IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

An approach using Association Rule Mining Technique for frequently matched pattern of an Organization's web log data

Siddharth Jain[*1], Ruchi Dave[2], Devendra Kumar Sharma[3]
[*1,2]Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
[3]Assistant Professor, Anand International College of Engineering, Jaipur
siddhuas_pce@yahoo.com

Abstract
Web mining allows you to check for patterns in data through content mining, structure mining, and usage mining. Two dynamic areas of today's research are data mining and the WWW. A combination of the two areas sometimes referred to as Web mining. Association rule mining is a very important data mining model studied extensively by the database and data mining community. Frequent set mining was motivated by the problem of analyzing transaction data in any Organization like Educational Institute. The purpose of find out frequent patterns in Web log data is to find information about the navigational actions or performance of the users. Web Server record a log entry for every hit of access data which is done by any user. A vast number of requests are registered in a web log file. Here, in paper we concentrate to find out the problems in existing past techniques. The chance of improving the quality of log file by reducing the size of the web-log files is increases and we received some simple and quality sample log file.

Keyword – Web usage Mining, Association Rule, Web log file, Pattern.

Introduction
Data generated by members of an Organization is valuable source of knowledge about several aspects of their network activity. Data mining tools can bring many new potential for the analysis of web access log files and also the activity of the user's hits. Every time a page in your web site is requested, your web server stores a line on the access log. The access log is basically a relation with a few columns that record what regard as important information is. This data can then be mined for different purposes. Pre-processing comprise important components like integration, data cleaning, user identification and session identification to change the raw log data into a format that will be easier and efficient in mining. The main difficulty which arises is that whenever we analyze any web-log file it requires more time and space to scan the database or all access details which is available in web-log file. So, we need a technique which scans the whole log file database as fast as possible. Mining of data streams should be an incremental process with the high update rate, which means new repetitions of mining results are built based on old mining results. So the results will not have to be recalculated each time whenever user's request is received. These problems are analyzed and discovered in a new Pre-processing technique.

Literature Review
Web log files include large number of records which is usually different and huge. For pattern discovery we have to assemble the records of different web-log files. Without using content mining, structure mining, and usage mining one cannot expect to find the meaningful patterns. To analyze usage data without pre-processing makes the web-log files more complex and meaningless .So we are using some preprocessing methodology. The results which we received after applying methodology, shows that the size of the web-log files reduces and also the log files have more useful records in comparison to the initial log file.
So, in this section we have discussed some previous work which is related on pre-processing techniques and web-log files. Some works like user authentication and identification, pattern discovery & analysis and also path completion and some researcher works on the number of hits of some websites. But here we worked on preprocessing the web log files which is used for mining process. This paper concentrates only on log content collections, Analysis pattern, cleaning, merging, user identification and session identification process to improve the quality of web log file.

Selecting Proper Schema for Design
This research paper include methodology which is used to reduce the size of web-log files to increase

the quality of log files with the help of data cleaning and merging the number of web-log files and also access the contents of user identification contents to check the records which is accessed by different and unique users.

Proposed Methodology

We know that web data mining architecture have following steps:

- Integration and merging of sample web log files.
- Pre-processing technique executed in which data cleaning, user identification and session identification occurs.
- After that pattern discovery.
- Then Pattern Analysis.
- And at last analysis reports.

So, this step is the proposed methodology which we used to make our web-log files more useful and small.

Web Log File

Proxy Server Log file is collected from Proxy server and www.anandeducation.org web site log file are composed and analyzed in this work. The data of the month Feb 2012 to June 2012 log file are merged to find source log file. This web site focuses on engineering education and also provided useful information about education. This Log File is the input for the pre-processing point. Almost 2212 visitors visited this web site during these months. The size of the log file increases day by day. The log file which we use in this research paper contains 129 days web log data. The total number of records found is 1,148,872 from this log file. The results of the analysis of these log files are mentioned in table.

Data cleaning

In this paper we used data cleaning technique to remove the useless or unwanted records those are not as useful as other data like user identification, user session, number of hits, number of visits, etc. So we propose a comparative algorithm for data cleaning which is very useful and also which reduces the size of web log file and improve the quality of contents in the log file.

Algorithm

Step 1: Begin

Step 2: Repeat to read multiple Log files from Proxy server web-logs

Step 3: Merge all Log files in to single sample web-log file

Step 4: Start data cleaning (If such type of extension like <jpeg, mpeg, js, gif, css> , < different error like http 404, refuse connection from proxy server> found then remove from single sample web-log records.

Step 5: Analyze the result log file from previous log file.

Step 6: Repeat step 4 and 5 until end of sample log file.

Step 7: End

Test

We have analyzed the visits history according to number of single site with different users and date.

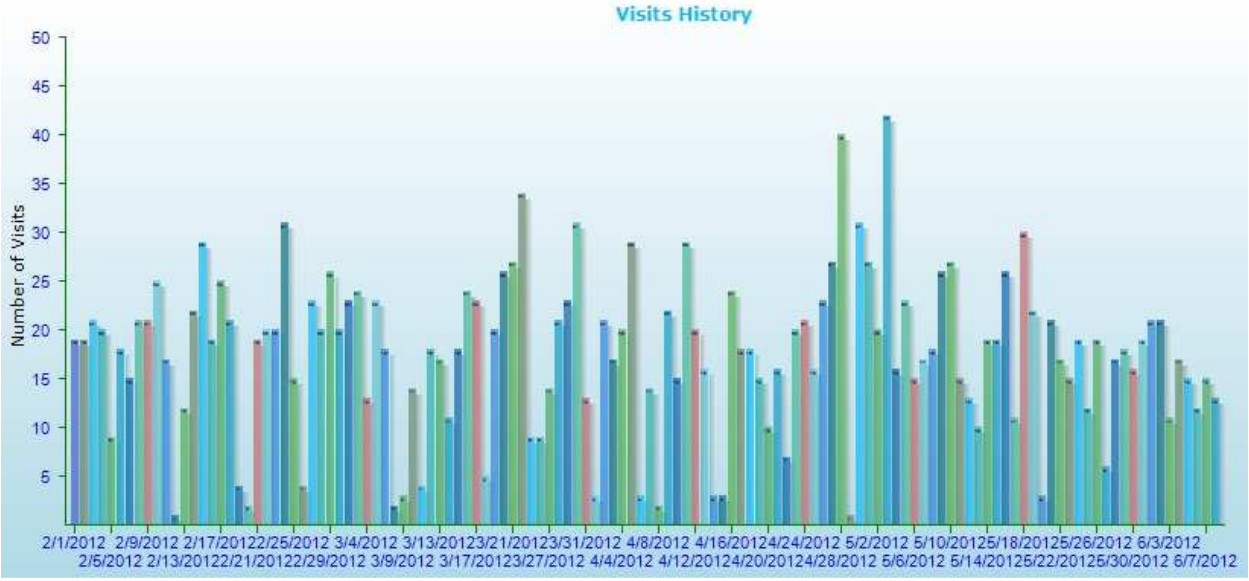

Fig-1

We have used the below proxy server sample log file for analyzing the result web-log file:

log20120215.txt - Notepad

File Edit Format View Help

```
192.168.1.175 - f21 [15/Feb/2012:08:21:09 +0530] "GET http://in.msn.com/?
rd=1&ucc=IN&dcc=IN&opt=0 HTTP/1.1" 200 31121 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:09 +0530] "GET
http://apac.rel.msn.com/default.aspx?di=9&pi=9574&ps=95101&pageid=220419&mk=en-
in&tp=in.msn.com&fk=C&gp=P&optkey=default&parsergroup=hops HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:09 +0530] "GET
http://apac.rel.msn.com/default.aspx?di=9&pi=9574&ps=95101&pageid=220419&mk=en-
in&tp=in.msn.com&fk=C&gp=P&optkey=default&parsergroup=hops HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:09 +0530] "GET
http://b.scorecardresearch.com/b?rn=1329323457369&c7=http%3A%2F%2Fin.msn.com%2F%3Frd
%3D1%26ucc%3DIN%26dcc%3DIN%26opt%3D0&c1=2&c2=3000001 HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:09 +0530] "GET
http://b.scorecardresearch.com/b?rn=1329323457369&c7=http%3A%2F%2Fin.msn.com%2F%3Frd
%3D1%26ucc%3DIN%26dcc%3DIN%26opt%3D0&c1=2&c2=3000001 HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:11 +0530] "GET http://in.msn.com/IN_IAD.aspx
HTTP/1.1" 200 5343 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:12 +0530] "GET http://www.google.co.in/
HTTP/1.1" 200 22877 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:12 +0530] "GET http://www.google.co.in/csi?
v=3&s=webhp&action=&e=17259,18167,33551,34636,34784,35055,35211,36046,36121,36525,366
04,36683,36790,36813&ei=mB07T_7OHNDMrQejtN3hBg&imc=1&imn=1&imp=0&rt=xjsls.110,prt.125
,xjses.172,xjsee.328,xjs.344,ol.359,iml.125 HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:12 +0530] "GET http://www.google.co.in/csi?
v=3&s=webhp&action=&e=17259,18167,33551,34636,34784,35055,35211,36046,36121,36525,366
04,36683,36790,36813&ei=mB07T_7OHNDMrQejtN3hBg&imc=1&imn=1&imp=0&rt=xjsls.110,prt.125
,xjses.172,xjsee.328,xjs.344,ol.359,iml.125 HTTP/1.1" 200 0 "HTTP" ""
192.168.1.175 - f21 [15/Feb/2012:08:21:14 +0530] "GET
```

Fig-2

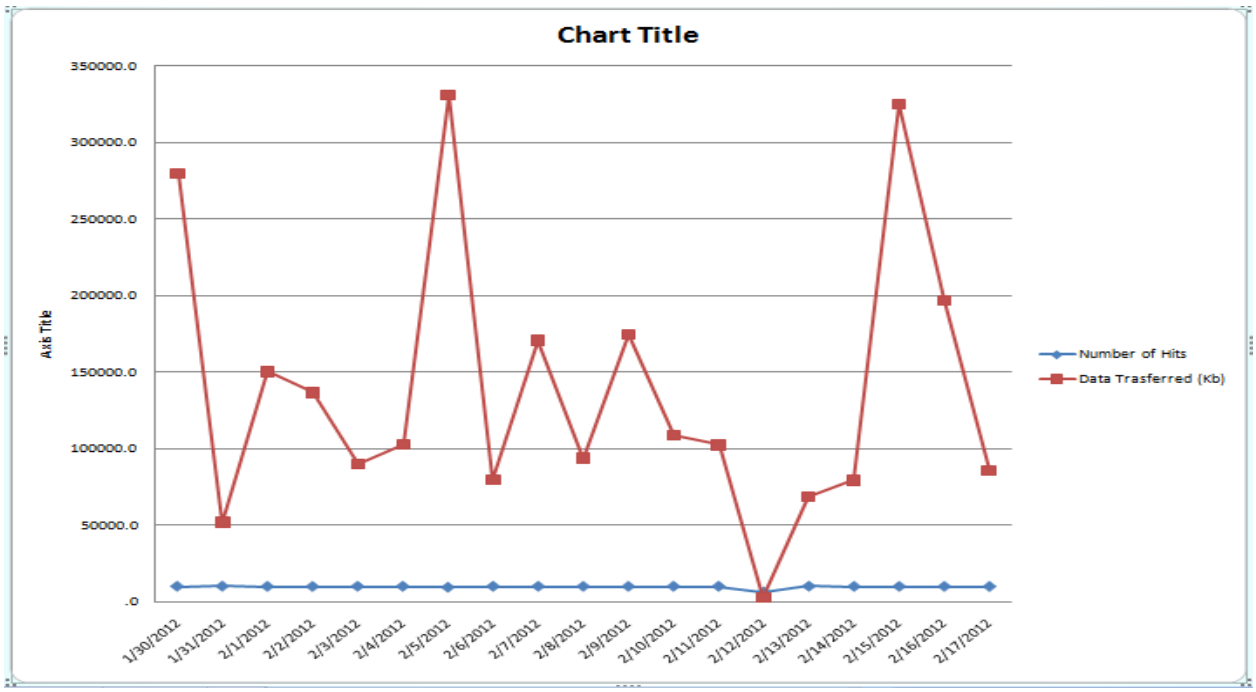

Fig-3

In fig-3 we find the hit ratio and data transfer rate of different types of web-log data according to date.

Conclusion

By using the resultant log file we have seen that the server have a best web-log data for making a good database for web-log mining. We summarized the web log data with introducing algorithm and found better results and also for improving the quality of the web-log file

References

[1]. J. Han, and M. Kamber, "Data Mining Concepts and Technique". Morgan Kaufmann Publishers, 2000

[2]. Anitha.A, 2010. "A New Web Usage Mining Approach for Next Page Access", International Journal of Computer Applications (0975-8887), Vol. 8, No.11, pp.7-10.

[3]. H. Srinath and S. S. Ramanna, "Web Caching: A technique to speed up access to web contents," Resonance Vol.7, No.7. , 2002.

[4]. Baglioni.M, U. Ferrara, A. Romei, S. Ruggieri, and F. Turini1, 2003. "Preprocessing and Mining Web Log Data for Web Personalization", Advances in Artificial Intelligence, Springer, Vol.2829, pp.237-249.

[5]. Li Chaofeng , 2011. "Research and Development of Data Preprocessing in Web Usage Mining", International Journal of computer applications, pp.1311-1315.

[6]. Sanjay Bapu Thakare et al., 2010. "An Effective and Complete Preprocessing for Web Usage Mining", International Journal on Computer Science and Engineering, Vol. 02, No. 03, pp.848-851

[7]. Papoulis, A., Probability, Random Variables, and Stochastic Processes, NY: McGraw Hill, 1991.

[8]. Mrs. Bharati M. Ramageri," Data Mining Techniques and Applications", "Department of Computer Application, Yamunanagar, Nigdi Pune, Maharashtra", Vol. 1 No. 4 301-305.

[9]. Brijesh Kumar Bhardwaj, Saurabh Pal "Data Mining: A prediction for performance improvement using classification" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011