FREE USER BEHAVIOR INFORMATION FROM CENTRAL DATABASE USING WEB USAGE MINING

Kamlesh S. Jetha
Assistant Professor, Department of I.T.
Anantrao Pawar College of Engineering & Research, Pune
Pune, India

Abstract: Web mining is the recent variation of data mining technique which is concerned to web data that results through various web activities. This application of data mining technique is used to discover patterns from the Web. Web usage mining is an important technology for understanding user's behaviors on the web which is the process of extracting useful information from server logs. User leaves some valuable information in web logs when user browses web pages. The knowledge from the data collected in log file is automatically discovered through web usage mining. The proposed concept in the paper is an attempt to apply an efficient web mining algorithm for web log analysis. E-commerce web portals demands security from search engines to identify context of the problem. The results received through web mining analysis get applied to the class of problems. Through Improved APrioriAll, candidate sets are found to be much smaller in stage wise comparison. The scanning of database block gets extremely reduced in E-web miner. E-web miner has much better performance with time and space parameters as compared with other algorithms.

IndexTerms: **Web mining, Candidate sets, Improved APrioriAll, E-Web miner, Web log**

I. INTRODUCTION

Web Mining is the application of data mining techniques which discover patterns from the Web. _Web usage mining is one of the important technologies for understanding user's behaviors on the Internet. It is the process of extracting useful information from server logs what users are looking for on Internet. The another aspect of the use of web usage mining is, while browsing web pages user leaves some valuable information in web logs. The log file automatically generates information about the traversed data. Such structured data are generated dynamically giving a sense of continuity over a period of time, as it has been collected from web pages. Web mining allows users to search for patterns in data through Web structure mining (WSM), Web content mining (WCM) and usage mining (WUM).

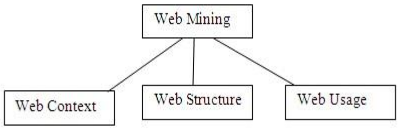


Fig. 1: Web Mining Taxanomy

The process of examining data related to the node and connection structure of particular web site that may be linear or hierarchical is web structure mining. Web content mining is used to examine the contents of data collected by search engines and web spiders. The actual linkage access information of web pages such as number of hits, subscription of pages or visits to discover interesting usage patterns from Web data. Web usage mining captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining process involves three types of tasks: Data Pre-processing, Pattern discovery of searched data and finally pattern analysis.

II. WEBLOG ANALYSIS:

Any kind of user traversed information is recorded by web server into web log file for corresponding access data. Web log mining is outcome of web usage mining. The log files provide data about typical behaviour of the users while browsing web pages, the operating system used on host, error handling methods and time required for particular web usage successful/unsuccessful transaction. The web logs are categorized in two forms. The first one is server side web log, that reveals information about availability as well as vulnerability of web servers, security loop holes of servers and user friendliness of the website. The client can improve their web browsing performance if they will get directed about frequency of usage of particular web page for data prefetching and caching.

III. SEQUENTIAL PATTERN ANALYSIS

The set of web pages are revealed by traversal pattern visited by a user in a session, which is used for data prefetching and caching purpose. The access pattern of the user can never be monitored in time synchronous manner because data generated in web log is asynchronous. A special case of structured data mining is sequential pattern mining, which finds statistically relevant patterns between examples where the values are delivered in a sequence. Sequential pattern mining is applied on information to predict customer behaviours. A user can invoke many sessions. A sequential pattern over total number of clients may emerge in many sessions. Support is also known as coverage considered as percentage of clients that are creating sequential patterns.

Discovering Rule has two associated measures as given below-

Example: If a basket contains apples and cheese, then it also contains beer.

1. Confidence- When the 'if' part is true, how often is the possibility to get 'then' true?
2. Support- how much of the database contains the 'if' part?

IV. APRIORI ALGORITHM

Apriori is a classic algorithm to extract frequent mining data set and association rule learning over transactional databases. It uses bottom-up approach, which attempts to find subsets which are common to at least support of the item set. Candidate generation method is used to extend one item at a time from frequent item sets and similarly groups of candidates are tested through data. When successful extensions are not found against data, then Apriori algorithm terminates its execution.

Apriori algorithm: Program flow chart can also be described with the help of logical flowchart as described below- .

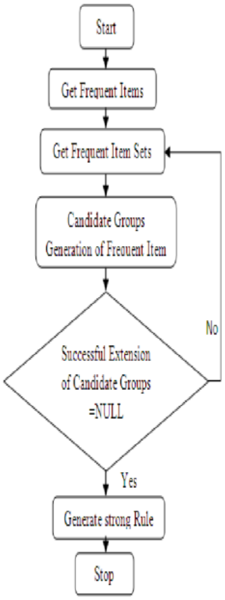


Figure 2. Apriori Algorithm

V. APRIORIALL ALGORITHM

There are two families of algorithms *count-all* and *count-some*. The count-all algorithms count all the large sequences, including nonmaximal sequences. The non-maximal sequences must then be pruned out (in the maximal phase). AprioriAll listed below is a count-all algorithm, based on the Apriori algorithm for finding large itemsets. Apriori- Some is a count-some algorithm. The intuition behind these algorithms is that since we are only interested in maximal sequences, we can avoid counting sequences which are contained in a longer sequence if we first count longer sequences. However, we have to be careful not to count a lot of longer sequences that do not have minimum support. Otherwise, the time saved by not counting sequences contained in a longer sequence may be less than the time wasted counting sequences without minimum support that would never have been counted

APrioriAll Algorithm is explained below:

Step1- L_1 = large 1-sequences; // Result of itemset phase

Step 2- for (k = 2; L_{k-1} 0; k++) do

Step 3- begin

C_k =New Candidates generated from L_{k-1} (see below) for each customer-sequence *c* in the database **do** Increment the count of all candidates in C_k that are contained in *c*.

L_k = Candidates in C_k with minimum support.

Step 4- end

Answer = Maximal Sequences in $_k$ L_k ;

Apriori Candidate Generation Algorithm is explained below:

The apriori-generate function takes as argument L_{k-1}, the set of all large (k-1)-sequences. It works as follows. First join L_{k-1} with L_{k-1}

Step 1- insert into C_k

Step 2 -select p.litemset$_1$, ..., p.litemset$_{k-1}$, q.litemset$_{k-1}$

Step 3 -from L_{k-1} p, L_{k-1} q

Step 4 - where p.litemset$_1$ = q.litemset$_1$, . . ., p.litemset$_{k-2}$ = q.litemset$_{k-2}$;

Step 5 - Next delete all sequences c C_k such that some (k-1)-subsequence of c is not in L_{k-1}

VI. E-WEBMINER ALGORITHM

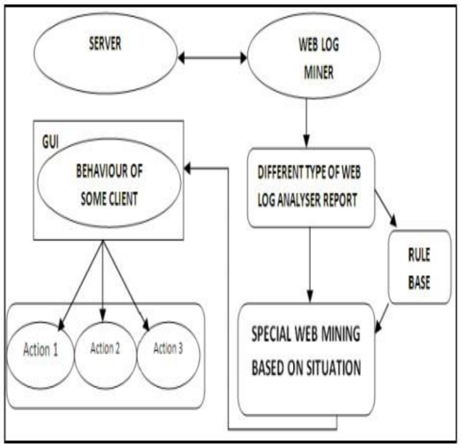


Figure 3. Architecture of E-web Miner

The E-Web miner carries support and confidence of sequential pattern of web pages and candidate set pruning to reduce the repetitive scanning of database containing the web usage information and thus reducing the time.

Algorithm of E-Web Miner:

Step 1- Arrange the web page set of different users in increasing order,
Step 2- Store all web page sets of user in string array A.
Step 3-frequency =0, max=0;

Step 4- for i=1 to n

```
for j=0 to (n-1)
if substring (A[i], A[j])  frequency=frequency+;
end if;
B[i] =Frequency;
end for
if max <= frequency max=frequency;
end if
end for
```

Step 5- Find all position in Array B where value is equal to Max and select the corresponding Substring from A.

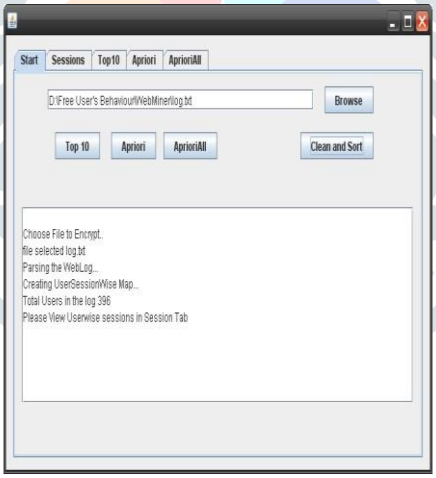
Step 6- Produce output of all substrings with their position which is the desired output.

The algorithm produces the partial result with a limited set of information which turns out to be a great demerit of the algorithm. The algorithm only list the item sets and the IP addresses from where these item sets were accessed. This partial information may be helpful in some cases but not always; this raises an expectation for better and efficient algorithm.

VII. EXPERIMENTAL RESULTS

When these algorithms are applies on web server, it provides very effective result. In below figure it shows the experimental result in web log form. The comparison of result with Apriori and APrioriAll is clearly represented through this analysis.

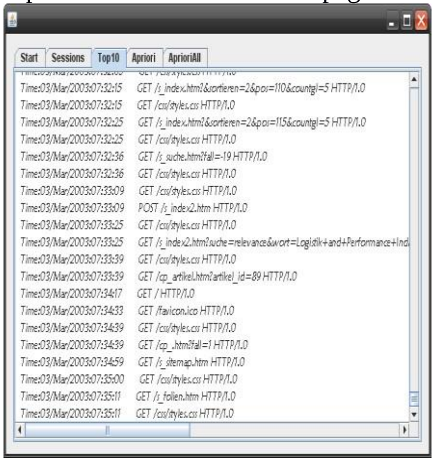
The sessions have been selected from browser as per user's choice and then it asks to select web log file which gets encrypted.



```
Choose File to Encrypt..
file selected log.txt
Parsing the WebLog...
Creating UserSessionWise Map...
Total Users in the log 396
Please View Userwise sessions in Session Tab
```

Figure 4. Web log file encryption

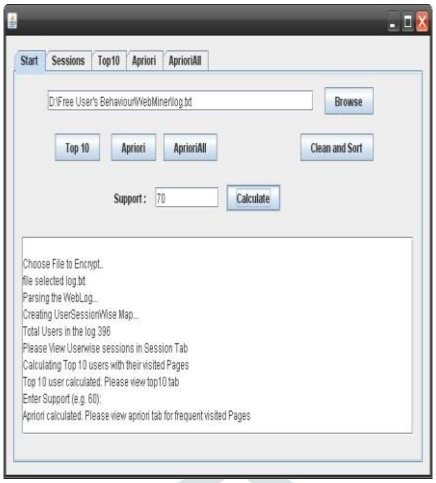
By click event on start button, it calculates top10 users with their visited pages.



```
Time:03/Mar/2003:07:32:15    GET /s_index.htm?&sortieren=2&pos=110&countgl=5 HTTP/1.0
Time:03/Mar/2003:07:32:15    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:32:25    GET /s_index.htm?&sortieren=2&pos=115&countgl=5 HTTP/1.0
Time:03/Mar/2003:07:32:25    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:32:36    GET /s_suche.htm?fall=-19 HTTP/1.0
Time:03/Mar/2003:07:32:36    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:33:09    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:33:09    POST /s_index2.htm HTTP/1.0
Time:03/Mar/2003:07:33:25    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:33:25    GET /s_index2.htm?suche=relevance&wort=Logistik+and+Performance+Ind
Time:03/Mar/2003:07:33:39    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:33:39    GET /cp_artikel.htm?artikel_id=89 HTTP/1.0
Time:03/Mar/2003:07:34:17    GET / HTTP/1.0
Time:03/Mar/2003:07:34:33    GET /favicon.ico HTTP/1.0
Time:03/Mar/2003:07:34:39    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:34:39    GET /cp_.htm?fall=1 HTTP/1.0
Time:03/Mar/2003:07:34:59    GET /s_sitemap.htm HTTP/1.0
Time:03/Mar/2003:07:35:00    GET /css/styles.css HTTP/1.0
Time:03/Mar/2003:07:35:11    GET /s_folien.htm HTTP/1.0
Time:03/Mar/2003:07:35:11    GET /css/styles.css HTTP/1.0
```

Figure 5. Top10 users visited web pages

Then we can generate Apriori result and APrioriAll result with use of support and confidence as depicted in figure 6 and 7.

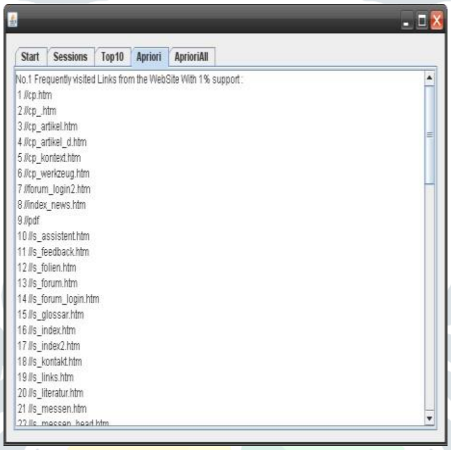


D:\Free User's Behaviour\WebMiner\log.txt

Support : 70

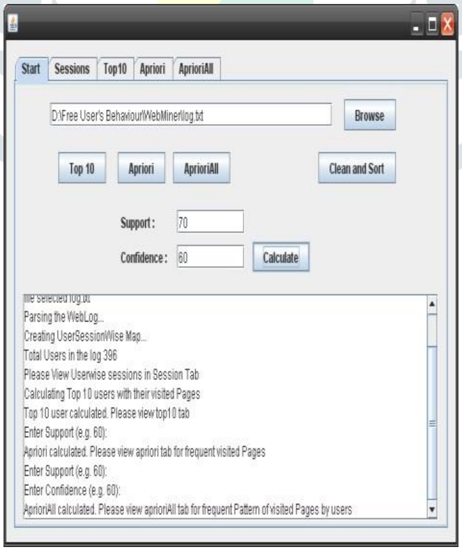
```
Choose File to Encrypt..
file selected log.txt
Parsing the WebLog...
Creating UserSessionWise Map...
Total Users in the log 396
Please View Userwise sessions in Session Tab
Calculating Top 10 users with their visited Pages
Top 10 user calculated. Please view top10 tab
Enter Support (e.g. 60):
Apriori calculated. Please view apriori tab for frequent visited Pages
```

Figure 6 a) Apriori Result Analysis with support



```
No.1 Frequently visited Links from the WebSite With 1% support :
1 //cp.htm
2 //cp_.htm
3 //cp_artikel.htm
4 //cp_artikel_d.htm
5 //cp_kontext.htm
6 //cp_werkzeug.htm
7 //forum_login2.htm
8 //index_news.htm
9 //pdf
10 //s_assistent.htm
11 //s_feedback.htm
12 //s_folien.htm
13 //s_forum.htm
14 //s_forum_login.htm
15 //s_glossar.htm
16 //s_index.htm
17 //s_index2.htm
18 //s_kontakt.htm
19 //s_links.htm
20 //s_literatur.htm
21 //s_messen.htm
22 //s_messen_head.htm
```

Figure 6 b) Apriori Result



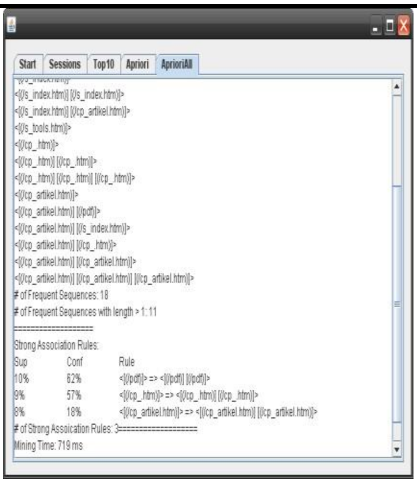
D:\Free User's Behaviour\WebMiner\log.txt

Support : 70

Confidence : 60

```
file selected log.txt
Parsing the WebLog...
Creating UserSessionWise Map...
Total Users in the log 396
Please View Userwise sessions in Session Tab
Calculating Top 10 users with their visited Pages
Top 10 user calculated. Please view top10 tab
Enter Support (e.g. 60):
Apriori calculated. Please view apriori tab for frequent visited Pages
Enter Support (e.g. 60):
Enter Confidence (e.g. 60):
AprioriAll calculated. Please view aprioriAll tab for frequent Pattern of visited Pages by users
```

Figure 7 a) APrioriAll Result Analysis with support and confidence



```
<[(/s_index.htm)]>
<[(/s_index.htm)] [(/s_index.htm)]>
<[(/s_index.htm)] [(/cp_artikel.htm)]>
<[(/s_tools.htm)]>
<[(/cp_.htm)]>
<[(/cp_.htm)] [(/cp_.htm)]>
<[(/cp_.htm)] [(/cp_.htm)] [(/cp_.htm)]>
<[(/cp_artikel.htm)]>
<[(/cp_artikel.htm)] [(/pdf)]>
<[(/cp_artikel.htm)] [(/s_index.htm)]>
<[(/cp_artikel.htm)] [(/cp_.htm)]>
<[(/cp_artikel.htm)] [(/cp_artikel.htm)]>
<[(/cp_artikel.htm)] [(/cp_artikel.htm)] [(/cp_artikel.htm)]>
# of Frequent Sequences: 18
# of Frequent Sequences with length > 1: 11
==================
Strong Association Rules:
Sup        Conf        Rule
10%        62%         <[(/pdf)]> => <[(/pdf)] [(/pdf)]>
9%         57%         <[(/cp_.htm)]> => <[(/cp_.htm)] [(/cp_.htm)]>
8%         18%         <[(/cp_artikel.htm)]> => <[(/cp_artikel.htm)] [(/cp_artikel.htm)]>
# of Strong Assoication Rules: 3==================
Mining Time: 719 ms
```

Figure 7 b) AprioriAll Result

VIII . Conclusion:

The proposed algorithm is a data mining algorithm and an improved version of existing APriori algorithm. There are many algorithms for data mining and most of them exploit the concept of Apriori algorithm, however the results they produce does not meet the expectation, motivated by the fact and also the significance of Web log analysis from the perspective of a web-designer the concept of mining algorithm for web log analysis has been proposed. Apart from the research work conducted, an evaluation tool has been developed to authenticate the work and compare the results. From the result analysis it is quite clear that proposed system is better as compared to the existing systems.

IX. References:

[1] Tong, Wang and Pi-lian, “Web log mining by Improved AprioriAll algorithm” World Academy of Science, Engineering and Technology, Vol 4 20011 pp 97-100.

[2] P Yadav, P Keservani “An Efficient Web Mining Algorithm for Web Log Analysis-E Web Miner” 978 -1-4577-0697 2012IEEE.

[3] K Vanitha, R Santhi “Using Hash Based Apriori Algorithm to Reduce the Candidate 2 Item Sets for Mining Association Rule” Journal of Global Research in Computer Science Vol 2 No 5 2011.

[4] Sachin Pardeshi, Ujjwala Patil. “Central web mining services – public and free access log files” World Journal Science and Technology April 2012, 2(3), ISSN 2231-2587.

[5] U.M. Patil and S.N. Pardeshi. "A survey on user future request prediction: Web Usage Mining" (UETAE) International Journal of Emerging Technology and Advanced Engineering, Vol. 2 (3), pp. 121-124 March 2012.

[6] Anitha, “A New Web Usage Mining Approach for Next Page Access Prediction”, International Journal of Computer Applications, Volume 8– No.11, October 2010.

[7] V. Sujatha, Punithavalli, “Improved User Navigation Pattern Prediction Technique From Web Log Data”, Proscenia Engineering 30, 2012.

[8] Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu, “A novel prediction model based on hierarchical characteristic of web site”, Expert Systems with Applications 38, 2011.

[9] A.V.Krishna Prasad and Dr.S.Ramakrishna, “Retrieving business applications using Open Web APIs Web Mining dashboard application case study” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (3),198-210,pp. 2010.

[10] M. Jalali, N. Mustapha et al,” WebPUM: A Web-based recommendation system to predict user future movements”, in international journal Expert Systems with Applications 37, 2010.

[11] Bam shad Mobster, Robert Cooley, Jaideep Srivastava “Automatic Personalization Based on Web Usage Mining” Published in: Magazine Communication of the ACM, Volume 43 Issues 8, Aug 2000.

[12] Alexandra’s Nanopolous, Dmitri’s Katsaros and Yannis Manolopolous “Effective prediction of web-user accesses: A data mining approach,” in Proc. Of the Workshop WEBKDD, 2001..