International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Efficient Approach for Web Search Personalization in User Behavior Supported Web Server Log Files Using Web Usage Mining

A Swapna, K Gurnadha Guptha, K Geetha

Department of Computer Science and Engineering, Sri Indu College of Engineering and Technology, JNTU Hyderabad, Hyderabad, India

ABSTRACT

In the present word web is the colossal capacity of data and it will continue expanding with the developing of web innovations. However, the person ability to peruse, get to and comprehend content does not increment with that string. Henceforth it ends up plainly complex to site proprietors to introduce appropriate data to the clients. This prompted give customized web administrations to clients. One of the notable methodologies in giving web personalization is Web Usage Mining. In this paper, our thought process of web use mining is to find clients' get to examples of site pages naturally and rapidly from the immense server get to log records, for example, often went to hyperlinks, much of the time got to site pages and clients gathering. Likewise, we proposed another strategy for finding clients' get to designs and prescribe it to the client.

Keywords : Web Usage mining; Web Intelligence; Web Personalization; log files ;Apriori algorithm; Web Log Cleaning Algorithm; Sessionization Algorithm.

I. INTRODUCTION

Web personalization is a generally new and troublesome field for web content conveyance. To satisfy desires of guests, clients, and faithful clients, web world is painted to offer astounding redid benefits all through their association with the framework. The effects of personalization and suggestion framework are frequently experienced by the fast prominence that this region has picked up inside the past couple of years. Clients ideally visit those sites that see their requirements, offer them quick esteem included modified administrations and basic access to required information in a basic graspable organization. Web personalization and suggestion framework assume a huge part in meeting this objective.

The corporate world looks towards the expansive volume of value-based and cooperation data produced by the web for R&D that encourages the formation of most recent imaginative focused administrations and items [1]. In today's e-business world, the greater part of the primary web-based business players have custom fitted the web personalization and suggestion framework including Yahoo!, Amazon, eBay, Netflix, News Weeder, IBM and a lot of something beyond.

The World Wide Web (WWW) is a colossal asset of numerous sorts of data in fluctuated groups which is extremely valuable for the examination of business advance, which is critical at this point days to remain in the opposition of business. Scientists are starting to examine human conduct in this appropriated Web information stockroom and are attempting to construct models for understanding human conduct in virtual situations. Information mining, frequently called Web mining when connected to the Internet, is a procedure of extricating shrouded prescient data and finding significant examples, profiles, and patterns from substantial databases. Web mining is an iterative procedure of finding information and is turned out to be a significant procedure for the understanding purchaser and business movement on the Web. There are three

subcategories for mining web data. These subcategories are

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

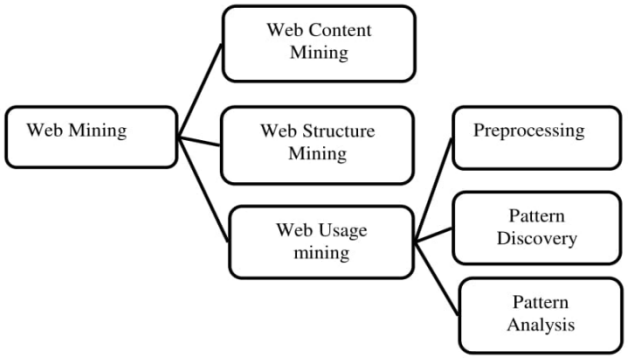

Figure 1: Web Data Mining Structure

Web personalization characterized in 2 subcategories initial one is active personalization, users expressly provide content to the system in order to get customized services/features. Once experiencing the advantages, users is also a lot of willing to submit info while not caring regarding its consequences. Moreover, second is passive personalization, the user is usually oblivious of what info is being occupied. Since personalization is achieved by suggests that of intensive info regarding users, thence privacy standards like P3P [2] should be used.

Web usage mining consists of 3 phases: preprocessing, pattern discovery, and pattern analysis [7]. The primary phase of net mining method is preprocessing, that in the main includes knowledge cleanup to remove unnecessary entries and data/fields from net access logs [8]. Successive phase is pattern discovery, which has strategies and algorithms developed from many fields like statistics, data mining, machine learning and pattern recognition. Generally, there are several data processing techniques notably for net personalization supported classification, clustering, sequential pattern mining, association rule discovery and Mark off models [9, 10]. Among them, sequential pattern mining technique is standard and wide used knowledge analysis technique in net usage mining. Pattern analysis is that the final innovate the net usage mining method. The aim of pattern analysis is to filtrate uninteresting rules or patterns found in the pattern discovery phase.

II. METHODS AND MATERIAL

Web Logs

At the point when a web client interfaces with the web and presents a demand, at that point his/her navigational data called as we get to log (now and then additionally called as weblogs in some writing) is put away in a web log record. The three unique wellsprings of web log document are web servers, intermediary servers, and customer programs [8]. We have utilized intermediary server logs to do tests for customized suggestions. Taking after is the example section from the intermediary server having squid joined web log design:

192.168.80.26 - - [05/Sep/2014:17:21:30 +0530] "GET http://www.excel-easy.com/vba.html HTTP/1.1" 200 3808 “http://www.google.co.in/url?sa=t&rct=j&q=macro%20in%20excel&source=web&cd=1&cad=rja&uact=8&sqi=2&ved=0CCcQFjAA&url=http%3A%2F%2Fwww.excel-easy.com%2Fvba.html&ei=saQJVImgPMqOuASLyYLICg&usg=AFQjCNFEZeyEk7sF_jOZdYU826TIN__d5g&bvm=bv.74649129,d.c2E" "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0)" TCP_MISS:DIRECT

The entry reflects the information as follow:-

- Remote information processing address: it's the ip address of client’s machine (host address).
- Username: it's denoted by “- - “. it's relevance only accessing password protected content.
- Timestamp: Date and time of client’s request.
- Access request: The request created by the consumer. Here, it's a “GET” request for the file “http://www.excel-easy.com/vba.html” mistreatment “HTTP/1.1” protocol.
- Status code: The ensuing standing code, e.g. two hundred denotes the success.
- Bytes transferred: a variety of bytes (e.g. 3808) transferred to the consumer.
- Referrer: it's the address of the previous page that coupled the user to this page.
- User Agent: It denotes the web browser and platform employed by the user.

In this paper, we tend to propose associate degree economical and novel design for web search personalization mistreatment net usage mining. It

doesn't want any express feedback from the user to find out his/her interest. The planned design contains varied modules that embrace preprocessing, Net Access Sequence (WAS) generation and user profile creation, Discovering fascinating usage patterns mistreatment planned economical serial access pattern mining algorithmic rule and at last module for customized recommendations. The planned algorithmic rule doesn't generate costlier WAP-tree at any stage and additionally it eliminates the necessity of projected information. this protects area and time. The new approach for sessionization ends up in the generation of correct frequent patterns. Once a similar user problems same/similar question, the system generates improved recommendations.

The rest of the paper is organized as follow. In section two we tend to introduce a study of connected work, Section three explains planned associate degree design, 6, the conclusion of the paper is mentioned

2. RELATED WORK

The focus of connected work to review and distinction the accessible technique to predict the net user behaviour.

Jagan and Rajagopalan [9] describe the net usage mining and algorithms used for providing personalization on the net. During this paper cantered the info pre-processing and pattern analysis on the web searching the association rule mining algorithms.

Ladekar A. Pawar A. et al. [10] describe an internet mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being hugely employed in net mining. The results obtained prove the hardiness of the algorithm proposed during this paper.

Parvatikar S. and Joshi B. [11] this paper cantered on net Usage Mining square measure the user navigation patterns and their use of net resources. The various stages concerned during this mining method and with the comparative analysis between the pattern discovery algorithms Apriori and FP-growth algorithm.
In the greater part of the endeavours by specialists, web use digging is utilized for web personalization on a specific site whose structure and substance is known ahead of time. In this paper, we concentrate on web seek personalization utilizing web use mining, where mining is connected on intermediary server logs to such an extent that every client will acquire customized proposals. The proposals are enhanced when a similar client fires the same/comparative inquiry.

3. PROPOSED WORK

Some researchers concentrate on the use of ontologies for customized web search. Hyperlink-based approaches have conjointly been used in literature. Some web search personalization analysis aims to enhance the initial page ranking algorithm. Some techniques use express feedback from a user concerning their preferences and interests [5, 17, and 18]. Some strategies area unit supported mapping a user question into a group of classes that represents user's search intention [2]. Several of the papers concentrate on personalization services for one web site. We propose architecture for web search personalization using net usage mining while not user's express feedback. It uses efficient data cleanup algorithm using java regular expressions, totally different approach for sessionization and efficient proposed consecutive access pattern mining algorithm. It recommends sites from one or a lot of websites counting on URLs in previous sessions, for a specific user.

As appeared in figure 2, the proposed engineering for web look personalization (utilizing web utilization mining), contains taking after primary strides:

Data gathering from intermediary server logs and information cleaning: Raw information (i.e. weblogs) of the intermediary server is cleaned by evacuating immaterial passages.

- User distinguishing proof: User is recognized through the mix of IP address and Agent.
- Sessionization: Different client's site page solicitations are portioned into the sessions.
- Web Access Sequence (WAS) era and client profile creation: Web get to succession is constructed and the conduct based client profile is made.
- Generation of continuous examples for customized suggestions, utilizing proposed mining calculation: Using the conduct based client profile and WAS, visit examples are created utilizing the proposed

consecutive get to example mining calculation. The proposed mining calculation depends on CSB mine.

3.1 Data Collection and Cleaning

The information cleaning, one of the real assignments in preprocessing stage incorporates the evacuation of unimportant passages and information/fields from web get to logs [8]. E.g. the passages that have the status of "mistake" or "disappointment" are evacuated. And, the solicitations with the augmentations .gif, .jpg , .jpeg, .JPG, GIF,.JPEG, .png , .cms , .css , .xbm , .robot.txt , .wav , . mpg , .promotions (from publicize servers), .swf are additionally evacuated. The sections which have a status code of 200 arrangement are considered as cleaned database. In web utilization mining process, effective information cleaning strategy is basic. This is accomplished utilizing "java customary expressions" which diminishes coding and runtime. The proposed web cleaning calculation is:

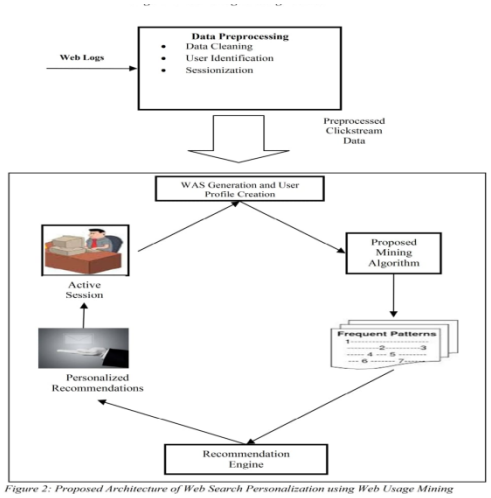

Figure 2: Proposed Architecture of Web Search Personalization using Web Usage Mining

Web Log Cleaning Algorithm

Input: Proxy server logs with squid combined log format

1) Start
2) Read the next weblog entry from web log file
3) Tokenize weblog entry using a regular expression as client IP, date, time, requested URL, web page status code, web page size returned and referrer.
4) If web page size is zero ignore the entry and go to step 2, otherwise
5) If returned status of the entry is other than in between 200-299, ignore the entry and go to step 2, otherwise
6) If entry refers to a page user is accessing is a multimedia object file like movie/sound /image etc., ignore it and go to step 2, otherwise
7) If referrer field is blank, go to step 2, otherwise
8) If referrer URL is any search engine (say Google) having a user query (with one or more words) save the separated fields of weblog entry in database with marking, Otherwise, Save the separated fields of weblog entry in database without any marking
9) End

3.2 User Identification

In proposed framework, it is important to recognize the distinctive clients for personalization. A client can be distinguished in view of IP locations: one IP deliver relates to one client. Yet, to be more precise, we distinguish the remarkable clients through mixes of IP address and the client specialist. So to recognize one of kind clients, a few tenets are utilized [9, 4, 8, and 19]:

- If IP address is new, it will be considered as a new user.
- If IP address is same but user agent is different, it will be considered as a new user.
- If IP address and user agent both are same, it will be considered as the same user.

3.3 Sessionization

Every client's movement on the web can be sectioned into various sessions. This procedure is called as sessionization. The point of sessionization is to discover different sessions for various clients [20]. The session is an arrangement of solicitations for site pages, made by a similar client over a specific route timeframe. The cleaned weblog database is utilized for session's creation. There are some normal standards to recognize client session [19, 21]:

- For a new user, a new session will be considered.
- For the same user if the referrer page of a requested page is null then a new session is assigned.
- If the time between page requests exceeds certain limit then it is considered as a new session. Generally, this time limit is 25.5 or 30 minutes. Many researchers use the timeout value of 30 minutes, the default timeout by Cooley.

In proposed system, a special approach has been used. If constant/similar question is given by the user on the

same date then requested address is supplementary to the previous session and it'll not be considered as a replacement session. I.e. it'll be thought-about as a continuation of the previous session. In our experimentations, it's found that this approach generates a lot of correct net Access Sequences (WAS), once a user searches for an extended time. The dotted rounded parallelogram in figure three shows a session. It's a collection of requests for numerous web content of various websites, created by constant user for same/similar question over a fundamental measure on constant date.

In the system, the timeout limit of twenty seven (average of 25.5 and thirty minutes) minutes is eliminated. within the experiments, it's found that once session timeout price of 27+ X, wherever X = some constant price e.g. 1, 2, 3…7, is applied to the weblogs, then several sessions square measure found to be quite this timeout limit. This leads to the formation of quite one session by constant user for the same/similar question. This can be because; by this timeout price it's tough to create a click stream linkage between last requested address among 27+X and initial requested address of next session accessed when 27+X minutes. As each sessions can generate separate WAS, the ensuing frequent patterns generated by projected mining formula won't be correct. Therefore within the system, we tend to use a special approach. If same/similar.

The question is given by the user on constant date then requested address is supplementary to the previous session. Thought-about it'll not be thought-about as a replacement session. In our experimentations, it's found that this forms less range of sessions and also the ensuing frequent patterns square measure correct.

Sessionization Algorithm

Input: Cleaned weblogs after user identification

1) Start
2) For each referrer field with special marking (search words)

i) Start of Session
ii) Find all Requested URL (RURL) of the same date
iii) For each RURL

Find all set of pages accessed recursively by considering RURL as referrer in next recursion

iv) End of Session
3) End

3.4 WAS Generation and User Profile Creation

Web Access Sequence (WAS) i.e. to discover set of pages got to together in time requested form. Every one of a kind website page is spoken to by a novel character and a web succession (string) is worked for all pages got to together in time requested form.

Client data is for the most part put away in two essential sorts of profiles: intrigue based client profile and conduct based client profile. The intrigue based profile can be portrayed by various models, for example, weighted vector show, chain of command sort demonstrate, weighted semantic net model and so on. The conduct based client profile can be characterized by the client conduct i.e. by putting away client's perusing designs [14]. The proposed framework utilizes conduct based client profile with alterations. This client profile stores IP address, client question, and client's perusing designs. The proposed WAS building and client profile creation calculation is:

WAS building and User Profile Creation Algorithm

Input: Preprocessed click stream database

1) Start
2) Get user query
3) Identify user
4) Remove stop words from user query
5) For each search word (SWORD) in user query

For each session containing the SWORD

i) Prepare WAS
ii) Add WAS, IP address and Query to WAS file

6) End

The WAS file generated in this step is used by the proposed mining algorithm as an input to generate frequent patterns.

3.5 Proposed Mining Algorithm

In Apriori-based calculations, it is important to examine the database different circumstances to acquire visit examples and dangerous applicant arrangements are produced particularly for countless. [12, 22, 11, 23]. Some example development based calculations require building anticipated databases while other example development based calculations include the development of various trees amid the total mining

process [12, 11, 24]. The CSB mining calculation does not create any hopeful succession, nor does it manufacture memory concentrated WAP-trees at any stage [1]. It creates sub-restrictive arrangement base to produce visit designs. Era of successive example stops when all produced arrangement can be converted to frame a solitary grouping for given support. We have considered execution time and memory utilization as the two measures for assessing the execution of consecutive example mining calculations [12]. So the execution of proposed calculation is enhanced by considering these two measures. Memory space is spared by maintaining a strategic distance from the utilization of any information structure to store the primary appearance of the (image may likewise be alluded as an occasion, for example, any page or URL) and killing the need of recursive digging for sub-CSB. Steps followed in the calculation keep away from development of anticipated database and utilization of conservative information structure brings about sparing memory space and time required for their operation. Toward the end, the created SAP information structure holds visit examples of all lengths.

Proposed Mining algorithm:

Input: WAS file

1) Start
2) Build Web Access Sequence set from WAS file
3) Calculate count of each symbol and store in Web Access Sequence set
4) Define Support value (Threshold)
5) Remove symbols with countless than support value from Web Access Sequence set
6) Build Unique Symbol list from Web Access Sequence set
7) Prepare Conditional Suffix table for each symbol of unique symbol list by linking unique symbol list elements with web access sequence set entries.
8) Generate Sequential Access Patterns starting with length 1 up to length ‘n’ with its count >= support value
9) End

In our framework, proposed mining calculation is connected after client fires a question and the created examples are utilized for suggestions without producing whenever devouring tree structure. There is no compelling reason to store the mining designs. At whatever point an inquiry is terminated, mining will be connected on related sessions and examples will be created for proposals live. i.e. without setting up any static model. So the framework creates all the more tweaked outcomes.

III. RESULTS AND DISCUSSION

Recommendation Engine

As shown in figure 3, recommendation engine generates personalized recommendations from frequent patterns generated by proposed mining algorithm. The patterns consist of numbers, where each number represents a unique page. Every unique page accessed by the user is assigned a unique page id

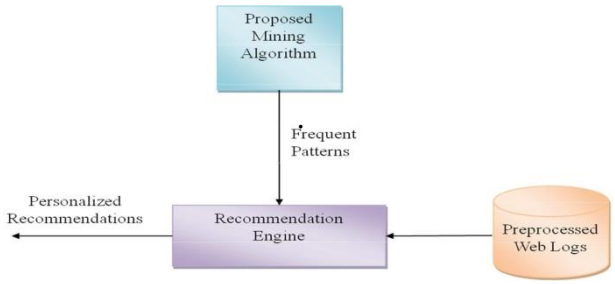

Figure 3: Recommendation Engine

Personalized Recommendation Algorithm:

Input: Frequent Patterns and Preprocessed click stream data

1) Start
2) For each frequent pattern For each frequent item
i. Find the corresponding unique web page
ii. Recommend the page
3) End

Each time a user submits a question , the user is 1st known by informatics and agent combination. Stop words from a question are removed after. Then previous user sessions are probe for similar words given in user question for a similar user. For this session WAS is ready and appended to WAS file. This WAS filing acts as an input to proposed mining rule. This rule generates frequent patterns that are used by suggestation engine to recommend URLs

IV.CONCLUSION

In this paper, we have proposed a proficient and novel engineering for web seek personalization utilizing web use mining, without client's express input. The engineering utilizes the proposed consecutive get to

example mining calculation. For execution assessment, the aftereffects of proposed mining calculation are contrasted and the consequences of CSB-mine calculation. Exploratory outcomes demonstrate that the proposed calculation performs superior to CSB-mine calculation. The outcomes demonstrate noteworthy change in normal memory utilization and furthermore change in the execution time. Additionally the exploratory outcomes demonstrate that, if sessionization is done without time restrain estimation of 30 minutes, more exact sessions (especially for longer seeking) are shaped. Utilizing these sessions, clients inquiry based conduct profile is construct. It is additionally demonstrated that, utilizing this profile the proposed mining calculation creates precise successive get to designs. At that point each time at whatever point a client issues the same/comparable inquiry.

V. REFERENCES

[1]. Baoyao Zhou, Siu Cheung Hui , Alvis Cheuk Ming Fong, "Efficient Sequential Access Pattern Mining for Web Recommendations",International Journal of Knowledge-Based and Intelligent Engineering Systems, ACM, Vol. 10Issue 2, April 2006, pp. 155-168.

[2]. Fang Liu, Clement Yu , Weiyi Meng, "Personalized Web Search For Improving Retrieval Effectiveness" , CIKM02, pp. 1-35

[3]. Tan B., Shen X., Zhai C., "Mining Long-term Search History to Improve Search Accuracy", Proceedings of KDD-06, 2006, pp. 718–723.

[4]. Eirinaki M., Vazirgiannis M., "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol. 3, No. 1, February 2003, pp. 1–27.

[5]. Sieg A., Mobasher B., Burke R, "Web Search Personalization with Ontological User Profiles", CIKM'07, ACM , Lisboa, PortugalNovember 6–8, 2007, pp. 525-534.

[6]. Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining", Communications ofthe ACM, Vol. 43, No. 8, August 2000, pp. 142-151.

[7]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, ACM SIGKDD, Vol. 1, Issue 2, Jan 2000pp. 12-23.

[8]. K. R. Suneetha, R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, Vol .9, No.4, April 2009, pp. 327-332.

[9]. Bamshad Mobasher, "Data Mining for Web Personalization" The Adaptive Web, LNCS 4321, Springer-VerlagBerlin Heidelberg,2007, pp. 90–135.

[10]. K. Suneetha , M. Usha Rani, "Performance Analysis of Web Page Recommendation Algorithm Based on Weighted Sequential Patterns and Markov Model", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No. 3, January 2013, pp. 250-257.

[11]. Jian Pei, Jiawei Han, Behzad Mortazavi-Asi, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proceeding of International Conference Data Engineering (ICDE 01), April 2001, pp. 215-224.

[12]. Nizar R. Mabroukeh and C.I. Ezeife, "A Taxonomy of Sequential pattern Mining Algorithms", ACM Computing Surveys, Vol. 43, No. 1, November 2010, pp. 3:1-3:41

[13]. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu, "Mining Sequential Patterns by Pattern-Growth:The PrefixSpan Approach", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, November 2004, pp. 1424-1440.

[14]. Cui Wei, Wu Sen, Zhang Yuan , Chen Lian-chang, "Algorithm of Mining Sequential Patterns for Web Personalization Services",The DATA BASE for Advances in Information Systems, Vol. 40, No. 2 ,May 2009, pp. 57-66.

[15]. R. Kousalya, V. Saravanan, "Personalizing User Directories Through Navigational Behavior of Interesting Groups and Achieving Mining Tasks", Journal of Theoretical and Applied Information Technology, Vol. 67, No. 2, 20 September 2014, pp. 321-333.

[16]. Zhicheng Dou, Ruihua Song, Ji-Rong Wen, "A Large-scale Evaluation and Analysis of Personalized Search Strategies", WWW 2007,

May 8-12, 2007, ACM 978-1-59593-654-7/07/0005, pp. 581-590.

[17]. Feng Qiu, Junghoo Cho, "Automatic Identification of User Interest for Personalized Search", Proceedings of the 15th International World Wide Web Conference, WWW 2006, Edinburgh, Scotland, May 2006, pp. 727-736.

[18]. Liu F., Yu C., Meng W., "Personalized Web Search for Improving Retrieval Effectiveness", IEEE Transactions on Knowledge and Data Engineering, 16(1), 2004, pp. 28-40.

[19]. R. Suguna, D. Sharmila, "User Interest Level Based Preprocessing Algorithms Using Web Usage Mining", International Journal onComputer Science and Engineering (IJCSE), Vol. 5, No. 09, Sep 2013, pp. 815-822.

[20]. K. Sudheer Reddy, G. Partha Saradhi Varma, M. Kantha Reddy, "An Effective Preprocessing Method for Web Usage Mining", International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014, pp. 412-415.

[21]. Maryam Jafari 1, Farzad Soleymani Sabzchi 2 , Amir Jalili Irani, "Applying Web Usage Mining Techniques to Design Effective Web Recommendation Systems: A Case Study", ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 2, No.8, March 2014, pp. 78-90.

[22]. Srikantaiah K. C., Krishna Kumar N., Venugopal K. R. and L. M. Patnaik, "Bidirectional Growth Based Mining and Cyclic Behavior Analysis of Web Sequential Patterns", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.3, No.2, March 2013, pp. 49-68.

[23]. Hengshan Wang, Chen Yang, Hua Zeng, "Design and Implementation of a Web Usage Mining Model Based on Fpgrowth and PrefixSpan", Communications of the IIMA, Vol. 6, Issue 2, 2006, pp. 71-86.

[24]. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl and Hua Zhu, "Mining Access Patterns Efficiently from Web Logs", Knowledge Discovery and Data Mining, Current Issues and New Applications, Lecture Notes in Computer Science, Vol. 1805, Springer, 2000, pp. 396-40