Volume 4, Issue 2, February 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Implementation on Web Log Mining

Bhaiyalal Birla*, Sachin Patel
IT & PCST, Indore
India

Abstract* — Web *is a treasure of information and data, where large amount of data is available in different formats and structures. Finding the useful data from the web is a complex task, therefore the data mining algorithm is working to identify the pattern and information from the data. The presented paper is evaluation and implementation of various techniques which are available for web data analysis. Therefore this paper includes pattern mining algorithms for evaluation and the implementation of frequent pattern analysis from the web data using Apriori Algorithm and for improve the performance we proposed a new modified Apriori algorithm and implement it. And finally we compare the Apriori and new modified Apriori results in this paper.

***Keywords*—web usage mining, frequent pattern mining, Apriori, personalization and web data.**

I. INTRODUCTION

Internet is a large source of data and information; the data on web is frequently accessed and changed. Important and knowledgeable information extraction form the World Wide Web is the application of data mining techniques. According to objective and purpose, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

- **Web usage mining:** This domain allows for the collected works of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server.
- **Web content mining:** This technique is also known as text mining, generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the significance of the content.

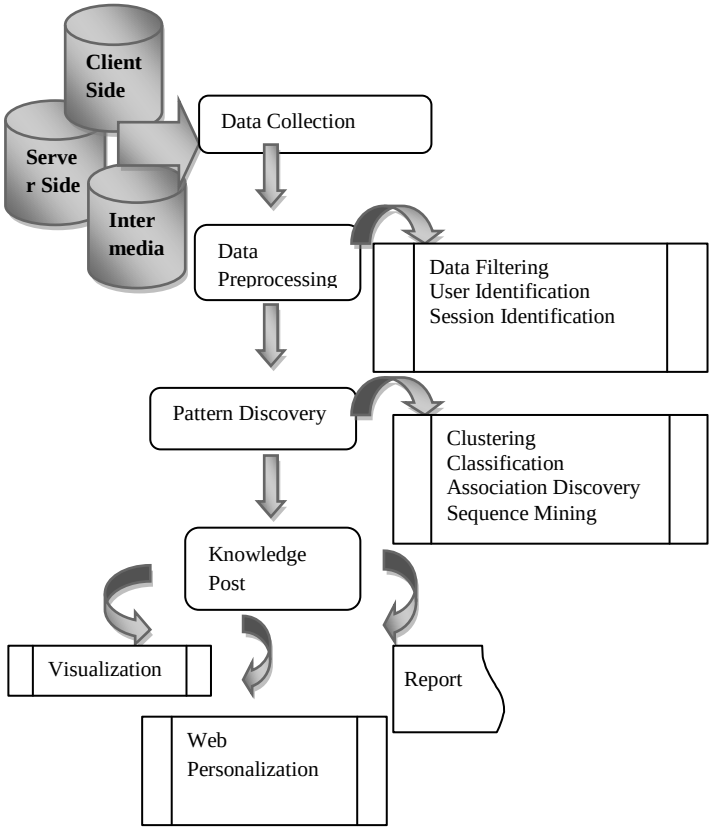

Fig. 1 Shows the Web Usage Mining Process

- **Web structure mining**: That is one of three categories of web mining, it is a tool used to recognize the connection between web pages linked by information or direct link connection. This organization of data is discoverable by the condition of web structure schema through database techniques for Web pages. This relationship allows a search engine to pull data concerning to a search query directly to the connecting Web page from the Web site the content rests upon.

In this working domain we concentrate over the web uses mining and information pattern detection. Figure 1 shows the basic process of uses mining process and the applications. The data is collected from the web servers, these web server manage the web access log for all the websites that are managed under the web server. This data is cleaned and managed in a specific format and may be used for different kind of information extraction using different schemes.

II. BACKGROUND STUDY

This section includes the previous methods that are consumes the frequent item set mining algorithms for web log pattern extraction. In order to find information from the web access log files ***L.K. Joshila Grace et al*** provides a study according to his article, Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analysing these log files gives a neat idea about the user. This paper gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn give way to an effective mining. It also provides the idea of creating an extended log file and learning the user behaviour.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. ***B.Santhosh Kumar et al*** presents a research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of memory usage and time usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm.

Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. In this paper ***S. Veeramalai et al*** analyse the pattern using different algorithms like Apriori, Hash tree and Fuzzy and then we used enhanced Apriori algorithm to give the solution for Crisp Boundry problem with higher optimized efficiency while comparing to other algorithms.

Web mining is used to discover relevant, useful and hidden information from the Web data. In the web usage mining, a category of web mining, we focused on knowledge discovery from the usage data of individual web site. It is an emerging field of research due to increase attraction of users towards World Wide Web. By understanding the behaviour of the user navigation pattern on a website, website personalization can be done thereby increasing the benefit of business, advertisement, and many more purposes. ***Amit Kumar Mishra et al*** attempts to personalize the website by the use of Self Organized Map and clustering technique user navigation behaviour based on past history.

III. WEB PERSONALIZATION

Personalization includes using technology to accommodate the differences between individuals. Once restricted mainly to the Web, it is becoming a factor in education, health care (i.e. personalized medicine), television, and in both "business to business" and "business to consumer" settings. Social Network websites use personal data to provide relevant advertisements for their users. Websites like Google and Facebook are using account information to give better services.

There are three categories of personalization:

- Profile / Group based
- Behavior based (also known as Wisdom of the Crowds)
- Collaboration based

Web personalization models include rules-based filtering, based on "if this, then that" rules processing, and collaborative filtering, which serves relevant material to customers by combining their own personal preferences with the preferences of like-minded others. Collaborative filtering works well for books, music, video, etc. However, it does not work well for a number of categories such as apparel, jewelry, cosmetics, etc. Recently, another method, "Prediction Based on Benefit", has been proposed for products with complex attributes such as apparel. [6]

In order to personalize a web site, the system should be able to distinguish between different users or groups of users. This process is called user profiling. In case no other information than what is recorded in the web logs is available, this process results in the creation of aggregate, anonymous user profiles since it is not feasible to distinguish among individual visitors. However, if the user's registration is required by the web site, the information residing on the web log data can be combined with the users' demographic data, as well as with their individual ratings or purchases. The final stage of log data pre-processing is the partition of the web log into distinct user and server sessions. A user session is defined as "a delimited set of user clicks across one or more web servers", whereas a server session, also called a visit, is defined as "a collection of user clicks to a single web server during a user session". If no other means of session identification, such as cookies or session ids is used, session identification is performed using time heuristics, such as

setting a minimum timeout and assumes that consecutive accesses within it belong to the same session, or a maximum timeout, assuming that two consecutive accesses that exceed it belong to different sessions. More details on the user and session identification process can be found in [8].

This section provides the overview of the previous works that supports and provides the guidelines to utilize different models over the web access logs.

IV. ALGORITHMS USED

This section provides the algorithms used that are required implementation work.

A. Apriori Algorithm

During the literature survey we found various frequent set mining algorithms are implemented for web personalization using proxy web access log mining. Additionally various other sequential and frequent pattern mining approaches are developed. But they not provide the sufficient performance for large data set.

The Apriori Algorithm is an influential algorithm for mining frequent item-sets for Boolean association rules.

- *Frequent Item-sets:* The sets of item which has minimum support (denoted by L_i for i^{th}-Item-set).
- *Apriori Property:* Any subset of frequent item-set must be frequent.
- *Join Operation:* To find L_k, a set of candidate k-item-sets is generated by joining L_{k-1} with itself.

Find the frequent item-sets: the sets of items that have minimum support– A subset of a frequent item-set must also be a frequent item-set

– i.e., if {AB} is a frequent item-set, both {A} and {B} should be a frequent item-set -Iteratively find frequent item-sets with cardinality from 1 to k (k-item-set)

- Use the frequent item-sets to generate association rules.
- Join Step: C_k is generated by joining $L_{k-1}$with itself.
- Prune Step: Any (k-1)-item-set that is not frequent cannot be a subset of a frequent k-item-set

Table 1 Shows the Apriori Algorithm Pseudo-Code

Variables: C_k: Candidate item-set of size k L_k: frequent item-set of size k L_1= {frequent items};

Process:
For (k = 1; L_k!=∅; k++) do begin
C_{k+1}= candidates generated from L_k;
For each transaction t in database do
Increment the count of all candidates in C_{k+1}
Those are contained in t
L_{k+1}= candidates in C_{k+1} with min_support
End
Return $\cup_k L_k$;

B. Modified Apriori Algorithm

The traditional Apriori algorithm is most frequently used by different researchers and groups to mine log data. This algorithm has some problem with their performance we observe that when the item set are increased then the time and memory required is increased exponent manner. To overcome this problem we propose a new Modified Apriori algorithm.

Table 2 Shows the Modified Apriori Algorithm Pseudo-Code

Variables: C_k: Candidate item-set of size k L_k: frequent item-set of size k L_1= {frequent items};

Process:
For (k = 1; L_k!=∅; k++) do begin
C_{k+1}= candidates generated from L_k;
For each transaction t in database do
If (t== input set) then
{
Increment the count of all candidates in C_{k+1}
}
Those are contained in t
L_{k+1}= candidates in C_{k+1} with min_support
End
Return $\cup_k L_k$;

Accuracy

The performance of algorithms are evaluated using N cross Validation method, based on this method accuracy is calculated using the total number of correctly classified objects verses the total sample produced to classify. The mathematical expression can be written as for calculating the performance in terms of accuracy as:

$$Accuracy = \frac{total\ no\ of\ correctly\ classified\ instaces * 100}{total\ No\ of\ instances}$$

Table 3 Accuracy of Apriori and Modified Apriori

Data set size(instances)	Apriori	Modified Apriori
508	74.21	78.35
1754	67.37	72.62
3890	69.29	73.23
7291	68.12	72.13
10289	73.41	78.41

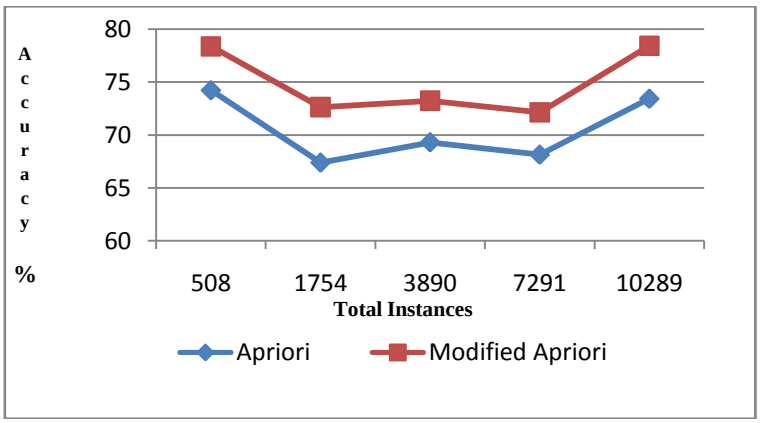

Fig. 2 Accuracy graph of frequent item set mining

Memory Usage

The total memory resources required to execute the given algorithm is known as Memory usage of the system. Which is defined as the peak memory required executing the system or algorithm is known as memory uses. By using the below given results that is observed that the memory consumption of the data models are increases as size of data is increases.

Table 4 Shows the Memory Usage of Apriori and Modified Apriori

Data set size (instances)	Apriori	Modified Apriori
508	32783	29380
1754	33712	30732
3890	34817	32783
7291	37892	34181
10289	36721	37381

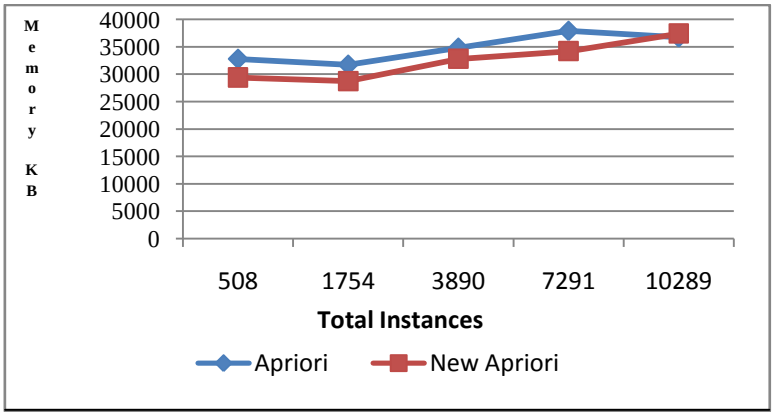

Fig. 3 Memory Usage Graph of Apriori and Modified Apriori

Build time

Total time required to develop data model using the input data is known as model build time. That is estimated using the elapse time between initialization of algorithm and finishing the model building. Table and graphs given below shows the model build time in terms of milliseconds. According to the evaluated results the model build time is increases as the size of data set is increases in all cases.

Table 5 shows the model build time of the system

Data set size (instances)	Apriori	Modified Apriori
508	25.27	17.26
1754	27.08	19.87
3890	68.22	24.67
7291	109.2	52.27
10289	298.87	166.12

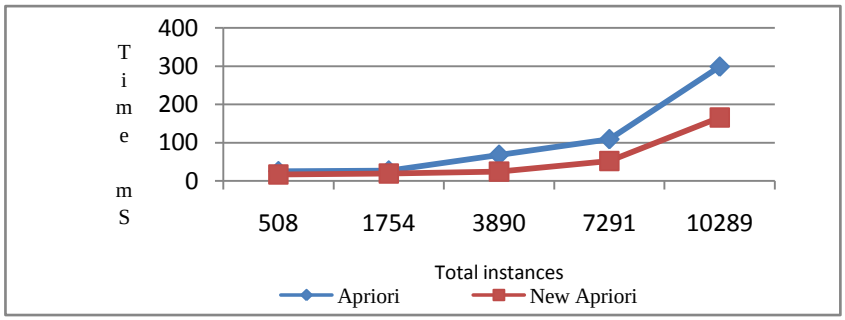

Fig. 4 Models Build Time of Apriori and Modified Apriori Algorithm

Search time

Search time or prediction time is defined as the time required predicting a value after accepting some parameters. Actually this time is too small thus here provided results are combined results for predicting the complete data set size.

Table 6 Decision Time of Apriori and Modified Apriori Algorithm

Data set size (instances)	Apriori	Modified Apriori
508	21.92	20.41
1754	32.13	29.26
3890	43.23	39.83
7291	53.55	47.32
10289	72.67	63.43

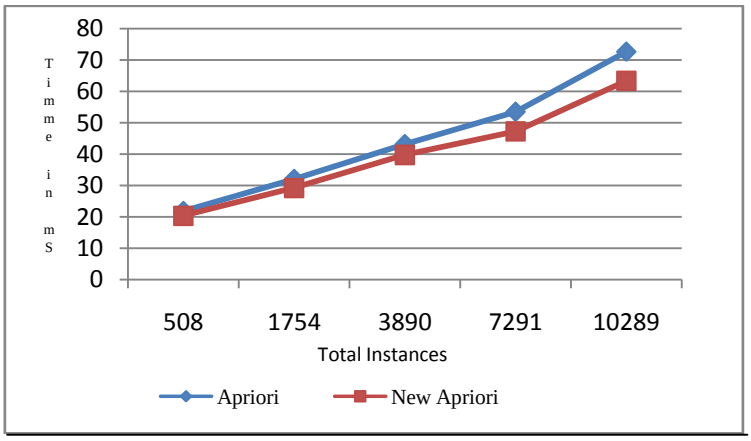

Fig. 5 Decision Time Graph of Apriori and Modified Apriori Algorithm

V. CONCLUSION

The proposed work is indented to find the effect of various data mining techniques over the web log analysis. That may lead to design and development of Apriori and Modify Apriori algorithm. After implementation of the Apriori algorithm that is observed that the performance of find frequent pattern is less effective than similar algorithms, therefore some improvements are proposed and implemented on Apriori algorithm, So Finally we implement modify Apriori

algorithm. And we observe some interesting facts are required to discuss that may helpful for other research objectives. The found facts are as:

1. Found that when the size of data increases the developed model perform poor results and some are provide more accurate results
2. In Frequent set mining our modify Apriori algorithm takes less time than Apriori algorithm for search patterns and building data models.
3. Memory used is directly proportional to the size of data in both kinds of algorithms used in web mining.
4. Accuracy is not much depends on the size of data it is most of the time depends upon type of data.

VI. FUTURE WORK

After implementation of web mining system that is observed that the targeted goals are achieved and successfully omit results as expected from the selected models. In future work the same algorithms are utilized to work with other application for utilizing the properties of both kinds of algorithm. During study some modifications are also proposed in Apriori algorithm that is much efficient and effective with respect to the traditional Apriori algorithm, in future work that is required to enhance more for reducing the search time and building time in the proposed algorithm.

ACKNOWLEDGEMENT

I would like to thank Mr. Sachin Patel (Professor, Patel College of Science and Technology, Indore) for their mentorship.

REFERENCES

[1] Simon Fong, Yan Zhuang, Jiaying He, Not Every Friend on a Social Network Can be Trusted: Classifying Imposters Using Decision Trees, 58 978-1-4673-5861-3/12/$31.00 ©2012 IEEE

[2] L.K. Joshila Grace, V. Maheswari, Dhinaharan Nagamalai, ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011

[3] B. Santhosh Kumar, K.V. Rukmani, Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms, Int. J. of Advanced Networking and Applications, 400, Volume:01, Issue:06, Pages: 400-404 (2010)

[4] S. Veeramalai, N. Jaisankar and A. Kannan, Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy, International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[5] Amit Kumar Mishra, Mahendra Kumar Mishra, Vivek Chaturvedi, Santosh Kumar Gupta, Jaiveer Singh, Web Usage Mining Using Self Organized Map, Volume 3, Issue 6, June 2013 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[6] Personalization: Collaborative Filtering vs Prediction Based on Benefit Theory, November 05, 2007,http://myshoppal.typepad.com/blog/2007/11/personalization.html

[7] Applying Web Usage Mining to a University Website Access Domain, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868, Foundation of Computer Science FCS, New York, USA

[8] New Approaches to Web Personalization, Ph.D. Thesis, Magdalini P. Eirinak, Athens University of Economics and Business Dept. of Informatics May 2006

[9] Study of Pre-processing Methods in Web Server Logs, Dr.Sanjeev Dhawan, Mamta Lathwal, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May2013

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

