# Pattern Discovery Using Association Rules

Ms Kiruthika M, Mr Rahul Jadhav
Associate Prof., Computer dept.
Fr CRIT, Vashi,
Navi Mumbai, India

Ms Dipa Dixit
Assistant Prof., IT dept.
Fr CRIT, Vashi,
Navi Mumbai, India

Ms Rashmi J
Lecturer, IT dept.
FrCRIT, Vashi
Navi Mumbai, India

Ms Anjali Nehete, Ms Trupti Khodkar
Fr CRIT, Vashi,
Navi Mumbai,India

***Abstract*— The explosive growth of Internet has given rise to many websites which maintain large amount of user information. To utilize this information, identifying usage pattern of users is very important. Web usage mining is one of the processes of finding out this usage pattern and has many practical applications. Our paper discusses how association rules can be used to discover patterns in web usage mining. Our discussion starts with preprocessing of the given weblog, followed by clustering them and finding association rules. These rules provide knowledge that helps to improve website design, in advertising, web personalization etc.**

***Keywords- Weblogs; Pattern discovery; Association rules.***

## I. INTRODUCTION

Association rule is one of the data mining tasks which can be used to uncover relationship among data. Association rule identifies specific association among data and its techniques are generally applied to a set of transactions in a database. Since, amount of data handled is extremely large, current association rule techniques are trying to prune the search space according to support count.

Rules discovery finds common rules in the format A→B, meaning that, when page A is visited in a transaction, page B will also be visited in the same transaction. These rules may have different values of the confidence and support [1].

Confidence is the percentage between the number of transactions containing both items of the rule and the number of transactions containing just the antecedent. Support is the percentage of transactions in the rule is true.

In the context of Web Usage Mining, association rules refers to set of pages which are accessed together with a minimum support value which can help in organizing Web space efficiently.

For example: Consider if 70% of the users who accessed

get/programs/courses/x.asp also accessed

get/programs/courses/y.asp, but only 30% of those who accessed get/programs/courses accessed get/programs/courses/y.asp, then it shows that some information in x.asp is making the clients access y.asp.

This inference helps the designers to decide on designing a link between the above two pages. The task of association rule mining has received a great deal of attention. Association rule mining is still one of the most popular pattern-discovery methods in KDD.

Hence, we would like to use association rules for pattern discovery analysis of Web Server Logs.

### *A.* Web Server Log

Web Servers are used to record user interactions whenever any request for resources are received.

A server log is a log file automatically created and maintains a history of page requests. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information [2].

But to understand the user behavior, analysis of these weblogs is a must. This analysis can help in understanding the user access patterns and can lead to grouping of resource providers, restructuring of websites, pinpointing effective advertising locations, targeting specific users for specific advertisements.

Unprocessed log are shown below:

#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer)

2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=323-21-603&q=3&y=2002&id=671 200 156 http/1.1 www.cs.depaul.edu

mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win +9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21) http://www.cs.depaul.edu/courses/syllabilist.asp depaul.edu/courses/syllabilist.asp

2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/default.asp – 200 16 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+ 98;+digext) http://www.cs.depaul.edu/news/news.asp?theid=573

2002-04-01 00:00:29 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) –

A sample log file converted into database in shown below in Table I.

## II. Scope and Applications

The user access log has very significant information about a Web server. A Web server access log contains a complete history of webpages accessed by clients. By analyzing these logs, it is possible to discover various kinds of knowledge, which can be applied to improve the performance of Web services.

Web usage mining has several applications and is used in the following areas:

1) It offers users the ability to analyze massive volume of click stream or click flow data.
2) Personalization for user can be achieved by keeping track of previously accessed pages which can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.
3) By determining access behavior of users, needed links can be identified to improve the overall performance of future accesses.

Web usage patterns are used to gather business intelligence to improve customer attraction, customer retention, sales, marketing, and advertisements cross sales. Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries.

## III. Proposed System

We would like to propose a system which would discover interesting patterns in these weblogs. Weblogs has information about accesses to various Web pages within the Web space associated with a particular server.

In case of Web transactions, association rules capture relationships among pageviews based on navigation patterns of users.

### *A.* Steps involved in the proposed system

Our proposed system would involve the following steps:

1) The input is a set of Weblogs for which we have to find association rules. We have chosen University Web server logs from www.cs.depaul.edu site
2) The server logs contain entries that are redundant or irrelevant for data mining tasks.
3) The Data cleaning process will select a subset of fields that are relevant for the task.
4) These selected attributes are then stored into a database.
5) Using a simple clustering approach these entries are divided into clusters or segmented.
6) Now, association rule mining is applied on these clusters, to obtain association rules having minimum support and confidence.
7) As a result of association rule mining, interesting patterns can be discovered and client's web usage can be evaluated.

TABLE I: A SAMPLE LOG FILE IN TABLE FORMAT.

| T no | Client IP | Date time | Method | Server IP | Port | URI Stem |
|---|---|---|---|---|---|---|
| 0 | 202.185.122.151 | 11/23/200 3 4:00:01PM | GET | 202.190.126.85 | 80 | /index.asp |
| 1 | 202.185.122.151 | 11/23/200 3 4:00:08 PM | GET | n202.190.126.85 | 80 | /index.asp |
| 2 | 210.186.180.199 | 11/23/200 3 4:00:10 PM | GET | 202.190.126.85 | 80 | /index.asp |
| 3 | 210.186.180.199 | 11/23/200 3 4:00:13 PM | GET | 202.190.126.85 | 80 | /tutor/include/style03.css |
| 4 | 210.186.180.199 | 11/23/200 3 4:00:13 PM | GET | 202.190.126.85 | 80 | /tutor/include/detectBrowser_cookie.js |

## IV. DESIGN

### A. Flow Diagram:

The flowchart for pattern discovery using association rules is given in fig 1.

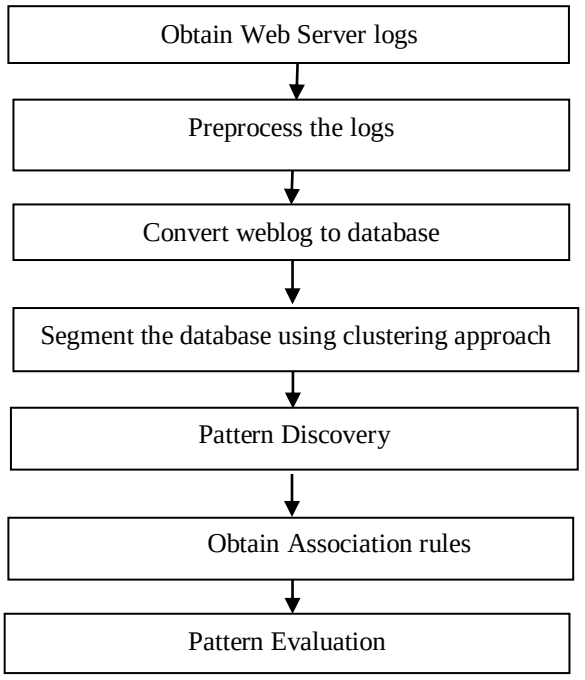


Figure 1 Flow diagram for pattern discovery of weblogs

Each of these blocks is explained in detail as follows:

*1) Obtain Web Server logs:*

Web server log is a file which is created and maintained by the webserver. We are analyzing the log file of the site: www.cs.depaul.edu. It is a text file. The file follows the extended log file format.

*2) Preprocessing the logs:*

The weblog created by the webserver contains details of all requests. It contains lot of irrelevant, incomplete data. Preprocessing involves removing such data.

*3) Conversion of log file to database:*

The weblog cannot be directly used for data mining. The dataset is converted to a database. This involves creating a database and then importing the log file to the MySQL database table.

*4) Segmenting the database:*

In this step, the database is segmented into clusters depending on the support count. After this a number of small clusters are obtained. Depending on the need, these clusters can be analyzed. Clustering web usage data allows the Web master to identify groups of users with similar behaviors for which personalized versions of the Web site may be created.

*5) Pattern Discovery:*

The next step is pattern discovery. Once the clusters are formed they are studied to recognize patterns within the entries of the clusters.

*6) Association rules:*

Association rules show relationship among different items. In case of Web mining, an example of an association rule is the correlation among accesses to various web pages on a server by a given client. Such association rules are obtained in this step

*7) Pattern Evaluation:*

The association rules obtained in the earlier step help in establishing relationships among data items. These association rules are evaluated to understand the information they provide. The interpretations of the rules provide useful knowledge.

### B. Implementation

The following diagrams illustrate the steps of implementation.

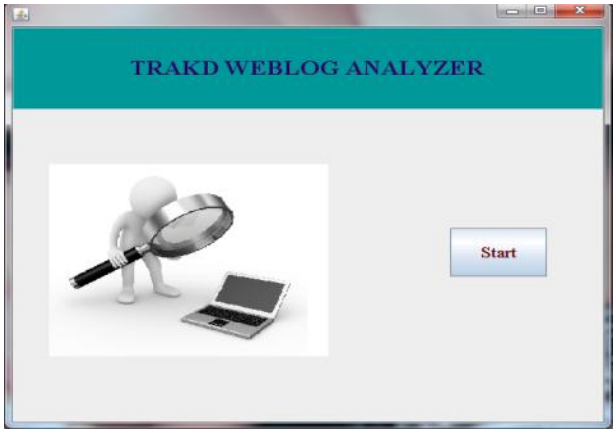


Figure 2 Welcome screen of weblog analyzer

**Step 1:** Weblog of University website hosted on a web server were obtained from www.cs.depaul.edu. There are 5061 records. The following figure3 shows the unprocessed weblog file.

```
#Software: Microsoft Internet Information Services 5.0
#Version: 1.0
#Date: 2002-04-01 00:00:00
#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status t
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=32
2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/default.asp - 200 16 http/1.1
2002-04-01 00:00:29 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depa
2002-04-01 00:00:29 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/default.asp - 200 94
2002-04-01 00:00:30 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http
2002-04-01 00:00:30 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 62 http/1.1 www.c
2002-04-01 00:00:30 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 6
2002-04-01 00:00:32 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /resources/ug_scholarships.asp section=
2002-04-01 00:00:34 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-
2002-04-01 00:00:35 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabisearch.asp -
2002-04-01 00:00:36 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/default.asp - 2
2002-04-01 00:00:40 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/nsf_scholarships.asp - 200 62
2002-04-01 00:00:44 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/2002/gradse2002
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /resources/gae_guide.asp |-|0|404
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /shared/404.asp 404;http://www.cs
2002-04-01 00:01:07 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/courses.asp dep
2002-04-01 00:01:09 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 71
2002-04-01 00:01:14 w010.z064221069.chi-il.ds1.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /people/facultyinfo.asp i
2002-04-01 00:01:15 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 post /cti/advising/display.asp - 200 1625
2002-04-01 00:01:20 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-
2002-04-01 00:01:31 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=31
2002-04-01 00:01:36 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 71
2002-04-01 00:01:39 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 post /cti/advising/display.asp - 302 2875
2002-04-01 00:01:43 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 get /cti/advising/display.asp - 200 47 htt
2002-04-01 00:01:47 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 post /courses/syllabilist.asp - 2
2002-04-01 00:01:53 12-249-142-45.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 443 get /courses/syllabisearch.asp -
2002-04-01 00:01:58 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=554-
2002-04-01 00:02:01 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=
2002-04-01 00:02:08 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 get /cti/advising/includes/faculty/account
2002-04-01 00:07:21 cache-mtc-ad01.proxy.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /people/default.asp - 200 47 ht
```

Figure 3 Unprocessed log file of www.cs.depaul.edu

**Step 2:** The next step is to convert the log file to database.

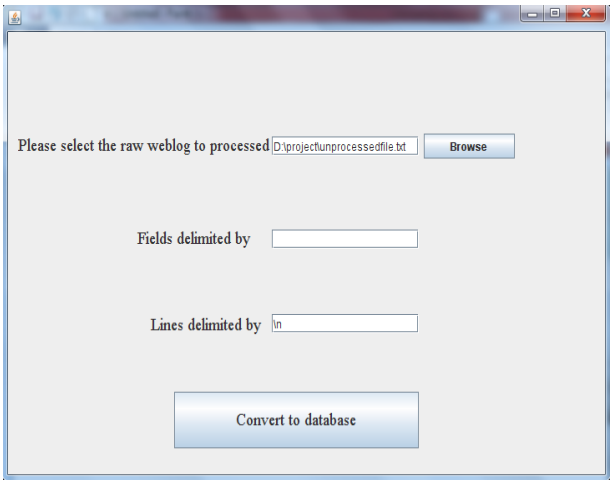


Figure 4 Selection of log file, specifying line and fields delimiters for the selected log file

The steps involved in the conversion of dataset to database are as follows:

- Log on to MySQL command line client.
- Create a table with all required attributes.
- Import the log files into database.

The MySQL commands are shown in figure below:

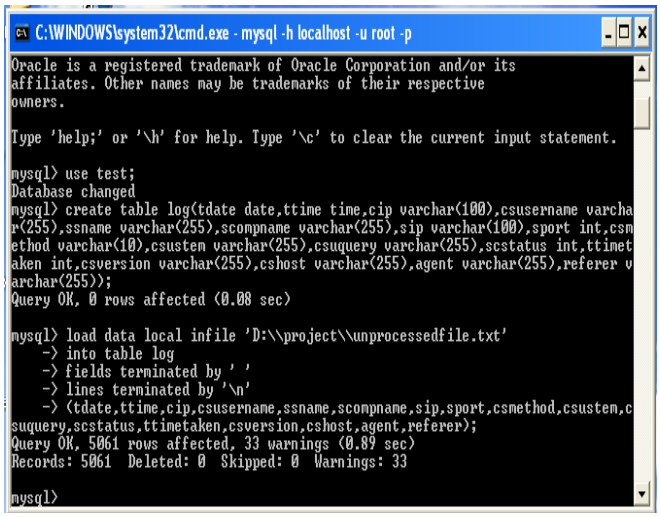


Figure 5 MySQL command prompt showing the commands used to create and load log database.

The database containing the entries of weblog is shown below in Fig 6

**Step 3:** The database has 5061 records. The count of entries for different IP addresses is obtained. There are entries having very low support count. Such entries need not be considered. The database is segmented into clusters having support count more than 20.

**Step 4:** The entries for IP addresses having support count greater than or equal to 30 are used for further analysis. There are 8 unique IP addresses having support count greater than or equal to 30.

These IP addresses are shown in fig 10.

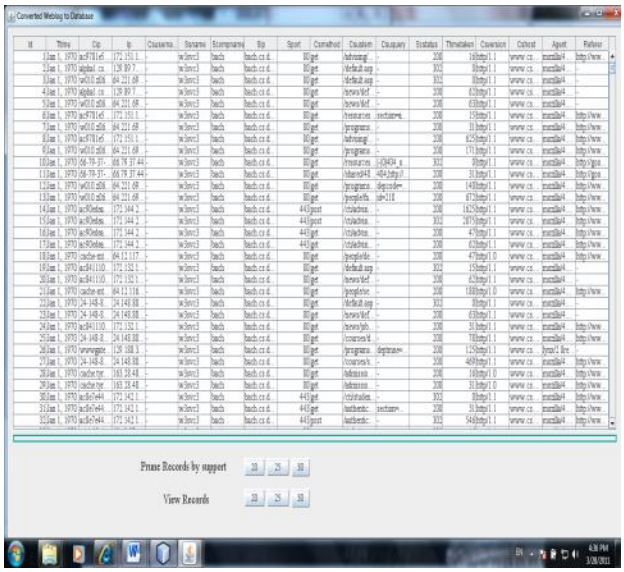

Figure 6 The database of weblog entries

The following fig 7 shows the entries of IP addresses having support count greater than or equal to 20.

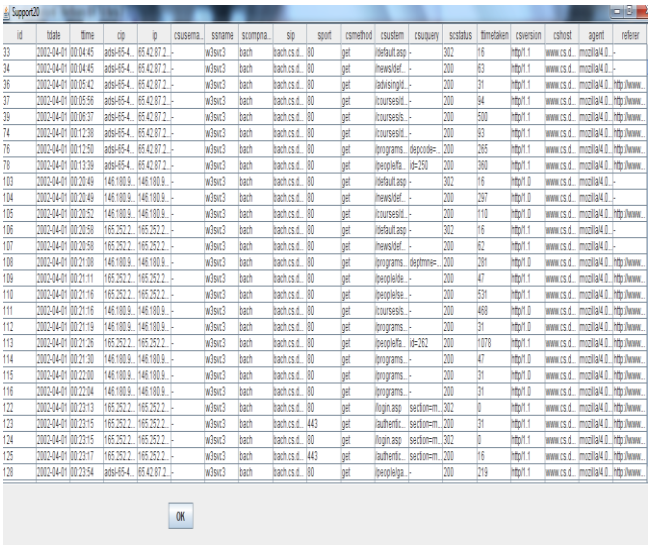

Figure 7 Weblog entries with IP addresses having support count greater than or equal to 20

The following fig 8 shows the entries of IP addresses having support count greater than or equal to 25.

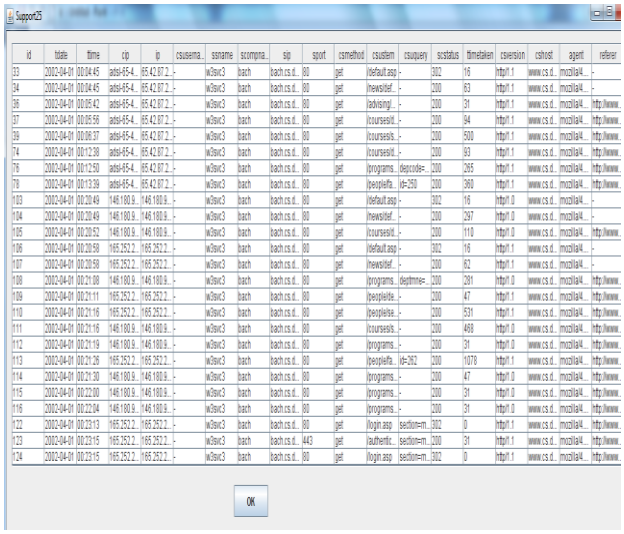

Figure 8 Weblog entries with IP addresses having support count greater than or equal to 25

The following fig 9 shows the entries of IP addresses having support count greater than to equal to 30

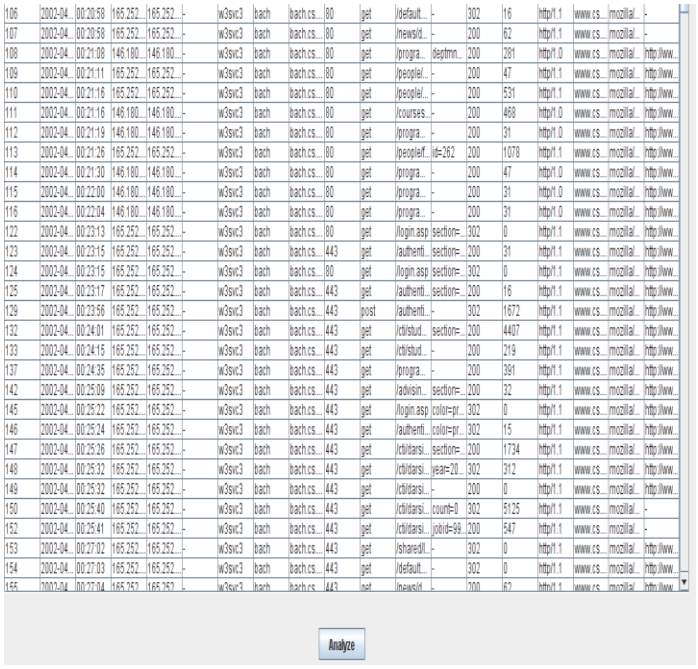

Figure 9 Weblog entries with IP addresses having support count greater than or equal to 30

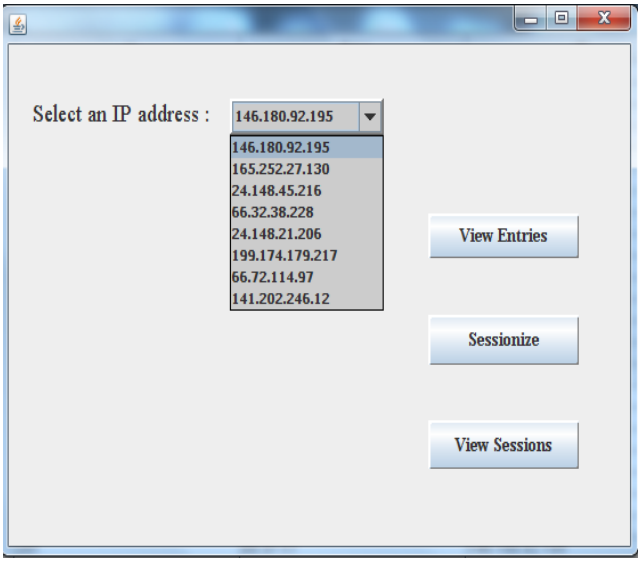


Figure 10 IP addresses having support count greater than or equal to30.

The entries of the IP address selected by the user can viewed by clicking on the 'view entries' button.

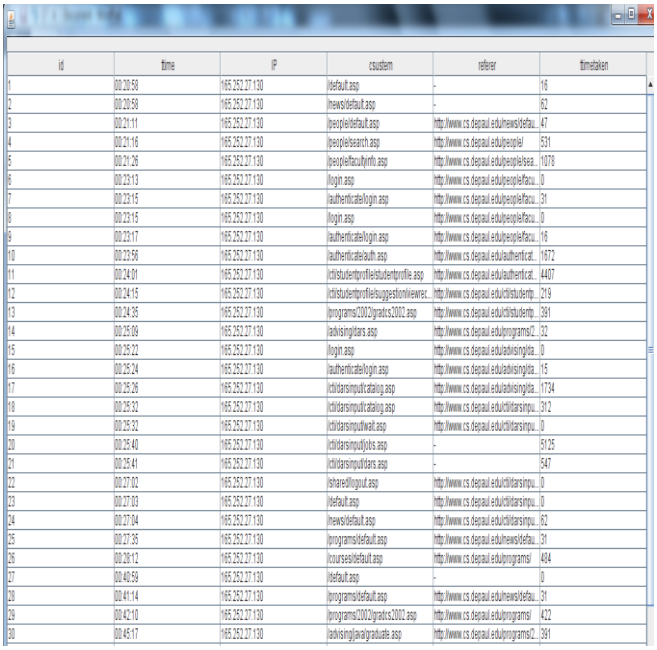

Figure 11 Entries of IP address selected from the drop down menu.

Sessions are identified. The session time is taken to be 5 minutes.

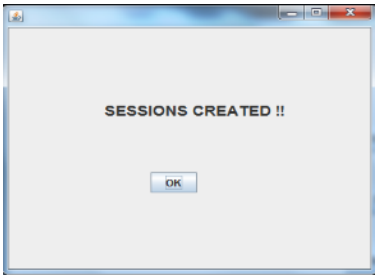


Figure 12 Screen which appears after sessions are created.

Depending on the pages requested the entries of each session are classified among 5 different predefined classes. The entries and the classification of entries of a session are shown in figure below.

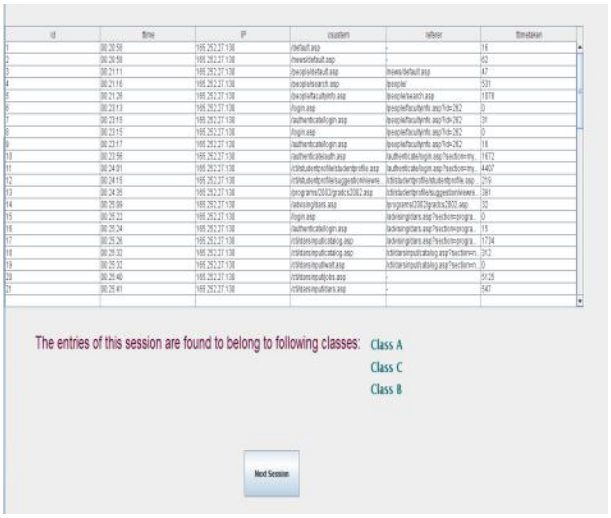


Figure 13 Entries belonging to session of 5 minutes and the classes to which the entries are classified.

After all sessions are viewed, the association rules can be viewed. These association rules show the relation between the IP address and the pages requested by the clients from that IP address. The association rules for IP address 165.252.27.130 are shown in the following figure.


Figure 14 Association rules for the IP address 165.252.27.130

## V. Conclusion

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years.

In this paper, implementation of a system for pattern discovery using association rules is discussed as a method for Web Usage Mining. Different transactions that are closely related to each other are grouped together by the use of clustering approaches on the preprocessed dataset.

The analysis of such clusters will lead to discovery of strong association rules. We obtained all significant association rules between items in the large database of transactions. The relation between different page requests was found.

The support and the confidence values of extracted rules are considered for obtaining the interest of the web visitors. Consequently, the number of hit can be increased by analyzing the visitor attitude.

The approach discussed in this paper, helps the web designers to improve their website usability by determining related link connections in the website.

## References

[1] M. Henri Briand, M. Fabrice Guillet, M. Patrick Gallinari, M. Osmar Zaaiane, "*Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support",* World Academy of Science, Engineering and Technology 48 2008.

[2] Mr. Sanjay Bapu Thakare, Prof. Sangram. Z. Gawali, "*A Effective and Complete Preprocessing for Web Usage Mining*", Expert Systems with Applications, 36(3), 6635-6644.

[3] Resul Daş, İbrahim Türkoğlu, "*Extraction of Interesting Patterns through Association Rule Mining For Improvement of Website Usability*", Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006 IEEE.

[4] Bamshad Mobasher, Namit Jain, Eui-Hong (Sam) Han, Jaideep Srivastava,"*Web Mining: Pattern Discovery from World Wide Web Transaction*", Proc. IEEE International Conference Multimedia Computing Systems, Hiroshima, Japan, June, 1996.

[5] B. Mobasher, R. Cooley, and J. Srivastava, "*Automatic personalization based on Web usage mining*" Communications of the ACM, vol. 43, pp. 142-151, 2000.

[6] C. R. Anderson, P. Domingos, and D. S.Weld, "*Adaptive Web Navigation for Wireless Device*" Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 879–884,2001.

[7] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "*Visualization of navigation patterns on a Web site using model-based clustering,*" Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.280-284, 2000.

[8] Dr.R.Lakshmipathy, V.Mohanraj, J.Senthilkumar, Y.Suresh, "*Capturing Intuition of Online Users using a Web Usage Mining*" Proceedings of 2009 IEEE International Advance Computing Conference (IACC 2009)Patiala, India, 6-7 March 2009.

[9] Kiruthika M, Dipa Dixit, Pranay Suresh, Rishi M. "An Approach to Convert Unprocessed Weblogs to Database Table"

[10] "Identifying User Behavior by Analyzing Web Server Access Log File" by K R Suneetha, R Krishnamoorthi.