A New Algorithm for Web Log Mining

Gajendra Singh
CS Dept, RGPV Bhopal, SSSIST Sehore
Bhopal, Madhya Pradesh, India

Priyanka Dixit
CS Dept, RGPV Bhopal, SSSIST Sehore
Bhopal, Madhya Pradesh, India

ABSTRACT

The enormous content of information on the World Wide Web makes it obvious candidate for data mining research. Data Mining Technique application is used to the World Wide Web referred as Web mining where this term has been used in three distinct ways; , Web Structure Mining, Web Content Mining and Web Usage Mining. Web Log Mining is one of the Web based application where it will facing with large amount of log data. In order to produce the web log through portal usage patterns and user behaviors, this research work implements the high level process of Web Log mining technique using basic rules. Web Log Mining consists of three main phases, namely Data Preprocessing, Pattern filtering and Pattern Analysis. As we know that server log files become a set of raw data where it's must go through with all the Web Log Mining phases to producing the final results. Here, Web Log mining, approach has been combining with the basic rules, to optimize the total execution time. Finally, this work wills present an overview of results analysis and Web Log Mining can use the findings for the suitable valuable actions.

Keywords

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

1. INTRODUCTION

Data mining is a technique used to deduce useful and relevant information to guide professional decisions and other scientific research. It is a cost-effective way of analyzing large amounts of log data, especially when a human could not analyze such datasets [1]. Mystification of the use the internet has made automatic knowledge extraction from Web log files a necessity. Information provided are interested in techniques that could learn Web users' information needs and preferences. This can improve the effectiveness of their Web sites by adapting the information structure of the sites to the users' behavior [2]. Recently, the advent of data mining techniques for discovering usage pattern from Web data (Web Usage Mining) indicates that these techniques can be a viable alternative to traditional decision making tools. Web Log Mining is the process of applying data mining techniques to the discovery of log patterns from Web data and is targeted towards applications. Web Log Mining mines the log data (Web server access logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, mouse clicks and any other data as the result of interaction with the Web) derived from the interactions of the users during certain period of Web sessions [3-5]. This work explores the use of Web Log mining techniques to analyze Web log records collected from proposed system. There are several commercial data Web mining tools to identify several Web access pattern by applying well known data mining techniques like Apriori Algorithm to the access logs of this educational portal. This includes descriptive statistic and Association Rules for the portal including support and confidence to represent the Web log [6-7].

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 [8-10]. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future.

2. PROPOSED WORK

It's already known that the association rule mining task is one of the most important tasks in among all other data mining tasks. It is support to wide range of ecommerce applications. Apriori algorithm mining association rules only in static transactional database, but in actual data mining, the data in database are always changed, and transaction data and access logs are always updated. Therefore, it is necessary to research proposed algorithm based on different concept and achieve highly efficient recommendation program for single item. Proposed Web Log Miner is the proposed web mining algorithm that removes the flaws of previous algorithm [1-4] and improves upon the time complexity of the earlier suggested algorithms. It provides an improved candidate set pruning as well. In fact, it has been shown successfully that it mines correct result of candidate set whereas the previous algorithms fails to deliver the correct result [11-12]. The algorithm has been designed independent of previous algorithms. Figure 1 is showing the comparatively diagram between two previous techniques with proposed technique like apriori, dynamic and proposed algorithm and it is clearly seen from figure that proposed algorithm removing some steps as compare other than two. Due to this proposed algorithm improving overall performance in term of time complexity. The main objective of the proposed concept is to reduce total number of elements in each candidate set without any repeating the step which is allowing changes in the larger log record sets. The observations or analysis are noted as follows: Firstly, candidate record set pruning gets reduced in steps. Secondly, by pruning, the number of element of candidate record set is decreased remarkable. Repetitive scanning of log database is totally eliminated. The Proposed Web Log Miner is proved to be highly efficient in terms of time. The study has been performed only on limited number of fictitious data. The higher number of web log records puts enough demand on CPU time too. If it is too large, then the server data base during preprocessing demand CPU Time. Basically this paper is going to be present general idea on a new proposed concept for log mining system which will enhance efficiency as compare existing log mining system.

The proposed concept is using data mining techniques. Data mining techniques have been successfully applied in many different fields including network management, manufacturing, process control, marketing, and fraud detection. Over the previous years, a growing number of research techniques have applied data mining to various problems in log mining. In this will apply to data mining for log mining field of web log mining. Presently, it is unfeasible for several computer systems to affirm efficiency to redundancy in data set with computers increasingly getting connected to easily accessible.

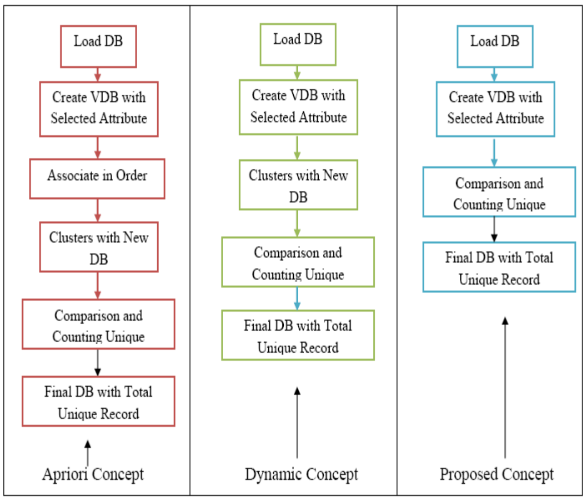

Load DB
Create VDB with Selected Attribute
Associate in Order
Clusters with New DB
Comparison and Counting Unique
Final DB with Total Unique Record
Apriori Concept
Load DB
Create VDB with Selected Attribute
Clusters with New DB
Comparison and Counting Unique
Final DB with Total Unique Record
Dynamic Concept
Load DB
Create VDB with Selected Attribute
Comparison and Counting Unique
Final DB with Total Unique Record
Proposed Concept

Figure 1: Comparative Diagram between Existing and Proposed Concept

2.1 Architecture of Proposed web Log Miners

This titled "A New Algorithm for Web Log Miners" is proposed for Web log mining by analyzing the principle of the log mining algorithm. Proposed algorithm will support to effectiveness, authorization, accuracy of log record set which is accessing from log database. As we know that, a previous algorithm requires more effort during log miners. This research introduces a new algorithm which is the reducing number of steps as compare previous. The user will start with start function and firstly it will enter user login and password. If it does not match then it will exit otherwise it will proceed for further process. In the next process user will load server log record data set. In this log record data set seven attribute have defined which is shown in table 1. Presented table1 is showing general view of log record.

Table 1: User Table

LogId	LogDate	LogTime	ExitDate	ExitTime	UserId	UserName	ProcessId	ProcessName
1	06/16/2008	18:30:29	06/16/2008	18:46:21	12274272	chandan_singh954@yahoo.com	7	Design Process

After load database form server, another virtual database will create with filtered attribute which is shown in table 2. Presented table 2 is showing the general record sets.

Table 2: Virtual Table

LogID	Name
1	chandan_singh954@yahoo.com
2	Rohit_nagar123@yahoo.com

After that selected first record set from virtual table (table 2) and compare this record set in whole virtual data table is its find the a counter will increase with one, after completing this, delete selected record set and move another record set. This process will continue till all record find uniquely with counter variable. After completing whole process finally another new database will create to show the information like table 3.

Table 3: Final Table

Name	NoOfVisit
chandan_singh954@yahoo.com	333
Rohit_nagar123@yahoo.com	333
Rajat_kumar543@gmail.com	334

The Architecture of proposed concept is shown in figure 2 for log miners of a server log using suitable user-defined concept.

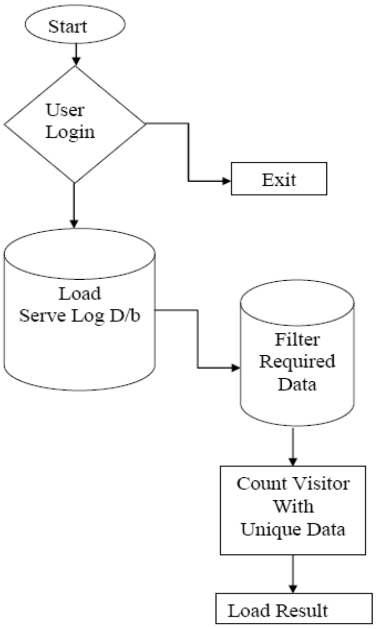


Figure 2: Architecture of Proposed Concept

Strength of the Proposed System:

- Proposed system is batter then comparing system to log mining performance.
- Proposed system is faster than comparing system in terms of execution time.
- Proposed system will be smaller than the compared system and easy to understand and implement.
- It will do not contain complex structure, control flow will be well defined and looping structure will be minimized. Due to the following facts it will take very less time for execution.

3. RESULTS ANALYSIS

For the experiments and contrast of three kind of model, we have used suitable the data sources from standard database which is available on different web sites. Here we set some parameters (Execution Time, Throughput, and CPU Utilization) which define on minimum support and different data volume. For these parameters we improved result of proposed model as compared previous one. Execution time is used to calculate the throughput of any technique. It indicates the speed of technique. The throughput of the any technique is calculated as the total data for execution divided by the total execution time. Proposed system considers the key value as a criterion to evaluate the performance of the proposed concept. The purpose of the proposed concept is to high efficiency. To achieve this by combining the basic rules with effective filtering technique to increase performance of the proposed system. The presented experimental results show the superiority of the proposed technique in terms of the processing time, and throughput. Desktop machine has been used to calculate experimental results which has following configuration (See table 3)

Table 3: Configuration

S. No.	Processor	Memory(Primary)	Platform	Software Application
1	Intel Pentium Dual Core E2200 2.20 GHz	1 GB of RAM	Window-XP SP2	Java (JDK Net Been 7.1)

In the experiments, the system executes a large log data with different data volume. There are three parameters used for calculating by the proposed system one is execution time, second is throughput, and third is CPU Consumption which is shown in table 4, 5, & 6 the proposed system has run hundred times approximately. In each time, same log data are respectively executed by existing system and **"Proposed system"** by copying them. Size of the selected data was same in each time. Finally, the outputs of the comparison system are execution time and throughput and CPU consumption which is noted in numeric form.

Execution Time: - "The Proposed System" system has been implemented on a number of data logs varying types of content and sizes of a wide range. Execution time of various logs data comparisons shown in table 4.

Table 4: Execution Time of Proposed Concept

S.NO	Data Volume	Proposed Algorithm
1	1000	2300
2	2000	4600
3	3000	6930

3.1 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).

Where Size is measuring in bytes and Execution times are measuring in execution time

For Example: Here selected file of 1000 Record.

Throughput of Proposed Concept

Encryption Throughput = 1000/2300 = 0.43

Table 5: Throughput Comparisons between Proposed and Existing Concept

S.NO	Data Volume	Proposed Algorithm
		CPU uses in % (Approx)
1	1000	59
2	2000	59
3	3000	59

3.4 CPU Consumption: "The Proposed System"

System has been implemented on a number of data logs varying types of content and sizes of a wide range. CPU utilization of 1000 to 3000 data logs comparisons shown in table 6.

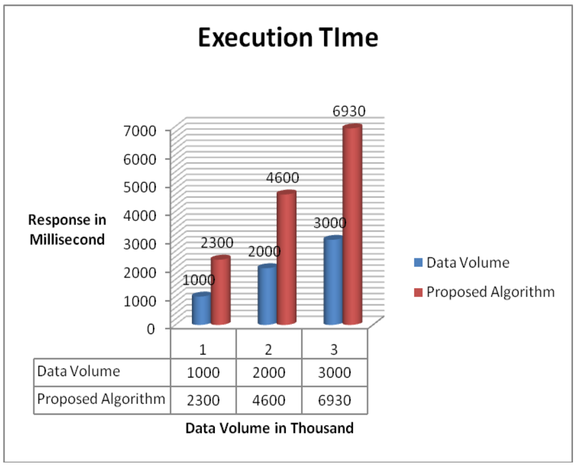
S.NO	Data Volume	Proposed Algorithm
		Throughput (Approx)
1	1000	0.434
2	2000	0.434
3	3000	0.432

Table 6: CPU Utilization Comparisons between Proposed and Existing Concept

4. SUMMARY

From the table 4, we are expecting: In the same data size, with minimum support reduce gradually, existing model execution time increases rapidly and the proposed model grow slowly. What's more, the time spending of improved model always much less than previous [2]. From the table 4, we are expecting: In the same minimum support, with the increase of data quantity, the time cost of existing model increases rapidly but the proposed model is not. Moreover, the former is the nearly 10 times of the latter under each date volume.

A graphical representation for the table 4 is shown in graph 1 with red line for 1000, for 2000 and for 3000 data volume for proposed algorithm. According to the graph, required time for the execution through proposed System.

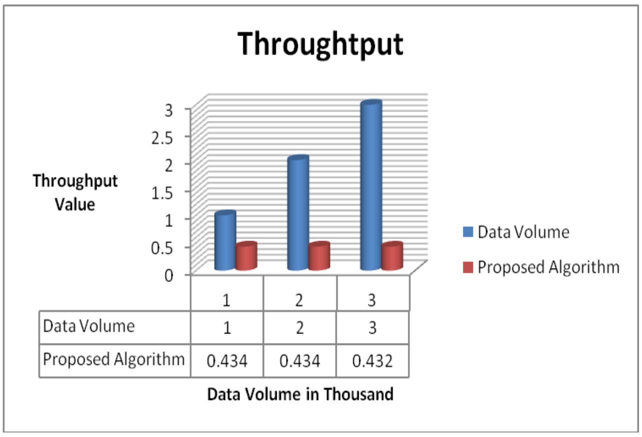


	1	2	3
Data Volume	1000	2000	3000
Proposed Algorithm	2300	4600	6930

Graph 1: Execution Time Comparison

4.1 Throughput

A graphical representation for the table 5 is shown in graph 2 with red line for 1000, for 2000 and line for 3000 data volume for existing as well as proposed algorithm.. According to the graph, total throughput for the execution through Proposed System is much better.

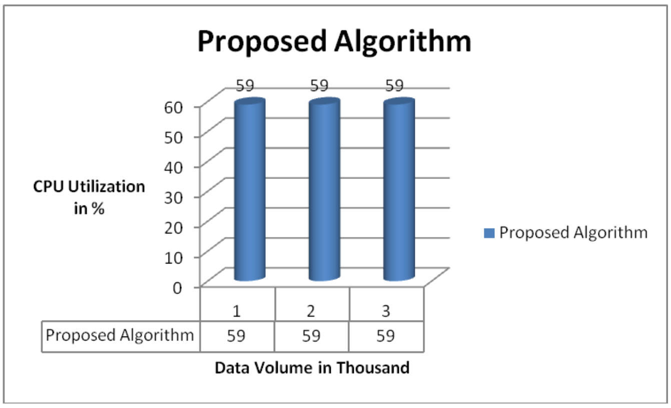


	1	2	3
Data Volume	1	2	3
Proposed Algorithm	0.434	0.434	0.432

Graph 5: Throughput Comparisons

4.2 CPU Utilization

A graphical representation for the table 6 is shown in graph 3 with blue line for 1000, for 2000 and line for 3000 data volume for existing as well as proposed algorithm. According to the graph, total CPU consumption during execution through Proposed System is much better.



	1	2	3
Proposed Algorithm	59	59	59

Graph 3: CPU Utilization Comparisons

5. RESULTS ANALYSIS

From the above discussion it can clearly see that the proposed Concept producing good results as compare existing concept which is defined in [2] and hence can be incorporated in the process of execution of large data logs. Also, I can see that the previous system have very less efficiency in terms of execution time and hence cannot be used for larger log record data set. The proposed system is good than previous system as they have higher efficiency. However it is also clear from table 4 to 6 and Graphs 1. To 3 that, by applying proposed concept to the log data set of different volume highly efficiency is obtaining as compare to different other concept. In execution time, CPU uses and RAM Uses the proposed system have quite good results as compared to different other system. Table 4 showing the execution time where various log data set are producing different time according to volume of log data set, if 2000 record set are executing through our proposed concept taking 4600 millisecond time to execute at the time of processing.

6. CONCLUSION

This research presents a Data base logs miner model as well as concept which is suitable for Data base log mining. The proposed connect has the following innovation:

- For a particular page, greatly reduce the search range. Can achieve real-time updates and improve the efficiency of the existing model.
- Minimum support has page pertinence, for the page with very small click rate can also find its associated rules.
- Mark association rules with flag, convenient location to the target page association rules.

As shows in experiment of the proposed model showing the improperness in the performance .This paper presents a rule discovery model which is suitable for web log mining. The model has the following innovation:

- For a particular data base, greatly reduce the search range. Can achieve real-time updates and improve the efficiency of the proposed model.
- Minsup has page pertinence, for the page with very small click rate can also find its associated rules.

- The algorithm not only gave the solution of "rare item" problem, but also avoids the problem of "combinatorial explosion". As shows in experiment, proposed algorithm's performance has improved.

In future trying to improve security level of the proposed algorithm .It will also try to resolve limitation of image type that mean any type of image will be encrypt and decrypt through proposed algorithm Further development of the algorithm to accommodate tighter generic security reductions for image encryption is therefore desirable. To be precise, the future work will evaluate performance of the proposed measures considering inconsistent data as well as a validation of sites using dynamic pages, and cookies.

7. HIGHLIGHTS OF FUTURE WORK

- Application of proposed techniques to pattern-growth sequential pattern mining algorithms.
- An in-depth research into concept generality, and the impact of the inclusion of semantics on countering complicated pattern queries with enhanced precision

8. 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] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In Proceedings of the 5th Int'l Conference on Extending Database Technology: Advances in Database Technology, pages 3–17, 1996

[12] Personalized Web Search by Mapping User Queries to Categories- Fang Liu Clement Yu Weiyi Meng Department of Computer Science, Department of Computer Science, Department of Computer Science, University of Illinois at Chicago University of Illinois at Chicago SUNY at Binghamton Chicago.