IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Web Page Navigation Mining Through Association Rules

Harsha Dubey[1], Dr. Megha Mishra[2], Dr. V. K. Mishra[3]

Research Scholar Department of Computer Science & Engineering, Shri Shankaracharya Technical Campus, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India[1]

Sr. Assistant Professor Department of Computer Science & Engineering, Shri Shankaracharya Technical Campus, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India[2]

Associate Professor Department of Computer Science & Engineering, Bharti College of Engineering & Technology, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India[3]

Abstract: Web mining is the utilization of information mining systems to naturally find and concentrate data from web archives and administrations. Web mining is three sorts: Web use Mining, Web content Mining and Web structure Mining. Web utilization mining is the Process of finding information from the cooperation created by the clients in the types of access logs, program logs, intermediary server logs, client session information, treats. The web server log document naturally made and kept up by a server comprising of a rundown of exercises it performed. The proposed framework is intended for website page forecast in suggestion framework and also it is useful for the investigation of web mining calculation to get incessant consecutive access design from the web log document of web server. The execution in light of the cleaning, fuzzy c means clustering and Association rules and the outcomes are noticed that the effective calculation for web log successive access in accurate example mining.

Keywords: Web Mining, Association rules, Clustering, Web Server log file.

1. INTRODUCTION

In the continuous growth and abundance of information available on the Internet, the World Wide Web has become a huge repository of information. Web mining is the application of data mining to large web data repositories. Web Mining is the use of data mining techniques to automatically discover and extract information from web documents and services. Web Mining comprises of three classes – Web Content Mining, Web Structure Mining and Web Usage Mining.

Web content Mining deals with the discovery of useful information from the web contents or the information or services. Web Structure Mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites. Web Usage mining mines the log data stored in a web server. Data used for web usage mining, can be collected in three parts: Server Side, Client Side and the proxy Collection.

The web mining allows for the collection of web access information for web page. This usage data provides the paths leading to accessed web pages. Web link prediction is the process to predict the web pages visited by the user based on the previously visited by the other users. Web pages may not only contain textual content but also other types of web data which may also audio files, video files and images. In the web navigation mining is present to improve the web services by removing useful data and knowledge from the web information.

2. RELATED WORK

In the Paper the web server log documents information are utilized for research work of web access forecast. The data gave by the information sources can be utilized to develop a few information to be specific clients, online visits, click-streams and server sessions. This strategy used to group comparable moves conduct to enhance the effectiveness of prediction. Forecast procedures were connected utilizing every bunch and utilizing the all information set. [1]

Preprocessing of web log record is the fundamental stride for web use mining. Cleaned Data in the wake of preprocessing is the base of example mining and example examination. Some preprocessing procedures are connected to enhance the nature of preprocessed information [2].

In the examination, information cleaning is performed by expelling the mistake demands and picture demand. A session has same IP address with as far as possible 30 minutes between back to back solicitations [5].

Successive access conduct for the clients can be utilized to enhance the execution of future access and the prefetching of much of the time page got to enhance the inertness time [8]. We realize that site development is constantly changed, numerous Prediction capability need not consider the conduct of recurrence but rather the site structure to dig web route designs for route expectation, dynamic mining methodology depends on the past mining comes about and framed new examples of the web information.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

3. PROBLEM DEFINITION

Due to their browsing patterns have to be extracted from web server log files. Web page prediction is the web usage mining by performing preprocessing of the data from a web site. One of the effectives of web caching has been reduced the network traffic, thereby minimize the client access latency, but it drawback is that it stores the pages without any prior knowledge. When the Big data log files are loaded in web page then its takes more time to clean the data. In the web page, the log data are not cleaned properly then the results are not given accurate.

4. METHODOLOGY

In the Mining, Web page prediction proposes a bracing approach for increasing the web server performance by analyze the user behavior. Prefetching and Prediction is done by the preprocessed to access the user log and it integrate into methods i.e. cleaning, Fuzzy c means clustering and Association rule to achieve the accurate web page prediction. In our work we use clustering technique proposed along with the association rule to give the better result.

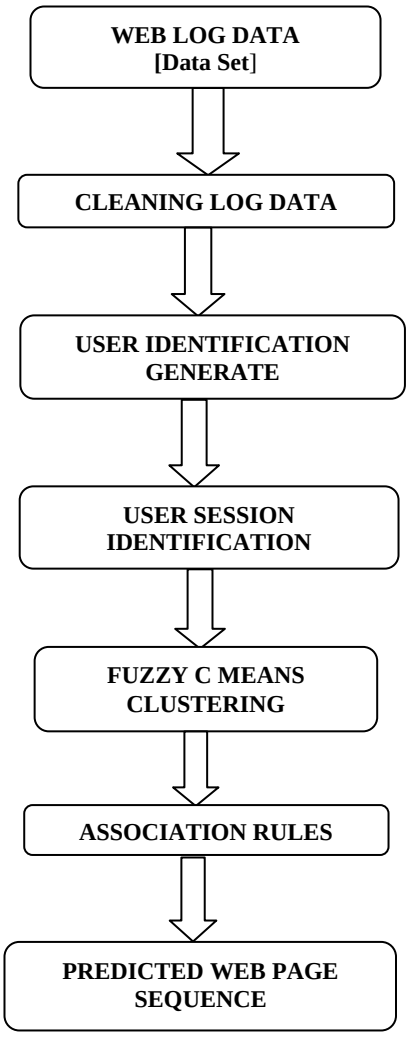


Figure 1: Proposed Model

WEB LOG DATA: - The logs files are maintained by the web server by the activity of the client who access the web server for a web site through the browser.

CLEANING LOG DATA: - In the log data, the entries in the log file for the unwanted view of image, graphics and multimedia by the users are removed.

USER IDENTIFICATION: - It is used to identify between the user and system to access it.

SESSION IDENTIFICATION: - It splits all the pages accessed by a user into the different session.

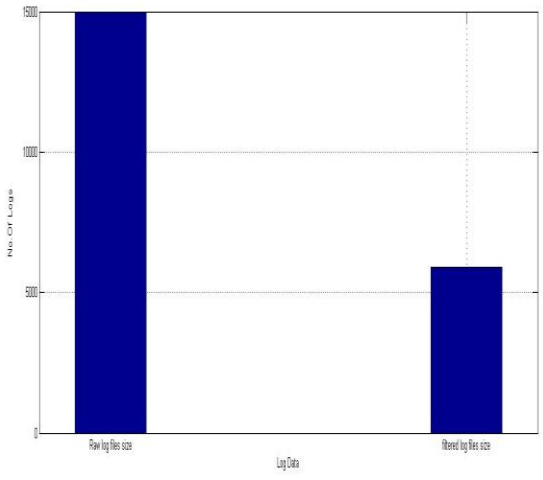
FUZZY C MEANS CLUSTERING:- It means a data clustering technique in which a dataset is grouped into n cluster with every data point in the data set belonging to every cluster.

ASSOCIATION RULES: - Association rule generation can be used to relate pages together in a single server session. The association rule may also serve as a heuristic for prefetching documents in order to reduce user perceived latency when loading a page from a remote site.

5. RESULT

The web mining process & Proposed Fuzzy c means technique becomes a major guide line of project implementation.
An experiment is performed on the web log data of NASA web log files. When an experiment is start, web site consists of 321 unique web pages. The web server log file data is collected date from July 1995 has been selected for further analysis.
There is total 5891 Original web log file size and total 606 valid records after filtered the data cleaning.
The server log files are :
204.31.113.138 - - [03/Jul/1996:06:56:12 -0800]
"GET /PowerBuilder/Compny3.htm HTTP/1.0" 200 593

Comparison between Initial and Filtered Log



Different Log Data

Figure 2: Comparison of Data Cleaning Before and After

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

The number of records resulted after cleaning phase is 606 and it is represent in Fig.2.

Record in original web log file Size	5891
Record in filtered log file Size	606
Number of Unique users Identified	214
Number of Different Unique Web Page	321

Fig.3 shows that the session length distribution of the web log dataset after the session is identified. The vertical axis stands the number of session length. The horizontal axis marked the length of session. In this figure shows that the each session length having many consecutive requests on the server.

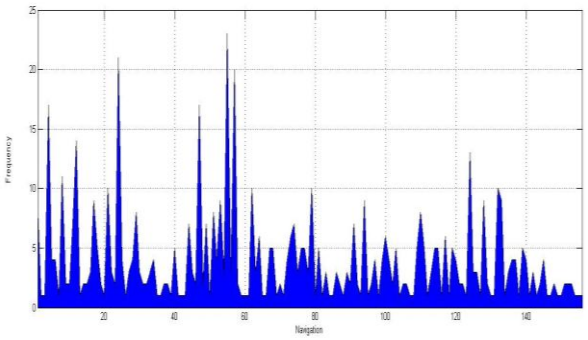


Figure 3: Graph shows the Frequency of each unique Navigation

Fig.4 shows the relation between the frequency of each page visited by the user and the page ID. Each web page has different number of times accessed by the user. Each web page is not the same and users always have similar transactions. Thus, the access history of user can be used to use for mining user access patterns.

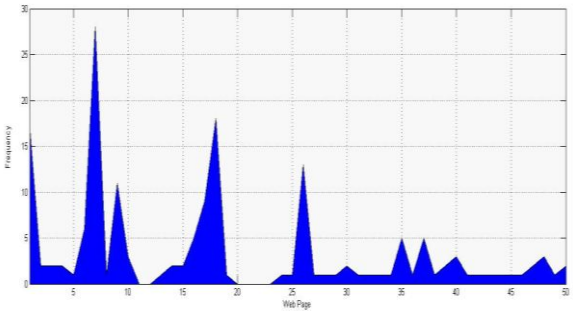
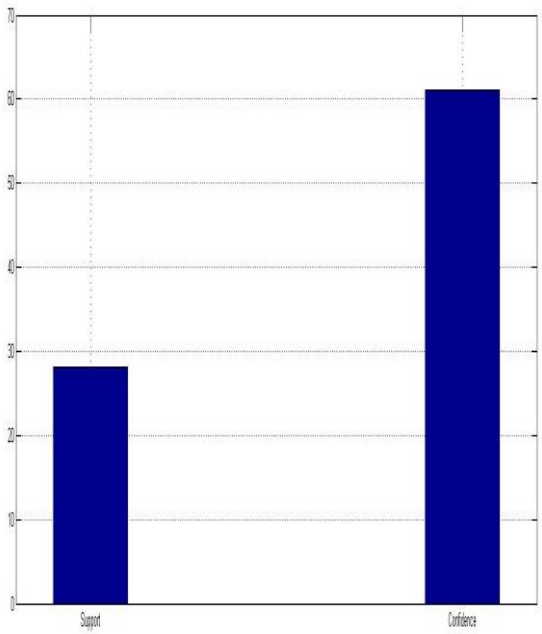


Figure 4: Graph shows Frequency of each unique web page.

For the test session, the corresponding cluster is chose and that cluster the web page is predicted which has the highest probability. In these, we use support and confidence in association rule:

Support (X) = Number of X transactions / Total Transactions
And Confidence (X, Y) = Support (X U Y)/ Support (X)
Where Support (XUY) means that the support of the union of the items in X & Y

Frequency Charts



Prediction Method
Figure 5: Accuracy of Web Page Prediction

6. CONCLUSION & FUTURE WORK

The problem of predicting user's access to the web server Web Page Perfecting has been widely used to reduce the user access latency problem on the internet; its Success mainly relies on the accuracy of web page prediction. Web usage mining allows the collection of web access information for the web page prediction. Data mining techniques like association rules, Sequential patterns, clustering can be used to discover the exact patterns.

The result of mining can be used to improve the website design and increase satisfaction which helps in different applications. My research work improve of site page access prediction accuracy by combing both Association rules and with use FCM calculation and Association rules and Fuzzy Clustering calculation can work together and give better web prediction results without compromise with accuracy. In Web page prediction, the next action corresponds to predicting the next page to be visited. The previous actions correspond to the previous pages that have already been visited.

The Future Scope is there are number of issues in preprocessing of log data. Analyzing web user access log files helps to improve the design of web components and web applications. Log includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time.
So Cleaning is done to speed up analysis as it reduces the number of records and increase the quality of the results in the analysis stage. More research can be done in preprocessing stages to clean raw log data and to identify users and to construct accurate sessions.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

REFERENCES

[1] Priyanka S.Panchal, Urmi D.Agravat "HYBRID TECHNIQUE FOR USER'S WEB PAGE ACCESS PREDICTION BASED ON MARKOV MODEL" IEEE, 4th ICCCNT 2013, Tiruchengode.

[2] Smriti Pandya, Rajesh Nigam "REVIEW PAPER ON WEB PAGE PREDICTION USING DATA MINING" IISTE, Vol.6, 2015.

[3] Prasad J. Koyande, Kavita P. Shirsat "COMPARISON OF DIFFERENT NAVIGATION PREDICTION TECHNIQUES" IJCSIT, Vol.6 (2), 2015

[4] Arshi Shamsi, Rahul Nayak, Pankaj Pratap Singh, Mahesh Kumar Tiwari, "WEB USAGE MIMNING BY DATA PREPROCESSING".IJCST Vol.3,Issue 1,Jan.-March 2012.

[5] Poornalatha G, Prakash S Raghavendra, "WEB PAGE PREDICTION BY CLUSTERING AND INTEGRATE DISTANCE MEASURES" IEEE/ ACM Trans. Syst., Man, Cybern. A syst., Humans, Volume 44,no.2 Sep 2012.

[6] Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan and Parinya Sanguansat, "DATA PREPROCESSING ON WEB SERVER LOG FILES FOR MINING USERS ACCESS PATTERNS", International Journal of Research and reviews in Wireless Communication (IJRRWC) Vol.2 No.2, June 2012.

[7] C.P.Sumathi, R.Padmaja Valli, T.Santhanam "AN OVERVIEW OF PREPROCESSING OF WEB LOG FILES FOR WEB USAGE MINING" Journal of Theoretical and Applied Information Technology 31st December 2011.Vol.34 No.2 IC.P.

[8] Deepti Razdan "THE NEXT PAGE ACCESS PREDICTION USING MARKOV MODEL", IJECCT Volume 1 Issue 1, September 2011.

[9] L.K.Joshila Gracel, V.Maheswari, Dhinaharan Nagamalai, "ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING", International journal of Network Security & its Applications, Vol.3, No.1, January 2011.

[10] Navin Kumar Tyagi, A.K.Solanki and Manoj Wadhwa, "ANALYSIS OF SERVER LOG BY WEB USAGE MINING FOR WEBSITE IMPROVEMENT", IJCSI International journal of computer science, Vol.7, Issue 4, No 8, July 2010.

[11] Priyanka Makkar, Payal Gulati, Dr. A.K.Sharma. "A NOVEL APPROACH FOR PREDICTING USER BEHAVIOUR FOR IMPROVING WEB PERFORMANCE", IJCSE International Journal on computer science and engineering, Vol.2, No.4,2010.

[12] Srivasta,J., Cooley,R., Deshpande,M., and Tan P.N.(2000). "WEB USAGE MINING: DISCOVERY AND APPLICATION OF WEB USAGE PATTERN FROM WEB DATA". Department of Computer Science and Engineering, University of Minnesota.

[13] Kosala ,R., Blockeel,H. (2000)."WEB MINING RESEARCH: A SURVEY." ACM SIGKDD (special Interest Group on Knowledge Discovery and Data Mining) Explorations. June, (2:1). Pp 1-10.