efficiency for Log mining technique and the proposed technique in creating the improved log mining technique is also discussed briefly. The proposed work is to combine the concept of log mining the ser's area of interest.

References

[1] K. Sudheer Reddy, G. Partha Saradhi Varma and S. Sai Satyanarayana Reddy Understanding the Scope of Web Usage Mining & Applications of Web Data Usage PatternsIEEE International Conference 2012

[2] RuPeng Luan*, SuFen Sun, JunFeng Zhang, Feng Yu, Qian Zhang A Dynamic Improved Apriori Algorithm and Its Experiments in Web Log Mining 9th IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012) 2012

[3] Mahendra Pratap Yadav, Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar An

Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner 1st IEEE Int’l Conf. on Recent Advances in Information Technology | RAIT-2012 |

[4] Indrajit Mukherjee, V. Bhattacharya, Samudra Banerjee, Pradeep Kumar Gupta and P. K. Mahanti Efficient Web Information Retrieval based on U sage Mining].1 Infl Conf. on Recent Advances in Information Technology I RAIT-20121

[5] S. Balaji and S. Sarumathi TOPCRAWL: Community Mining in Web search Engines with emphasize on Topical crawling Proceedings of the IEEE International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012

[6] K. R. Suneetha, Dr. R. Krishnamoorthi- “Identifying User Behavior by Analyzing Web Server Access Log File”, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009

[7] Mark E. Snyder, Ravi Sundaram, Mayur Thakur- “Preprocessing DNS Log Data for Effective Data Mining”, 2008.

[8] S. Sun and J. Zambreno, "Mining Association Rules with Systolic Trees," Proc. In!'1 Conf Field-Programmable Logic and Applications (FPL '08), Sept 2008.

[9] G. Stumme, A. Ho tho, and B. Berendt. Semantic web mining: State of the art and future directions. Journal of Web Semantics: Science, Services and Agents on the World WideWeb, 4(2):124–143, 2006.

[10] R. R. Sarukkai. Link prediction and path analysis using markov chains. In Proceedings of the 9th Intl. World Wide Web Conf. (WWW’00), pages 377–386, 2000.

[11] Ratnesh Kumar Jain , Dr. R. S. Kasana1, Dr. Suresh Jain, (July 2009)“Efficient Web Log Mining using Doubly Linked Tree,” International Journal of Computer Science and Information Security, IJCSIS, vol. 3.

[12] K. R. Suneetha, and R. Krishnamoorthi,(April 2009 “Identifying User Behavior by Analyzing Web Server ccess Log File,” IJCSNS International Journal of Computer Science and Network Security, vol. 9, pp. 327-332.