Web Log Miner

Laxmi Devi[1] Mr. Manoj Yadav[2]
[1]M.Tech Scholor [2]Assistant Professor
[1]Al-Falah School of Engineering and Technology Dhauj, Faridabad, Hariyana

Abstract— The efficiency of mining association rules is an important field of Knowledge Discovery in Databases. The Apriori algorithm is a classical algorithm in mining association rules. In this Paper, I have proposed a web browser named "Web Log Miner" that supports mining feature Which will mine Web usage using Apriori algorithm and presents a list of web pages based on user`s interest and cache.. This paper present that web usage can be mine by two ways time based and frequency based. In this study I have used Apriori algorithm which reduce the redundant generation of sub-item sets during pruning the candidate item sets, and can form directly the set of frequent item sets and eliminate candidates having a subset that is not frequent in the meantime. This paper explores the use of Apriori Algorithm in Web Usage Mining to analyze Web log records collected from E-Learning portal.

Key words: Association rules, Knowledge Discovery, Web Log Miner, Apriori Algorithm

I. 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 data, especially when a human could not analyze such datasets. Massification 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. 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 Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data and is targeted towards applications. Web Usage Mining mines the secondary 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. This study will explore the use of Web Usage Mining techniques to analyze Web log records collected from E-Learning portal. This includes descriptive statistic and Association Rules for the portal including support and confidence to represent the Web usage.

In this Paper, I have proposed a web browser named "Weblog Miner" that supports mining feature Which will mine Web usage using Apriori algorithm and presents a list of web pages based on user`s interest and cache.

In field of Data Mining, Association Rule Learning is a popular and well researched Method for discovering Interesting relations between variables in large databases. It is a study of analyzing and presenting strong rules discovered in databases using different measures of interestingness.

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule {onion, potatoes}=>{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy burger. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.

In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics. But this study will explore the use of Web Usage Mining techniques to analyze Web log records collected from E-Learning portal. This includes descriptive statistic and Association Rules for the portal including support and confidence to represent the Web usage.

II. STUDY OF EXISTING SYSTEM

A. APRIORI ALGORITHM

In computer science and data mining, Apriori is a classic algorithm for learning association rules.

Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

Following the original definition by Agrawal the problem of association rule mining is defined as:

Let I = {i1, i2, ..., in} be a set of n binary attributes called items. Let D = {t1, t2, ...,tn} be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $\cap = \emptyset$. The sets of items (for short itemsets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is I = {milk,bread,butter,beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table below. An example rule for the supermarket could be {milk,bread}=>{butter}

All rights reserved by www.ijsrd.com

meaning that if milk and bread is bought, customers also buy butter.

Note: this example is extremely small. In practical applications, a rule needs a support of several hundred transactions before it can be considered statistically significant, and datasets often contain thousands or millions of transactions

Transaction ID	Milk	Bread	Butter
1	1	1	0
2	0	1	0
3	0	0	1
4	1	1	0
5	0	1	0
6	0	0	1
7	1	1	1
8	0	1	1
9	1	1	0
10	1	1	0
11	0	1	1
12	1	0	0
13	1	1	0
14	1	0	1
15	1	1	0

Table 2.1: Transactions

B. ASSOCIATION RULE GENERATION

Association rule generation is usually split up into two separate steps:

(1) First, minimum support is applied to find all frequent item sets in a database.
(2) Second, these frequent item sets and the minimum confidence constraint are used to form rules.

While the second step is straight forward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over I and has size 2n- − 1 (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent item sets.

C. 2APRIORI ALGORITH PSEUDOCODE

```
Procedure Apriori (T, minSupport) { //T is the database and minSupport is the minimum support
L1= {frequent items};
for (k= 2; Lk-1 !=pie ; k++) {
Ck= candidates generated from Lk-1
//that iscartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not
//frequent
for each transaction t in database do{
 #increment the count of all candidates in Ck that are contained in t
Lk= candidates in Ck with minSupport
 }//end for each
 }//end for
returnUK LK;
}
```

As is common in association rule mining, given a set of itemsets (for instance, sets of retail

Transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

III. PROPOSED APPROACH

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straightforward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible itemsets (item combinations). The set of possible itemsets is the power set over I and has size 2^n-1 (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicitywhich guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent and Exploiting this property, efficient algorithms (e.g., Aprioriand Eclatcan find all frequent itemsets).

IV. DESIGN AND IMPLEMENTATION

This study is aim to develop a web browser named Weblog Miner that supports mining feature

WebLog Miner comprises of 2 separate words: Web log and Miner

WebLog: It is a file automatically created & maintained to list the actions performed or requested by the user and Miner is a tool that extracts the valuable information from the bulk.

A. ANALYZER

Web pages can be analyzed through below mentioned parameters:-

(1) Time based analysis

All rights reserved by www.ijsrd.com

(2) Frequency based analysis

1) TIME BASED ANALYSIS

It analyze web log with respect to access time of each web page & display result of top 5 web pages. It shows information about How long the particular web page was accessed in the last 7 days. User can see the details about highly accessed web sites and their accessed time in table also.

2) FREQUENCY BASED ANALYSIS

It analyze web log with respect to no. of access of each web page & display result of top 5 web pages. It shows information about How many times the particular web page was accessed in the last 7 days.

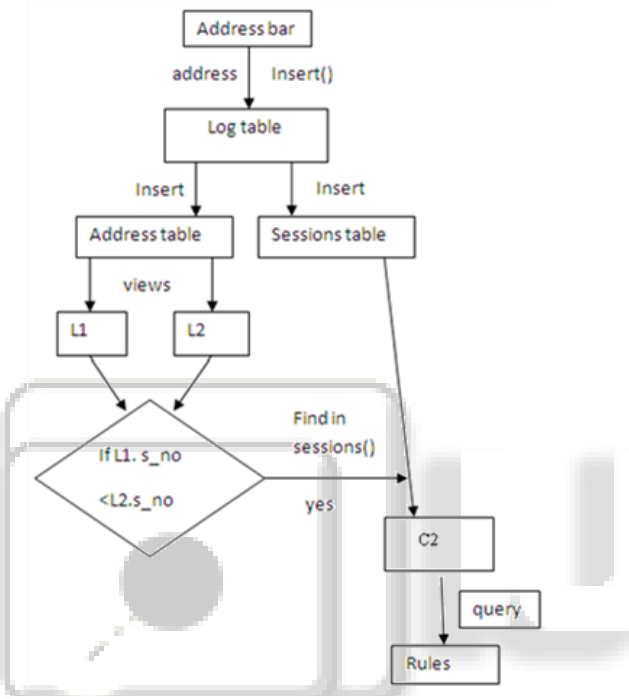

Fig 4.1: Flowchart showing methodology to develop proposed algorithm.

B. PROPOSED SOLUTION USING APRIORI ALGORITHM:

Apriori algorithm has sound theory base, but the focus is its efficient issue. Agrawal raised several improving methods in person, like Apriori Tid, Apriori All, etc. After that many optimized methods were raised based on the framework of Aprioir algorithm. This study introduced an algorithm that decrease the number of candidate items in the candidate item set Ck.

In the Apriori algorithm, Ck-1 is compared with support level once it was found. Item sets less than the support level will be pruned and Lk-1 will come out which will connect with itself and lead to Ck. The optimized algorithm is that, before the candidate item sets Ck come out, further prune Lk-1, count the times of all items occurred in Lk-1, delete item sets with this number less than k-1 in Lk-1. In this way, the number of connecting items sets will decrease, so that the number of candidate items will decline.

C. The Realization of Algorithm

According to the properties of frequent item sets, this algorithm declines the number of candidate item sets further. In other words, prune Lk-1before Ck occur using Lk-1. This algorithm can also be described as following:

Count the number of the times of items occur in Lk-1(this process can be done while scan data D); Delete item sets with this number less than k-1 in Lk-1to get Lk-1. To distinguish, this process is called Prune 1 in this study, which is the prune before candidate item sets occur; the process in Apriori algorithm is called Prune 2, which is the prune after candidate item sets occur to find out the k candidate item sets.

V. CONCLUSION

In this thesis, we have considered the Apriori Algorithm in mining Web usage .

In this paper proposed methodologies used for classifying the user using Web Usage data. This model analysis the users behaviors and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. The results of evaluations shows that using more efficient algorithms for finding similar .users lead to recommender system that provides more interesting recommendations for website users. Proposed solution can be extended by considering the effect of users‟ feedback for increasing the quality of recommendation.

This can be done, eventually, by introducing new parameters for the characterization of the Web Usage data.

REFERENCES

[1] Agrawal, R., Imielinski, T. and Swami, A. (1993).Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the International ACM SIGMOD Conference, Washington DC, USA, pages 207–216

[2] Agrawal, R. and Srikant, R. (1994).Fast Algorithm for Mining Association Rules.Proc. of the 20th VLDB Conference. Pp 487-499.

[3] Agrawal, R., and Srikant, R. (1995).Mining Sequential Patterns. In Proc. of the Eleventh International

[4] Magdalini Eirinaki, Michalis Vazirgiannis, and Iraklis Varlamis, "Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process," in Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'03), Washington DC, 2003.

[5] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, and Ali Mamat, "WebPUM: A Web-based recommendation system to predict user future movements," Expert Systems with Applications, vol. 37, pp. 6201-6212, 2010.

[6] Magdalini Eirinaki, Dimitrios Mavroeidis, George Tsatsaronis, and Michalis Vazirgiannis, "Introducing Semantics in Web Personalization: The Role of Ontologies," in Proc. EWMF/KDO'2005, 2005, pp. 147-162.

[7] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Computer

All rights reserved by www.ijsrd.com

Networks and ISDN Systems, vol. 28, no. (7–11), pp. 1007–1014, 1996.

[8] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, vol. 43, no. 8, pp. 142–151, 2000.

[9] F. Masseglia, P. Poncelet, and R. Cicchetti, "WebTool: An Integrated Framework for Data Mining," in Proceedings of the 9th International Conference on Database and Expert Systems Applications (DEXA'99), Florence, Italy, 1999, pp. 892-901.

[10] Ranieri Baraglia and Fabrizio Silvestri, "An online recommender system for large Web sites," in Proceedings of the IEEE/WIC/ACM international conference on Web, Beijing, China, 2004.

[11] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs," in Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999.

[12] Dimitrios Pierrakos, Georgios Paliouras, Christos Papatheodorou, and Constantine D. Spyropoulos, "KOINOTITES:A Web Usage Mining Tool for Personalization," in Proceedings of the Panhellenic Conference on Human Computer Interaction, 2001.

[13] B. Zhou, S. C. Hui, and K. Chang, "An intelligent recommender system using sequential Web access patterns," in IEEE conference on cybernetics and intelligent systems, 2004, pp. 393–398. International Journal of Web & Semantic Technology (IJWesT)

All rights reserved by *www.ijsrd.com*