Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Design and Implementation of New Web Log Miner Approach for Web Log Mining

Samidha D. Sharma
Dept. of IT,
Rajiv Gandhi Proudyogiki Vishwavidyalaya, India

Poulami Das
Dept. of IT,
Rajiv Gandhi Proudyogiki Vishwavidyalaya, India

Abstract: The enormous content of information on the World Wide Web makes it obvious candidate for data mining research. All most all the application of data mining techniques to the World Wide Web referred as Web mining where this term has been used in three distinct ways; Web Structure Mining , Web Usage Mining and Web Content Mining. E Learning is the Web based application where it will facing with large amount of data. In order to produce the server logs patterns and user performance, this work implements the high level process of Web Log Mining using some modification in basic Association Rules algorithm call modified Apriori Algorithm. Web Log Mining consists of three main phases, namely Preprocessing of Log Record, Discovering of Pattern and Analysis of Pattern. 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 Association Rules, with modified Apriori Algorithm to optimize the content of the server log record. Finally, this work will present an overview of results analysis and Web administrator can use the findings for the suitable valuable actions.

Key Words: server log file, data mining, Web mining, Web Usage Mining, Association Rules, Apriori algorithm.

I. INTRODUCTION

Web technology is not evolving in comfortable and incremental steps, but it is erratic, turbulent, and often rather uncomfortable. It is estimated that the Internet, arguably the most important part of the new technological environment, has expanded by about 2000 % and that is doubling in size every six to ten months [10]. In recent years, the advance in computer and web technologies and the decrease in their cost have expanded the means available to collect and store data. As an intermediate consequence, the amount of information (Meaningful data) stored has been increasing at a very fast pace. Traditional information analysis techniques are useful to create informative reports from data and to confirm predefined hypothesis about the data [11, 12]. However, huge volumes of data being collected create new challenges for such techniques as organizations look for ways to make use of the stored information to gain an edge over competitors. It is reasonable to believe that data collected over an extended period contains hidden knowledge about the business or patterns characterizing customer profile and behavior [13, 14]. With the rapid growth of the World Wide Web, the study of knowledge discovery in web information, modeling information and predicting the user's access on web site information has become very important. From the application point of view, business and administration, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-CRM, e-Business, e-Services, e-Newspapers, e-Education, Digital Libraries e-Government, and so on [3]. Web is becoming the necessity of the businesses and organizations because of its demand from the clients. Since the web technology largely feeds on ideas and knowledge rather than being dependent on fixed assets, it gave birth to new companies such as Google, Netscape, Yahoo, e-Bay, Expedia, Amazon, e-Trade and so on [6]. With the large number of companies using the Internet to distribute and collect information/data, knowledge discovery on the web sites has become an important research area. With the explosive growth of information sources available on the World Wide Web, it has become necessary for organizations to discover the usage patterns and analyze the discovered patterns to gain an edge over competitors. In the recent years the World Wide Web has become the preferred platform for developing Internet applications and thanks to its powerful communication paradigm which is based on browsing and multimedia contents, and to its open architectural standards which facilitate the integration of different types of content and systems [1, 2]. Current Web applications are very complex and high sophisticated software products, where its successes and failure depend on the quality, as perceived by users. A number of methods have been proposed for evaluating their effectiveness in content/information delivery. Content/information personalization, for instance, aims at tailoring the Web contents to the final recipients according to their profiles. Here some another approach is the adoption of Web Log Mining techniques for the analysis of the navigational behaviour of Web users by means of the discovery of patterns in the Web server log [4, 5]. Traditionally, to be effective, Web log mining requires some additional pre-processing, such as the application of methods of page annotation for the extraction of meta-data about page semantics or for the construction of Web site ontology. This proposed work, propose a novel approach to Web Log Mining. It has the advantage of integrating Web log mining goals directly into the Web application development process. Thanks to the adoption of a conceptual modelling method for Web application design, and of its supporting case tool, the generated/developed Web applications embed a logging mechanism that - by means of a synchronization tool - is able to produce semantically enriched Web log files [7]. This log, that call conceptual log, contains additional information with respect to standard Web server logs and some of this information is useful to the Web mining process which refers not only to the composition of Web pages in terms of atomic units of contents, to the conceptual entities Web pages deal with, but refers also to the identifier of the user crawling session, to the specific data instances that are published within dynamic pages, as well as to some data concerning the topology of the hypertext. Therefore, any extra effort is needed during or after the application development for reconstructing and analyzing usage behaviour. The main contribution of this work comes from existing frameworks. First one is the model-based design and development of an applications based on the Web Log Mining and its supporting tool [7, 8]. The second one is an evaluation of the developed applications based on data mining analytics that had started by collecting the application data. The evaluation of the developed application aimed at studying its suitability to respond to users' needs by observing their most frequent log pattern [9].

.

II. PROPOSED CONCEPT

The proposed approach is removing some steps as compare other than previous approaches which is defined in [5]. Due to this reason performance of the proposed approach is enhancing on selected parameters. The prime concerned of the proposed work 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 Chapter is going to be present a new proposed approach for log mining to enhance efficiency as compare previous log mining approach. The proposed approach is based on data mining techniques. The proposed work is for Web log mining by analyzing the principle of the log mining algorithm. Proposed approach support authentication & authorization with high efficiency effectiveness & accuracy over log record set which is accessing from server log database. It's known that, pervious approaches require to improvement on some performance parameter like efficiency and efficiency can be measure in terms of execution speeds. The proposed work proposes a new log mining approach. The proposed work is trying to enhance previous approach of log mining. In this proposed work have design an architecture where user first login with its id and password and will go for execute proposed approach. It has included some previous approach and a results portion which will compare to all approaches in to find out the differences between proposed approach and previous approach. The Flow Architecture of proposed concept is shown in figure 1 for log miners of a server log using suitable user-defined concept.

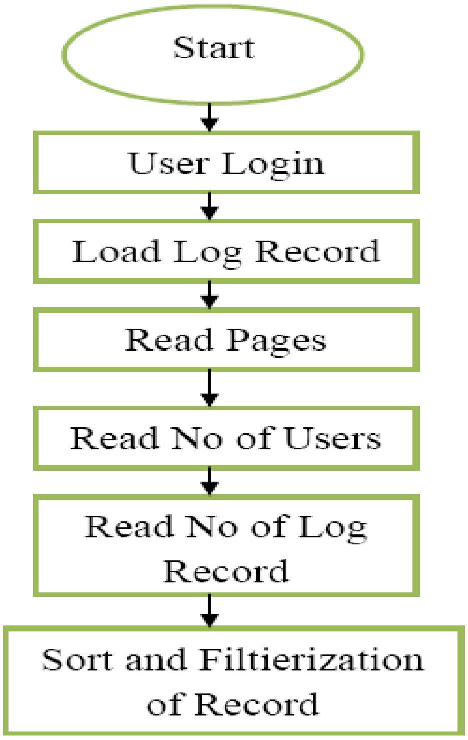


Figure 1: Flow Architecture of Proposed Approach

Some of the principle and nature which recommended approach is based on were introduced:
Theorem one: sort the dataset in default order on the basis of main identifier for quick access.
Theorem two: frequent item sets' all nonempty set must be frequently [2].
Theorem three: non-frequent item sets' superset is not frequent item sets [2].
Theorem four: filter unique items from item sets.

Collaborative of sorting algorithms based on nature one where collaborative filtering algorithm based on nature two, and content-based recommendation algorithm based on nature three. Nature four is based on collaborative filtering algorithm Latter Apriori algorithm applied. These four theorems are used to determine/find out the frequent item sets' generation rules. Web log can be updated at any time so that generation of all association rules do not require every time, association rules related with real-time mining is used for specific pages. In while same user wants access to another page at the same time then he will saw the huge difference. Suppose user want to find the a target page for P. the given minimal support is m = XY %. The steps of the given process are as flow:

- Input total number of Users (Ui) where { i = 1 to n}
- Visit Web Log Database (X),
- Users (Ui) Visit the Pages (Pj) where { j = 1 to n}
- Read total web log record (Nk) where { k = 1 to n} with Total Users (Uj) per Pages (Pj)
- Sort total users (Ui) corresponding pages (Pj).
- Delete users (Ui) from any type of mismatching web log record (Nk) from Web Log Database (X)
- Filter record set of Users (Ui) from web log record (Nk) corresponding arranges or create sets of pages (Pj).

Example: For example in this log data record following attributes are selected correspondence users.

- LogId
- LogData
- LogTime
- ExitDate
- ExitTime
- UserID
- UserName
- ProcessId
- ProcessName

Formal representation of all theses attribute within values in log record set is shown in Table I.

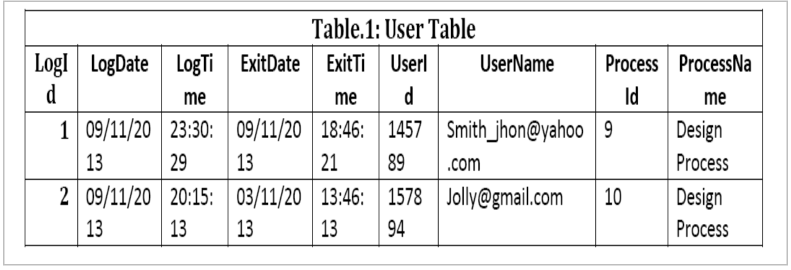


Table.1: User Table

LogId	LogDate	LogTime	ExitDate	ExitTime	UserId	UserName	Process Id	ProcessName
1	09/11/2013	23:30:29	09/11/2013	18:46:21	145789	Smith_jhon@yahoo.com	9	Design Process
2	09/11/2013	20:15:13	03/11/2013	13:46:13	157894	Jolly@gmail.com	10	Design Process

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

TABLE II: VIRTUAL TABLE

LogID	Name
1	Smith_jhon@yahoo.com
2	jolly@gmail.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 III.

TABLE III: FINAL TABLE

Name	NoOfVisit
Smith_jhon@yahoo.com	445
jolly@gmail.com	446
albertparera@gmail.com	432

Characteristic of the Proposed Approach:-

- Performance of Proposed Approach is batter.
- Easy to Understand
- No Complex Structure
- Less Execution Time

III. RESULTS

- Performance Analysis: This section presents the results of evaluating the efficiency of the proposed technique that is based on selected parameters. For an algorithm it is important to be efficient and secure. Efficiency of an algorithm is computed on the bases of time complexity and space complexity.
- Execution Time
- CPU Process Time
- Throughput
- Memory Utilization

The execution time is considered the time that an encryption algorithm takes to produce a cipher text from a plaintext. Execution time is used to calculate the throughput of an encryption scheme which is indicates to the speed of encryption process. The throughput of the encryption scheme/technique is calculated as the total plaintext in bytes encrypted divided by the execution time.

The CPU process time is the time where a CPU is committed only to the particular process for calculations which reflects to the load of the CPU. The higher CPU time is used in the encryption process/technique, CPU load will be higher. The memory deals with the amount of memory space it takes for the whole process of encryption and decryption. For the experiments and contrast of three kind of model. In this used suitable the data sources from standard database which is available on different web sites. Here set some parameters which define above on minimum support and different data volume. For these parameters improved result of proposed model as compared existing model. 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 technique. The prime factor of the proposed technique is to produce 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 over other technique in terms of the processing time, and throughput. Desktop machine has been used to calculate experimental results which has following configuration (See table IV).

TABLE IV: 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)

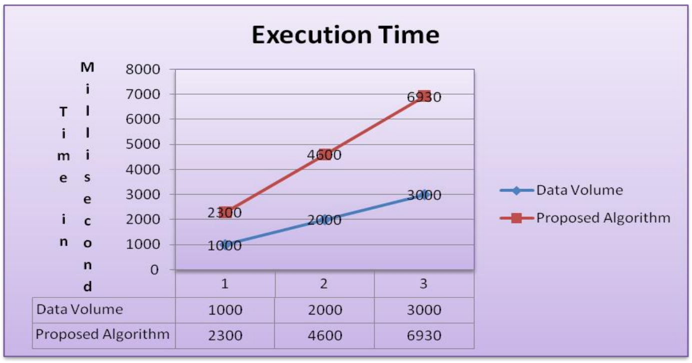
In the experiments, the system executes a large log data with different data volume. There are three parameters used for calculating by the proposed system one is execution time, second is throughput, and third is CPU Consumption which is shown in table 5, 6, 7. and 8 The proposed system has run hundred 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.

Tabular Analysis: - In this compared results were presented in the form of tables.

Execution Time: - "The Proposed System" and existing system have 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 V.

TABLE V: EXECUTION TIME COMPARISON BETWEEN PROPOSED CONCEPT & EXISTING CONCEPT

S.NO	Data Volume	Proposed Algorithm
		Execution Time in Millisecond(Approx)
1	1000	2300
2	2000	4600
3	3000	6930



Graph 1: Execution Time Analysis

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.

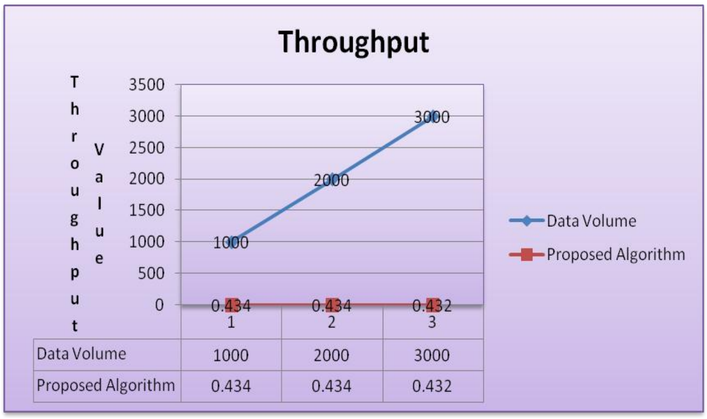
Throughput of Proposed Concept

Encryption Throughput = 1000/2300

= 0.43

TABLE VI: THROUGHPUT ANALYSIS OF PROPOSED CONCEPT

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

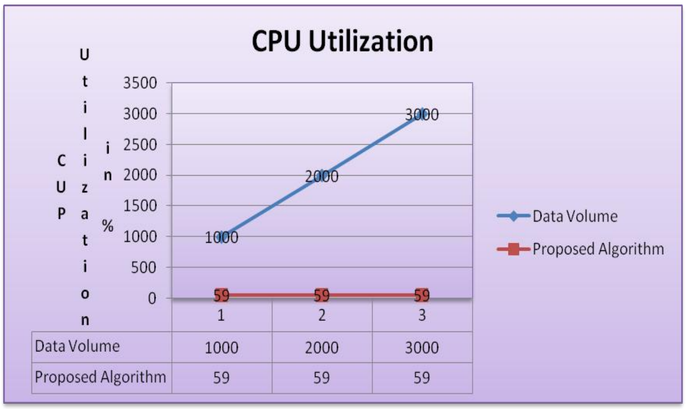


Graph 2: Throughput Analysis

CPU Consumption: "The Proposed System" have 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 VII.

TABLE VII: CPU UTILIZATION OF PROPOSED CONCEPT

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

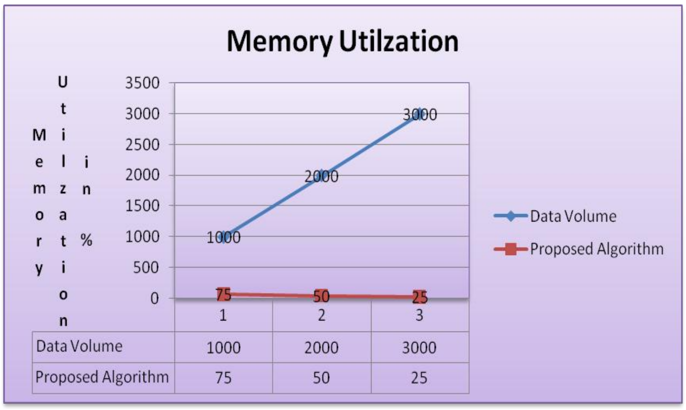


Graph 3: CPU Analysis

Memory Consumption: “The Proposed System” have been implemented on a number of log data set varying types of content and sizes of a wide range. Memory Utilization of 1000 to 3000 data logs comparisons shown in table VIII.

TABLE VIII: MEMORY UTILIZATION OF PROPOSED CONCEPT

S.NO	Data Volume	Proposed Algorithm
		Memory uses in % (Approx)
1	1000	75
2	2000	50
3	3000	25



Graph 4: Memory Analysis

Summary: From the table 2, 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 existing model. From the table 2, 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.

Results Analysis: From the above discussion it can clearly see that the proposed Concept producing good results and hence can be incorporated in the process of execution of large data logs record. Also, there can see that the earlier systems 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 as they have higher efficiency with effectiveness. However it is also cleared from table 2 to 4 and Graphs 1. To 4 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. Table 2 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.

IV. CONCLUSION

Web Log Mining is an active field for research and it will generate new hopes in internet based business. Web Log Mining applications are being used in Websites and this work totally focused on server log record. This work presents a brief introduction of Web mining technique, a part of the data mining technologies and also the implementation of the Web Log Mining tools. Any Server log files from server host, selected for this work. In order to perform the Web Log Mining, the methodology that being introduce in [5] it includes three main phases; Data Preprocessing (Server Log Record), Pattern Analysis of log record and Pattern Discovery from log record. All the three phases were done carefully to produce quality results. Data Processing phase for the Web Log Mining is a challenging task and basic Association Rules algorithm, with modified Apriori algorithm select as a proposed technique to produce the support and confidence of the different levels in Web log mining of server. The selection of the modified Apriori algorithm for performing Web usage mining on particular log record in severs is because of modified Apriori algorithm is a belonging from Apriori data mining technique for association based analysis. By applying this modifying algorithm to the Web log file or record, the relationship between the accessed pages can be mined. The Web log patterns efficiency in terms of time, CUP utilization and Memory Utilization of developed application also can analyze by using this algorithm where the descriptive statistic approach can’t perform this type of analysis. The results and findings patterns for this analysis are more reliable more accuracy as compare standard Apriori algorithm properties. The results or findings from this experimental analysis are surely useful for Web administrator in order to improve Web services, performance and effectiveness through the improvement of Web sites, including their structure, presentation and contents delivery. The valuable actions may contain of performing the Web pages value added modification.

REFERENCES

[1] 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

[2] 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 |

[3] 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

[4] 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

[5] 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

[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] 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

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

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

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

[11] 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.

[12] G. Stumme, A. Hotho, 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.

[13] 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.

[14] Ajit Abhraham, Vitorino Ramos, Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming, to appear in CEC´03 - Congress on Evolutionary Computation, IEEE Press, Canberra, Australia, 8-12 Dec. 2003.