International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-3, Issue-1, March 2014

A New Preprocessing Approach for Mining Association Rule from Large Database

Kalpana Wani, J.W. Bakal

Abstract: ***As we know now a day's internet plays vital role in serving the needs of the user's on web. We can search the information, we can order the items and can do many things online. Different options are available for this, but many times we don't have time to go through all alternatives and decide the best one. In such case a Web Page Recommendation System will be helpful to suggest the pages which are most relevant to your current search. The Server log files are generated as a result of an interaction between the client and the service provider on web. Server log file contains the massive hidden valuable information related to the visitors, if we mined this, it can be used for predicting the navigation behavior of the users. However the task of discovering frequent sequence patterns from the server log is challenging as it consist of huge data. Most of the time this data is incomplete and because of that it can't be processed further for generating accurate knowledge. Proposed system focuses on adopting an intelligent technique that can provide personalized web service for accessing related web pages more efficiently and effectively. Proposed system uses two intelligent algorithms for predicting the user behavior's namely FP Growth and Éclat. These algorithms save the time and space problem of existing system. Further from the frequent pages pattern Direct and Indirect Association Rules are generated and based on that Ranking is provided to pages which will help recommendation system to recommend similar search pages. This paper focuses on new approach for preprocess web log data.***

Index Terms: ***Association Rule, Indirect Association Rule, Parsing, Preprocessing.***

I. INTRODUCTION

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.[1] Based on the concept of strong rules, Rakesh Agrawal et al.[2] 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{Onions, 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 hamburger meat. 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.

Revised Manuscript Received on 30 March 2014.
* Correspondence Author
Prof. Kalpana Wani*, Department of Computer Engineering, Mumbai University, PIIT College, New Panvel, Navi Mumbai (M.H), India.
Dr. J.W.Bakal, Department of IT, Mumbai University, SS Jondhale College of Engineering, Dombivali, Mumbai (M.H), India.

Web usage mining, intrusion detection, Continuous production, and bioinformatics. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

The amount of information on World Wide Web is growing rapidly, as well as the number of Web sites and Web pages per Web site are also increasing day by day. Consequently, it has become more difficult to find relevant and useful information for Web users. Web usage mining is concerned with guiding the Web users to discover useful knowledge and supporting them for decision-making. In that context, predicting the needs of a Web user as we visits Web sites has gained importance. The requirement for predicting user needs in order to guide the user in a Web site and improve the usability of the Web site can be addressed by recommending pages to the user that are related to the interest of the user at that time. This Recommendation uses association rule as explain above to find the navigation behavior of the user.

II. PROPOSED SYSTEM

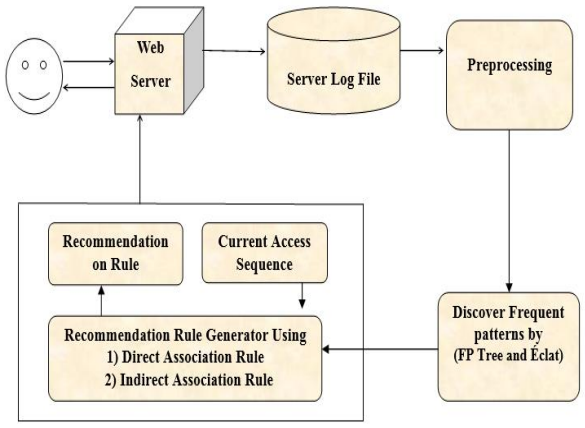

Fig.1. Architecture Diagram of Proposed System

Following steps are used for implementation:

Step 1: Data Cleaning & Preprocessing

Preprocessing is necessary, because Log file contain noisy &ambiguous data which may affect result of mining process. Some of web log file data are unnecessary for analysis process and could affect detection of web attack. Data preprocessing is an important steps to filter and organize only appropriate information before applying any web mining algorithm. Preprocessing reduce log file size also increase quality of available data. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleansing, user identification, session identification.

Step 2: Finding Frequent Pattern

First count the occurrences of pages in database and then reorder the pages with higher count page first and so on. Draw the FP Growth tree and start mining FP tree by considering the suffix as page with lowest count first. Find the conditional pattern base for each page and if length of conditional pattern base is less than some threshold value then find the frequent pattern using Éclat algorithm and if length is greater than or equal to the threshold value then use FP Growth algorithm for finding the frequent pattern of visited pages.

Step 3: Direct Association Rule Generation between the Frequent Access Pages

After getting the frequently visited page sequence filter the pattern by applying some minimum confidence and support value and form the one page association rule between pages of frequent pattern and calculate its support and confidence values and filter it with respect to minimum confidence and support value.

Step 4: Indirect Association Rule Generation between the Frequent Access Pages

In direct association rule of pages there are many pages which are connected to each other indirectly via some transitive pages. Find such pages which are indirectly connected and filter it with respect to minimum confidence and support value.

Step 5: Complex Association Rule Generation

A complex association rule exists if at least one of two component rules exists, i.e., either direct or complete indirect or both of them. The main quality features of both direct and in indirect rules—confidences—are combined within complex association rules. A complex association rule is characterized by the *complex confidence*, $con*(di \rightarrow *dj)$, as follows

$$Con*(di \rightarrow *dj) = \alpha \, .Con(di \rightarrow dj) + (1 - \alpha) \, . \, Con\#(di \rightarrow \#dj)$$

Where α is the direct confidence reinforcing factor $\alpha \in [0, 1]$ Calculate the complex confidence value for different values of α. Setting α we can emphasize or damp the direct confidence at the expense of the complete indirect one. The greater the value of α, the closer the complex confidence to the direct one. Depending on highest value of α rank is assigned to the pages and sends it to recommendation system.

III. SERVER LOG FILE FORMATS

Web Site Analyzer can use the information in HTTP, FTP, and other server log files to analyze a site. The log file formats that Web Site Analyzer can analyze include the following:

- NCSA (Common or Access, Combined, and Separate or 3-Log)
- W3C Extended (used by Microsoft IIS 4.0 and 5.0)
- SunTM ONE Web Server (iPlanet)
- IBM Tivoli Access Manager WebSEAL
- WebSphere Application Server Logs
- FTP Logs
- Custom Log File Format (field information defined by user)

A. NCSA Log Formats:

The NCSA log formats are based on NCSA httpd, and are widely accepted as standard among HTTP server vendors. *Web Site Analyser supports the following options:*

- Common (access log)
- Combined
- Separate with Date (three log) Switching among options is specific to each HTTP server implementation.

1) NCSA Common (access log)

The NCSA Common log format contains only basic HTTP access information. The NCSA Common Log, sometimes referred to as the Access Log, is the first of three logs in the NCSA Separate log format. The Common log format can also be thought of as the NCSA Combined log format without the referral and user agent. The Common log contains the requested resource and a few other pieces of information, but does not contain referral, user agent, or cookie information. The information is contained in a single file.

2) NCSA Combined Log Format

The NCSA Combined log format is an extension of the NCSA Common log format. The Combined format contains the same information as the Common log format plus three (optional) additional fields: the referral field, the user agent field, and the cookie field.

3) NCSA Separate (three-log format)

The NCSA Separate log format, sometimes called three-log format, refers to a log format in which the information gathered is separated into three separate files (or logs), rather than a single file. The three logs are often referred to as:

- Common log or access log
- Referral log
- Agent log

The three-log format contains the basic information in the NCSA Common log format in one file, and referral and user agent information in subsequent files. However, no cookie information is recorded in this log format.

a) Common or access log

The first of the three logs is Common log, sometimes referred to as the access log, which is identical in format and syntax to the NCSA Common log format.

b) Referral log

The referral log is the second of the three logs. The referral log contains a corresponding entry for each entry in the common log.

c) Agent log

The Agent log is the third of the three logs making up the three-log format. Like the referral log, the agent log contains a corresponding entry for each entry in the common log.

B. W3C Extended Log Format

An extended log file contains a sequence of lines containing ASCII characters terminated by either the sequence LF or CRLF. Log file generators should follow the line termination convention for the platform on which they are executed. Analyzers should accept either form. Each line may contain either a directive or an entry.

Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by whitespace, the use of tab characters for this purpose is encouraged. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself. Lines beginning with the # character contain directives.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The following directives are defined:
=>**Version:** <integer>.<integer>
The version of the extended log file format used.

=>**Fields:** [<specifier>...]
Specifies the fields recorded in the log.

=>**Software:** string
Identifies the software which generated the log.

=>**Start-Date:** <date> <time>
The date and time at which the log was started.

=>**End-Date:**<date> <time>
The date and time at which the log was finished.

=> **Date:**<date> <time>
The date and time at which the entry was added.

=>**Remark:** <text>
Comment information. Data recorded in this field should be ignored by analysis tools.
The directives Version and Fields are required and should precede all entries in the log. The Fields directive specifies the data recorded in the fields of each entry.
The following is an example file in the extended log format:
#Version: 1.0
#Date: 12-Jan-1996 00:00:00
#Fields: time cs-method cs-uri
00:34:23 GET /foo/bar.html12:21:16 GET /foo/bar.html
12:45:52 GET /foo/bar.html
12:57:34 GET /foo/bar.html
Fields
The #Fields directive lists a sequence of *field identifiers* specifying the information recorded in each entry. Field identifiers may have one of the following forms:
Identifier: Identifier relates to the transaction as a whole.
Prefix-identifier: Identifier relates to information transfer between parties defined by the value prefix.
Prefix (header)
Identifies the value of the HTTP header field header for transfer between parties defined by the value prefix. Fields specified in this manner always have the value <string>.
The following prefixes are defined:
c Client
s Server
r Remote
cs Client to Server.
Sc Server to Client.
Sr Server to Remote Server, this prefix is used by proxies.
Rs Remote Server to Server, this prefix is used by proxies.
X Application specific identifier.
The identifier cs-method thus refers to the method in the request sent by the client to the server while sc(Referer) refers to the eferrer: field of the reply. The identifier c-ip refers to the client's ip address.
Identifiers.
The following identifiers do not require a prefix
date : Date at which transaction completed, field has type <date>
time : Time at which transaction completed, field has type <time>
time-taken : Time taken for transaction to complete in seconds, field has type <fixed>
bytes : bytes transferred, field has type <integer>
cached : Records whether a cache hit occurred, field has type <integer> 0 indicates a cache miss.
Special fields for log summaries.
Analysis tools may generate log summaries. A log summary entry begins with a count specifying the number of times a particular even occurred. For example a site may be interested in a count of the number of requests for a particular URI with a given eferrer: field but not be interested in recording information about individual requests such as the IP address.
The following field Is mandatory and must precede all others:
count : The number of entries for which the listed data, field has type <integer>
The following fields may be used in place of time to allow aggregation of log file entries over intervals of time.
Time-from : Time at which sampling began, field has type <time>
time-to : Time at which sampling ended, field has type <time>
interval :Time over which sampling occurred in seconds, field has type <integer>
Entries
This section describes the data formats for log file field entries. These formats are chosen so as to avoid ambiguity, minimize the difficulty of generation and parsing and provide for human readability.Each logfile entry consists of a sequence of fields separated by whitespace and terminated by a CR or CRLF sequence. The meanings of the fields are defined by a preceding #Fields directive. If a field is omitted for a particular entry a single dash "-" is substituted.

C. SunTM ONE Web Server(iPlanet)

A log file in the One Web Server (iPlanet) format contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space. If a field is unused in a particular entry dash "-" marks the omitted field. Directives, which appear at the beginning of the log, define the information and format contained in the entries.

D. IBM Tivoli Access Manager WebSEAL

IBMR TivoliR Access Manager WebSEAL is a resource security manager for Web-based resources. To import data from IBM Tivoli Access Manager WebSEAL, you must configure WebSEAL to create logs in the NCSA Combined log format.

E. WebSphere Application Server (WAS) Log Format

WebSphere Application Server (WAS) version 4.0 provides an API for application level logging called Analytic Logging Service (ALS). You can use call this API from your own Web applications. The IBM WebSphere Personalization server uses this API to log data. This API supports three logging methods: HTTP, database, and file. The record format for logged data differs slightly based on the logging method used.

IV. DATA PREPROCESSING

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

As shown in the Proposed System Architecture Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Data preprocessing help us to extract proper format of data and generate the user sessions and their navigation paths.

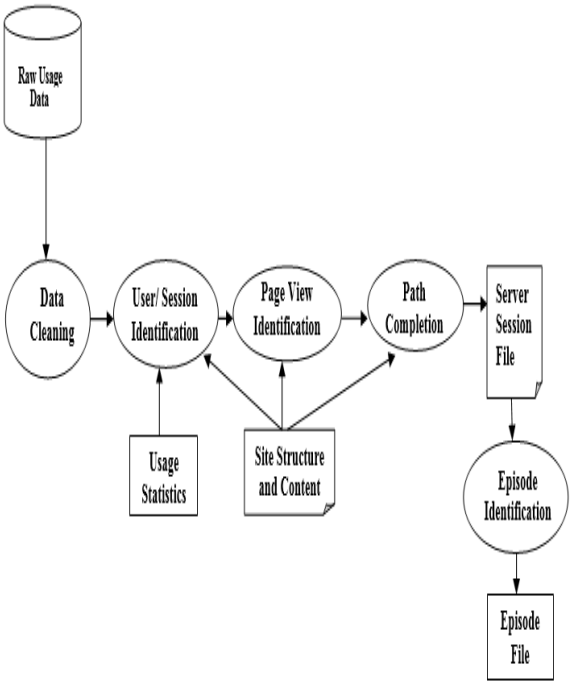

Fig.2 Data Preprocessing

V. SYSTEM IMPLEMENTATION

In our system instead of using any readymade tool for data preprocessing we have developed our own processing method which process web log data present in W3c format i.e. in Extended Log Format. Details of this format are explained above. Home page of proposed system is shown below. It consist of the display of the Architecture Diagram of the system and a brief explanation of the working of the system.

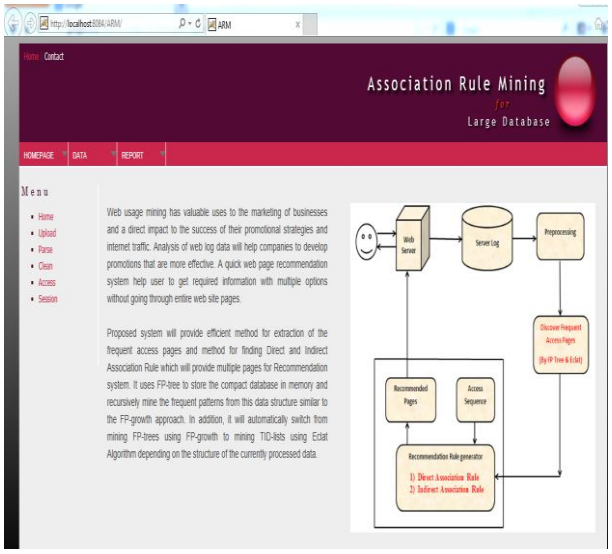

Fig.3 Home Page of Proposed System

On left hand side of the home page Menu options are given which are listed below

1) Home
2) Upload
3) Parse
4) Clean
5) Access
6) Session

A. Home

First option Home is the default option which display home page of the system. Second option is Upload

B. Upload

Second option is Upload, which facilitates to upload the weblog file in w3c format. We can browse the file from its location and click on upload button as shown below.

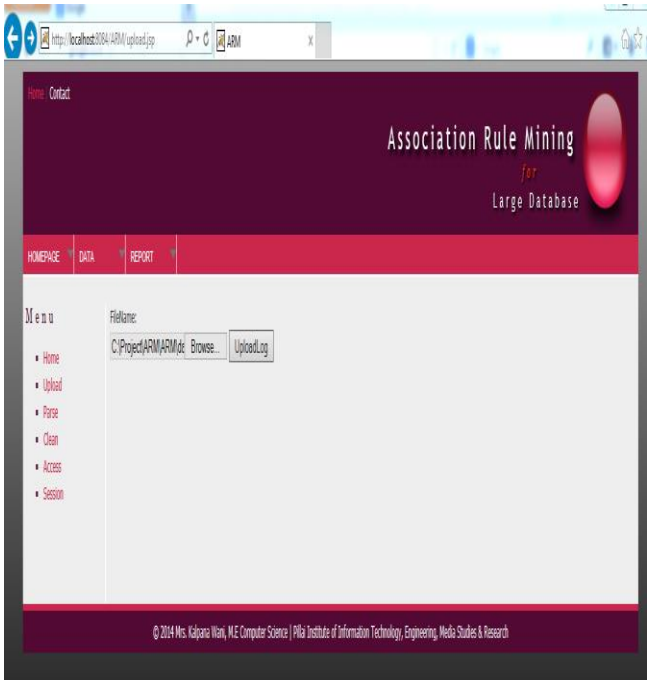

Fig.4 Web Log File Upload

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

After clicking on the upload button if file is uploaded successfully then message will be displayed about successful uploading of file and file contents are displayed as shown below. If file type is not matching then it display the error.

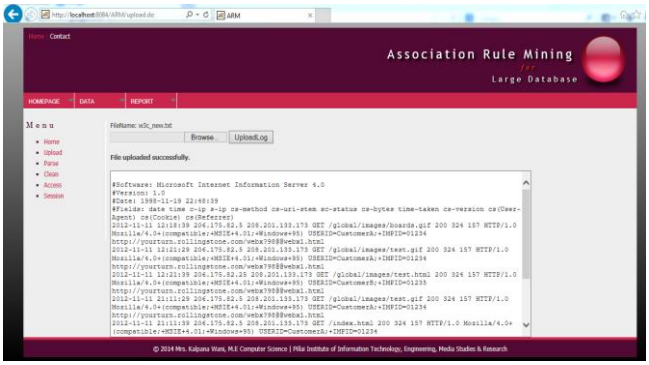

Fig.4 Web Log File Contents

C. *Parse*

Once weblog file is uploaded successfully then we can start preprocessing of the file by selecting the Parse option from the Menu. We have to use Parse option for uploaded file only, if we try to use it without uploading a file it will give error. Following snapshot show that when we select Parse option for uploaded file it automatically display the file name.

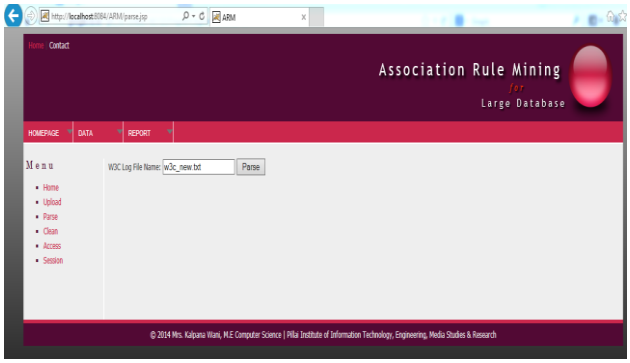

Fig.5 Web Log File Selection for Parsing of Data

After getting required file name click on parse button which will convert W3C format weblog file information in a tabular format as shown below.

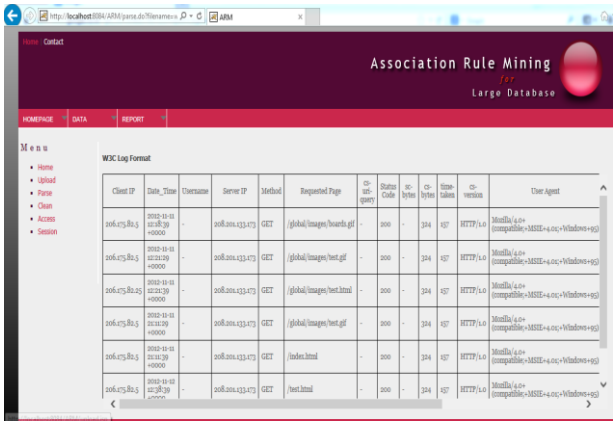

Fig.6 Tabular Representation of Web Log file data

D. *Clean*

After converting web log data into a tabular format we need to clean the data to know accurate information. In web log data many entries are not completed or sometimes we want information about some specific data format etc. Depending upon our requirement we can clean the data and retrieve the information. Different cleaning option provided in this system are given below.

a) Remove failed/invalid request.
b) Select only "GET" method
c) Remove multimedia object request.
d) Remove web robots' request.

We can select either one or multiple option depending upon our data requirements as shown below.

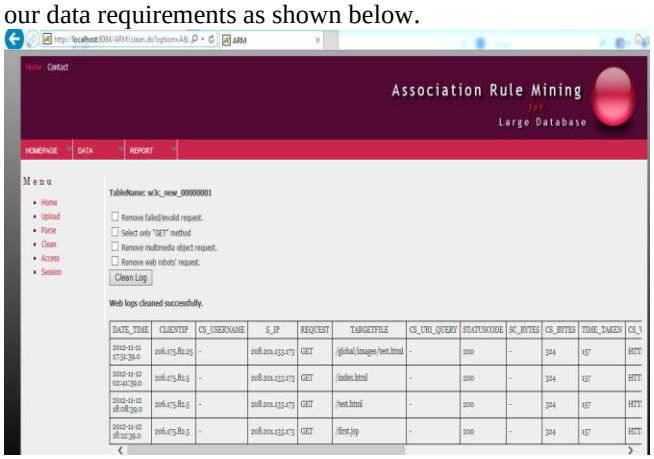

Fig.7 Web Log Data after cleaning

E. *Access*

To know the user access pages list we can use this option. It will display list of users as well as their access pages as shown below.

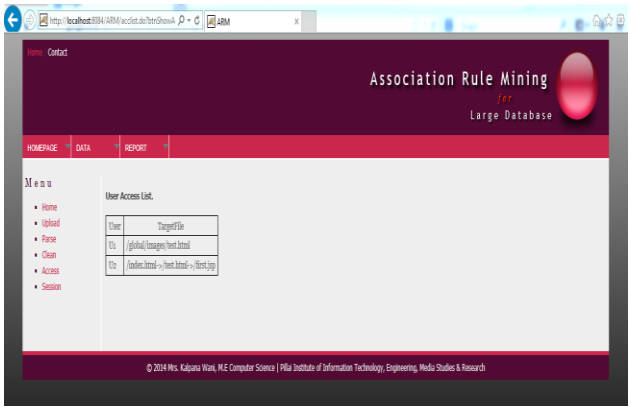

Fig.8. User Access List

F. *Session*

This option gives the details of visited users their sessions and the visited pages in respective sessions as shown in figure.

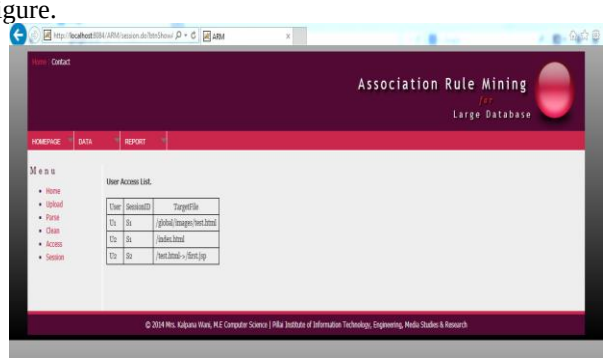

Fig.9 User Session

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Next task to be done is finding the frequent access patterns of the pages and then finding direct and indirect association rule between the frequent access pages to give ranking to them and then recommending for web page recommendation system. Rest of the module are under development.

VI. CONCLUSION

In today's world there is boom of online shopping and because of that if companies want to increase their business it's required to analyze the user behavior. This is possible only with help of web mining and its further processing to find frequent pattern and the association rule between them. This system will help to recommend appropriate page to user.

In future we are planning to complete remaining work by providing security to log file

ACKNOWLEDGMENT

I would like to thank my guide Dr.J.W. Bakal to provide a valuable guidance and inspiring me to publishing paper in International Journals and Conferences.

REFERENCES

1. Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.
2. Agrawal, R.; Imieliński, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207. Do i: 10.1145/ 170035. 170072. ISBN 0897915925.
3. Bina Kotiyalt, Ankit Kumar2, Bhaskar Pant3, R.H. Goudar4, Shiv ali Chauhan5 and Sonam Junee6." User Behavior Analysis in Web Log through Comparative Study of Éclat and Apriori", Proceedings of7'h International Conference on Intelligent Systems and Control (ISCO 2013)
4. N. VENKATESAN, RAMARAJ," FREQUENT ITEMSET MINING WITH BIT SEARCH", Journal of Theoretical and Applied Information Technology, 15 July 2012. Vol. 41 No.1
5. Ravi Bhushan and Rajender Nath," Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques", 978-1-4673-4529-3/12/$31.00c 2012 IEEE
6. PRZEMYSŁAW KAZIENKO." MINING INDIRECT ASSOCIATION RULES FOR WEB RECOMMENDATION", Int. J. Appl. Math. Computer. Sci., 2009, Vol. 19, No. 1, 165–186
7. Bamshad Mobasher, Robert Cooley, Jaideep Srivastava,"Automatic Personalization Based on Web Usage Mining", Communications of the ACM, New York, Volume 43, Issue 8, Aug 2000.
8. Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, IEEE, 2008.
9. Robert. Cooley, Bamshed Mobasher and Jaideep Srinivastava,"Web mining:Infonnation and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.
10. Perkowitz, M., Etzioni, 0.: Adaptive sites: Automatically learning from user access patterns. In: Proc. of the Sixth International WWW Conference, Santa Clara, CA. (1997).
11. Cooley, R., Mobasher, B., & Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. Knowledge Information Systems, 1(1), pp. 5-32. (1999).
12. H. Mannila, H. Toivonen. Discovering generalized episodes using minimal occurrences. In: Proc.Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.
13. Yan, T. W., Jacobsen, M., Garcia-Molina, H., and Dayal, U.1996. From user access patterns to dynamic hypertext linking. In Proceedings of the Fifth international World Wide Web Conference on Computer Networks and ISDN Systems (Paris, France). P. H. Enslow, Ed. Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1007-1014.
14. Ciesielski, V. and Lalani, A., Data mining of web access logs from an academic web site. In Proceedings of the Third
15. International Conference on Hybrid Intelligent Systems (HIS'03).
16. Kalpana Wani, Madhu N.Association Rule Mining for Large Database. In International Journal of Engineering Research and Technology in ISSN:2278-0181(ISO 3297:2007)volume-2 Issue10 oct 2013
17. http://www.w3.org/TR/WD-logfile.html

AUTHOR PROFILE

Prof. Kalpana Wani She has completed her B.E in Computer Engineering, having 7 yrs. Experience. She has published 2 International journal papers, 3 papers at International conferences and 3 papers at National conferences. She is Currently Working as Assistant Processor in Fr.C.R.I.T Vashi, Navi Mumbai and currently pursuing M.E. from PIIT New Panvel, Mumbai University.

Dr. J.W.Bakal He has completed M. Tech., Ph. D. in Computer Engineering, having 27 yrs. experience out of which 7 yrs. he has worked in industry and 20 yrs. in academics. He has published 20 International journal papers and 20 papers at International conferences and more than 50 papers at National conferences. He is chairman Board of Studies, Information Technology, University of Mumbai, Mumbai. He was Chairman of Board of Studies in Computer Engineering of University of Mumbai. He is member of Board of Studies in MCA, University of Mumbai. He is PG and Ph. D. Guide in University of Mumbai. He is member of Academic Council, Research and Recognition committee of University of Mumbai. Currently working as a Principal Shivajirao S. Jondhale College of Engineering, Dombivali. He worked as a Campus Director of Saraswati Education Society's Group of Institution at Diksal. He received a National award, "Achievement in Education excellence award" by NEHRDO, New Delhi. He is actively involved in the development of IETE Mumbai centre. He has served IETE Mumbai centre as Chairman, Vice Chairman, and Hony. Secretary for more than 8 yrs. during his tenure as a Hony. Secretary, IETE Mumbai centre got its own premises at the heart of city in Chembur. He has organized several International and National conferences successfully; ATC 2009 was one of it. During his tenure as a Chairman, IETE Mumbai Centre got the Second Best Centre Award for year 2011-2012 during ATC 2012 at Bangalore. He is member of many technical professional bodies such as Fellow of IETE, New Delhi, Member, IE, Kolkata, **Member,** International Association of Computer Science and Information Technology (IACSIT), Singapore. **Member**, IEEE, USA, Member**,** International Journal of Engineering Research and Industrial Applications (IJERIA), and so on. He was Convener of Diamond Jubilee Year Celebration of IETE at Mumbai Centre. Recently he has been elected as the member of IETE Governing Council 2013-2016, New Delhi and given responsibility as Co-chairman of Academic Committee of IETE New Delhi.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)