AN EFFICIENT AND NOVEL APPROACH FOR WEB SEARCH PERSONALIZATION USING WEB USAGE MINING

[1]KRISHNAKANT P. ADHIYA, [2]SATISH R. KOLHE
[1]Department of Computer Engineering, SSBT's College of Engineering & Technology, Bambhori, Jalgaon, Maharashtra (India)
[2]School of Computer Sciences, North Maharashtra University,
Jalgaon, Maharashtra (India)
Email: [1]kpadhiya@yahoo.com, [2]srkolhe2000@gmail.com

ABSTRACT

Generally each user has a different background and different needs when he/she searches for any information on the web. For the same query issued by different users, most of the search engines return the same results to all users, without considering the individuals need. This problem is solved using personalized web search. In this paper, we propose the architecture for web search personalization using web usage mining without user's explicit feedback. The proposed architecture uses an efficient sequential access pattern mining algorithm based on CSB-mine algorithm, a new approach for sessionization and modified user profile. The proposed mining algorithm does not generate any costlier tree and it avoids construction of projected database at any stage. Our experimental results on synthetic data set show that proposed algorithm performs better than the CSB-mine algorithm in terms of run time and average memory usage. Then experiments are carried out on proxy server logs for personalized recommendations, using the proposed architecture. The experimental results indicate that the system generates accurate sequential access patterns and improved recommendation of web pages.

Keywords: *Web Search Personalization, Web Usage Mining, Sequential Access Pattern.*

1. INTRODUCTION

The huge and continually increasing amount of information on the web creates many challenges for the researchers of web search [1, 2]. One of the problems of search engines is that when queries are issued by different users, most of them return the same results to the users without considering who have submitted the query. Thus, all users get the same result for the same query, even if they have widely different interests, needs and backgrounds. For example, a biologist or a medical student may fire the query "mouse" for rodent's information, while engineers may fire the same query for computer peripheral device "mouse". But both types of users get the same results from search engines.

This problem is solved using personalized web search and different search results are provided to the users considering their needs. Different types of personalization techniques and strategies have been developed [3]. The website can be customized as per the interest and need of user, which is one of the aspects of web personalization. The need of specific user is learned through the analysis of the user's surfing behavior on web and correlating it with other web data such as, structure, content and user profile data [4]. The aim of personalized web search is to present search results to a particular user, based on that user's interest and need [5].

The components of web personalization include [4, 6]:

- The preprocessing of web data such as content data, usage data, user profile data and structure data.
- Extracting statistical information and discovering interesting usage patterns using various data mining techniques.
- The actions to be carried out, recommended by such personalization systems.

Web mining, the application of data mining techniques is used to discover and extract information or knowledge from different web resources. Web mining is categorized mainly into three categories: web structure mining, web usage mining (i.e. web log mining) and web content mining. The web usage mining is a process of discovering the knowledge or patterns by applying various data mining algorithms on web usage data. The web usage data are stored as raw web logs at various locations such as client browsers, proxy servers or web servers [7, 8]. As shown in figure 1, web usage mining consists of three phases: preprocessing, pattern discovery, and pattern analysis [7].

The first phase of web mining process is preprocessing, which mainly includes data cleaning to remove unnecessary entries and data/fields from web access logs [8]. The next phase is pattern discovery, which includes methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Generally, there are many data mining techniques particularly for web personalization based on classification, clustering, sequential pattern mining, association rule discovery and Markov models [9, 10]. Among them, sequential pattern mining method is popular and widely used data analysis technique in web usage mining. Pattern analysis is the final phase in the web usage mining process. The aim of pattern analysis is to filter out uninteresting rules or patterns found in the pattern discovery phase**.**

Web Logs

When a web user interacts with the web and submits a request, then his/her navigational information called as web access log (sometimes also called as web logs in some literature) is stored in a web log file. The three different sources of web log file are: web servers, proxy servers, and client browsers [8]. We have used proxy server logs to carry out experiments for personalized recommendations. Following is the sample entry from the proxy server having squid combined web log format:

```
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=0CCcQFjA
A&url=http%3A%2F%2Fwww.excel-easy.com%2
Fvba.html&ei=saQJVImgPMqOuASLyYLICg&us
g=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 IP address: It is the IP address of client’s machine (host address).
- Username: It is denoted by “- - “. It has relevance only when accessing password protected content.
- Timestamp: Date and time of client’s request.
- Access request: The request made by the client. Here, it is a “GET” request for the file “http://www.excel-easy.com/vba.html” using “HTTP/1.1” protocol.
- Status code: The resulting status code, e.g. 200 denotes the success.
- Bytes transferred: Number of bytes (e.g. 3808) transferred to the client.
- Referrer: It is the URL of previous page that linked the user with current page.
- User Agent: It denotes the web browser and platform used by the user.

In this paper, we propose an efficient and novel architecture for web search personalization using web usage mining. It does not need any explicit feedback from user to learn his/her interest. The proposed architecture contains various modules which includes preprocessing, Web Access Sequence (WAS) generation and user profile creation, discovering interesting usage patterns using proposed efficient sequential access pattern mining algorithm and finally module for personalized recommendations. The proposed algorithm does not generate costlier WAP tree at any stage and also it eliminates the need of projected database. This saves space and time**.** The new approach for sessionization results in generation of accurate frequent patterns. When the same user issues same/similar query, the system generates improved recommendations.

The rest of the paper is organized as follow. In section 2 we introduce a study of related work, section 3 explains proposed architecture, section 4 presents performance evaluation of proposed mining algorithm, section 5 presents experimental results for personalized recommendations and in section 6, conclusion of the paper is mentioned.

2. RELATED WORK

Mobasher [9], distinguished between personalization and customization systems. The author provided brief descriptions of conventional strategies for web personalization which includes collaborative, content-based and rule-based filtering systems. The author overviewed web personalization process including various phases of a data mining cycle such as preprocessing , pattern discovery, and finding out useful patterns or knowledge though pattern analysis. Finally, it is shown that various pattern discovery techniques including sequential pattern discovery can be effectively used in a web personalization system.

Jian Pei et. al. [11], proposed PrefixSpan (Prefix-projected Sequential Pattern mining) algorithm. The algorithm does not involve generation of candidate sequences at any stage of execution. But it suffers from the major effort of constructing a projected database. [11, 12, 13].

Baoyao Zhou et. al. [1], proposed Conditional Sequence Base mining (CSB-mine) algorithm. It is sequential access pattern mining algorithm. It does not generate any candidate sequence like prefix span algorithm. At any stage it does not build any memory intensive WAP-tree for storing web access sequences unlike WAP-mine algorithm. The first preprocessing step builds conditional sequence base. Event queues are built from conditional sequence base, which are used to generate sub-conditional sequence base recursively to obtain frequent patterns. Using CSB-mine algorithm, authors proposed a recommendation system which performs better than WAP-mine based system.

Fang Liu et. al. [2], proposed a technique which combines user profile and a general profile. The technique involves construction of user profile based on the search history and construction of general profile based on the ODP category hierarchy. The user's search intension is represented by mapping a user query into a set of categories by combining both these profiles. It is shown that the technique of combining both profiles produces more accurate result than using a single profile alone.

Bamshad Mobasher et. al. [6], proposed web personalization using web usage mining. Authors described and compared different web usage mining techniques, based on usage clustering, transaction clustering and association rule discovery. It is shown that all techniques are promising, with their own unique characteristics.

Magdalini Eirinaki et.al. [4], presented web personalization process using web mining, and emphasized particularly on web usage mining. Authors discussed web usage mining process along with data mining techniques which include sequential pattern discovery, clustering, association rules and classification.

Cui Wei et. al. [14] proposed a hybrid web personalization system based on clustering and contiguous sequential patterns. In the proposed framework the log files are clustered using self-organizing maps (SOM). Clusters identified in SOM represent the most interesting categories of web pages. Then for each cluster, contiguous sequential pattern (CSP) mining is used for further personalization. The proposed system is found to be in primary stage.

Jaideep Srivastava et. al. [7], described the various phases in web usage mining such as data preprocessing, pattern discovery and pattern analysis, in detail. Authors also described the dimensions and applications areas where web usage can be used.

R. Kousalya et. al. [15], provided techniques and mining tasks to personalize web directories using navigation patterns. Authors provided a metric to compute the similarity between two navigation patterns and a methodology to discover interesting groups. Mining tasks include the discovery of categories and the sequences of categories in the web directory. Experimental results show effectiveness of their approach.

Zhicheng Dou et. al. [16], presented a framework for evaluating the personalized strategies on a large scale. Authors presented the performance of click-based and profiles-based web search personalization strategies. The experimental results show that all strategies outperform the normal web search and proved that different queries must be handled in different manner. It is also shown that generally click-based methods perform better than profiles-based methods.

In most of the efforts by researchers, web usage mining is used for web personalization on a particular website whose structure and content is known in advance. In this paper we focus on web search personalization using web usage mining, where mining is applied on proxy server logs such that each user will obtain personalized recommendations. The recommendations are improved when same user fires the same/similar query.

3. PROPOSED ARCHITECTURE

Some researchers focus on the use of ontologies for personalized web search. Hyperlink-based approaches have also been used in literatures. Some web search personalization research aims to improve the original page ranking algorithm. Some techniques use explicit feedback from a user regarding their preferences and interests [5, 17, 18]. Some methods are based on mapping a user query into a set of categories which represents user's search intention [2]. Many of the papers focus on personalization services for single website.

We propose architecture for web search personalization using web usage mining without user's explicit feedback. It uses efficient data cleaning algorithm using java regular expressions, different approach for sessionization and efficient proposed sequential access pattern mining algorithm. It recommends web pages from one or more web sites depending on URLs in previous sessions, for a particular user.

As shown in figure 2, the proposed architecture for web search personalization (using web usage mining), contains following main steps:

- Data collection from proxy server logs and data cleaning: Raw data (i.e. web logs) of proxy server is cleaned by removing irrelevant entries.
- User identification: User is identified through the combination of IP address and Agent.
- Sessionization: Different user's web page requests are segmented into the sessions.
- Web Access Sequence (WAS) generation and user profile creation: Web access sequence is built and behavior based user profile is created.
- Generation of frequent patterns for personalized recommendations, using proposed mining algorithm: Using the behavior based user profile and WAS, frequent patterns are generated using the proposed sequential access pattern mining algorithm. The proposed mining algorithm is based on CSB mine.

3.1 Data Collection and Cleaning

The data cleaning, one of the major tasks in preprocessing phase includes the removal of irrelevant entries and data/fields from web access logs [8]. e.g. the entries that have status of "error" or "failure" are removed. As well as, the requests with the extensions .gif, .jpg , .jpeg, .JPG, .GIF , .JPEG, .png , .cms , .css , .xbm , .robot.txt , .wav , .mpg , .ads (from advertise servers), .swf are also removed. The entries which have status code of 200 series are considered as cleaned database. In web usage mining process, efficient data cleaning method is essential. This is achieved using "java regular expressions" which reduces coding and run time. The proposed web cleaning algorithm is as follow:

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 web log entry using regular expression as clientIP, 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 goto step 2, otherwise
5) If returned status of the entry is other than in between 200-299, ignore the entry and goto 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 goto step 2, otherwise
7) If referrer field is blank, goto 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 web log entry in database with marking,
   Otherwise,
   Save the separated fields of web log entry in database without any marking
9) End
```

3.2 User Identification

In proposed system, it is necessary to distinguish the different users for personalization. A user can be identified based on IP addresses: one IP address corresponds to one user. But to be more accurate, we identify the unique users through combinations of IP address and the user agent. So to identify unique users, some rules are used [9, 4, 8, 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 a same user.

3.3 Sessionization

Each user's activity on web can be segmented into different sessions. This process is called as sessionization. The aim of sessionization is to find out various sessions for different users [20]. The session is a set of requests for web pages, made by same user over a certain navigation period of time. The cleaned web log database is used for session's creation. There are some common rules to identify user 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, different approach has been used. If same/similar query is given by the user on same date then requested URL will be added to the previous session and it will not be considered as a new session. i.e. it will be considered as a continuation of previous session. In our experimentations it is found that this approach generates more accurate Web Access Sequences (WAS), when a user searches for longer time. The dotted rounded rectangle in figure 3 shows a session. It is a set of requests for various web pages of different websites, made by a same user for same/similar query over a time period on same date.

In the system, the timeout limit of 27 (average of 25.5 and 30 minutes) minutes is eliminated. In the experiments it is found that when session timeout value of 27+ X, where X = some constant value e.g. 1, 2, 3…7, is applied to the web logs, then many sessions are found to be more than this timeout limit. This results in formation of more than one session by same user for the same/similar query. This is because; by this timeout value it is difficult to form a clickstream linkage between last requested URL within 27+X and first requested URL of next session accessed after 27+X minutes. As both sessions will generate separate WAS, the resulting frequent patterns generated by proposed mining algorithm will not be accurate. So in the system, we use different approach. If same/similar query is given by the user on same date then requested URL will be added to the previous session. It will not be considered as a new session. In our experimentations it is found that this forms less number of sessions and the resulting frequent patterns are accurate.

Sessionization Algorithm

Input: Cleaned web logs 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

After sessionization, next step is to build Web Access Sequence (WAS) i.e. to find set of pages accessed together in time ordered fashion. Each unique web page is represented by a unique character and a web sequence (string) is built for all pages accessed together in time ordered fashion.

User information is generally stored into two basic types of profiles: interest-based user profile and behavior-based user profile. Interest-based profile can be described by different models such as weighted vector model, hierarchy type model, weighted semantic net model etc. Behavior-based user profile can be defined by the user behavior i.e. by storing user's browsing patterns [14]. The proposed system uses behavior based user profile with modifications. This user profile stores IP address, user query and user's browsing patterns. The proposed WAS building and user profile creation algorithm is as follow:

WAS building and User Profile Creation Algorithm

Input: Preprocessed clickstream 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 algorithms it is necessary to scan the database multiple times to obtain frequent patterns and explosive candidate sequences are generated specifically for large number of itemsets. [12, 22, 11, 23]. Some pattern-growth based algorithms require to construct projected databases while other pattern-growth based algorithms involve the construction of number of trees during the complete mining process [12, 11, 24]. The CSB mining algorithm does not generate any candidate sequence, nor does it build memory intensive WAP-trees at any stage [1]. It generates sub-conditional sequence base to generate frequent patterns. Generation of frequent pattern stops when all generated sequence can be merged to form a single sequence for given support. We have considered execution time and memory consumption as the two measures for evaluating performance of sequential pattern mining algorithms [12]. So the performance of proposed algorithm is improved by considering these two measures. Memory space is saved by avoiding use of any data structure to store first appearance of symbol (symbol may also be referred as event, such as any web page or URL) and eliminating the need of recursive mining for sub-CSB. Steps followed in the algorithm avoid construction of projected database and use of compact data structure results in saving memory space and time required for their operation. At the end, the generated SAP data structure holds frequent patterns 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 count less 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 upto length ‘n’ with
    its count >= support value
9)  End
```

In our system, proposed mining algorithm is applied after user fires a query and the generated patterns are used for recommendations without generating any time consuming tree structure. There is no need to store the mining patterns. Whenever a query is fired, mining will be applied on related sessions and patterns will be generated for recommendations live. i.e without preparing any static model. So the system generates more fine tuned results.

3.6 Recommendation Engine

As shown in figure 4, 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 user is assigned a unique page-id.

Personalized Recommendation Algorithm:

Input: Frequent Patterns and Preprocessed clickstream 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 query, the user is first identified by IP and agent combination. Stop words from a query are removed afterword. Then previous user sessions are searched for similar words given in user query for the same user. For this session WAS is prepared and appended to WAS file. This WAS file acts as an input to proposed mining algorithm. This algorithm generates frequent patterns which are used by recommendation engine to recommend URLs.

4. PERFORMANCE EVALUATION OF PROPOSED MINING ALGORITHM

For the performance evaluation, we have implemented proposed sequential access pattern mining algorithm and CSB-mine algorithm using Java NetBeans 6.9 M1. The results are then compared with the results obtained by executing the PrefixSpan algorithm. We have used the standard code of PrefixSpan algorithm available under GPL2 license. The PrefixSpan algorithm is proposed by Jian Pei et.al. [11]. For the experimentation, we used the machine with Pentium Dual Core processor running at 2.00 GHz, 3GB RAM and Windows 7 operating system. The data set T25i10D10K used for performance evaluation is a synthetic data set, which is publicly available by IBM Quest data mining project. It is found that the number of sequential access patterns generated by all three algorithms is same. This shows the correctness of our implementations. Table 1 shows the results of CSB-mine algorithm and table 2 shows results of proposed mining algorithm. The results show 45 to 50 percent improvement in average memory consumption and 10 to 15 percent improvement in the execution time.

Execution time also includes the time required for reading a session file, building a data structure and then generating final sequential access patterns. Memory consumption includes the memory required to store all data structures and then to store final sequential access patterns.

5. EXPERIMENTAL RESULTS FOR PERSONALIZED RECOMMENDATIONS

To achieve personalized recommendations and to perform all the steps of web usage mining process, web logs of proxy server is used. It has squid combined log format. The experimentation is carried out on two days logs. Table 3 shows the results of preprocessing phase, proposed mining algorithm and recommendation engine.

In preprocessing phase data cleaning, user identification and sessionization tasks are carried out on those two days logs. The web log also contains entries related to two queries issued by an engineer, from a machine with IP address = 192.168.80.26. The first query is “Mouse” and other similar one is “What is Mouse”. After preprocessing, next phase i.e WAS generation and user profile creation is carried out. As shown in table 3, the proposed mining algorithm generates frequent patterns which are used by recommendation engine to recommend URLs.

Now, the same user fired the query “What is Mouse”, on next day. The user is first identified by IP and agent combination. Stop words are removed from a query. So query is converted to “Mouse”. Then previous user sessions are searched for similar words given in user query for the same user. For this session WAS is prepared and appended to WAS file. This WAS file acts as an input to our proposed mining algorithm. The mining algorithm again generates modified frequent patterns which are used by recommendation engine to recommend URLs. Table 4 shows preprocessed data of three days, modified frequent patterns and more fine tuned (improved) recommendation of web pages.

In other scenario, we have considered different user with IP address = 192.168.80.27. He has different interest and needs. E.g. he is interested to get information about rodents with the query “Mouse”. All tasks of web usage mining process are carried out on four days logs. The results (like table 3) are shown in table 5, and the recommended pages are now different than those presented in table 3 and table 4. This proves the personalized recommendations.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an efficient and novel architecture for web search personalization using web usage mining, without user’s explicit feedback. The architecture uses the proposed sequential access pattern mining algorithm. For performance evaluation, the results of proposed mining algorithm are compared with the results of CSB-mine algorithm. Experimental results show that the proposed algorithm performs better than CSB-mine algorithm. The results show significant improvement in average memory consumption and also improvement in the execution time. Also the experimental results show that, if sessionization is carried out without time limit value of 30 minutes, more accurate sessions (particularly for longer searching) are formed. Using these sessions, users query based behavior profile is build. It is also shown that, using this profile the proposed mining algorithm generates accurate sequential access patterns. Then every time whenever a user issues the same/similar query, he/she obtains more fine tuned personalized recommendations from our system.

Thus, the web usage mining technique using proposed architecture is a promising approach for better web search personalization. However it requires the squid combined web log format to carry out the web usage mining process.

In future work, we are combining the other user's experiences who previously issued the same/similar query to improve personalized recommendations.

REFRENCES:

[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. 10 Issue 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, Portugal* November 6–8, 2007, pp. 525-534.

[6] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining", *Communications of the 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-Verlag Berlin 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 on*

Computer 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. 396-407.

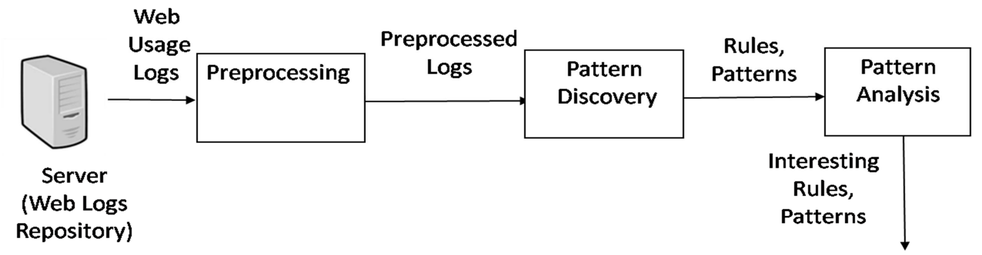

Figure 1: Web Usage Mining Process

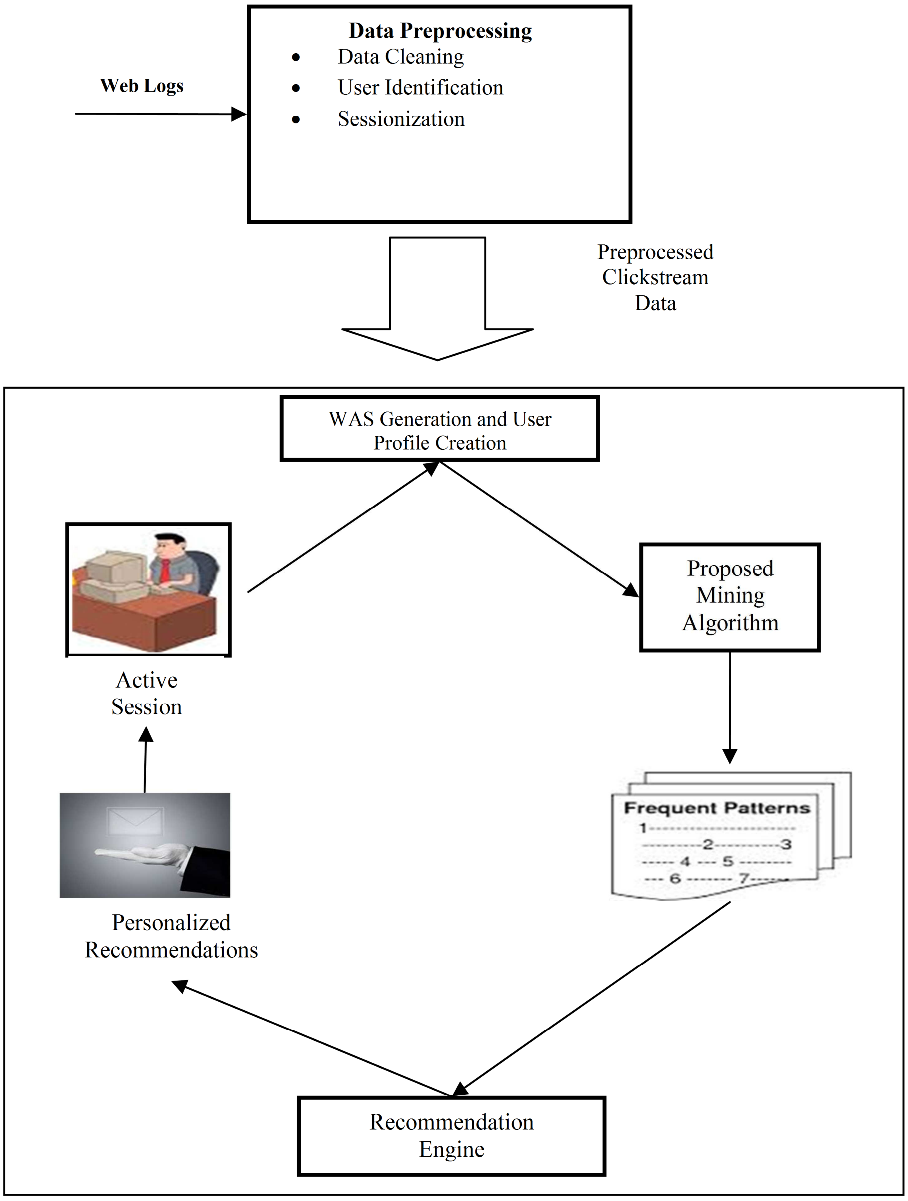

Figure 2: Proposed Architecture of Web Search Personalization using Web Usage Mining

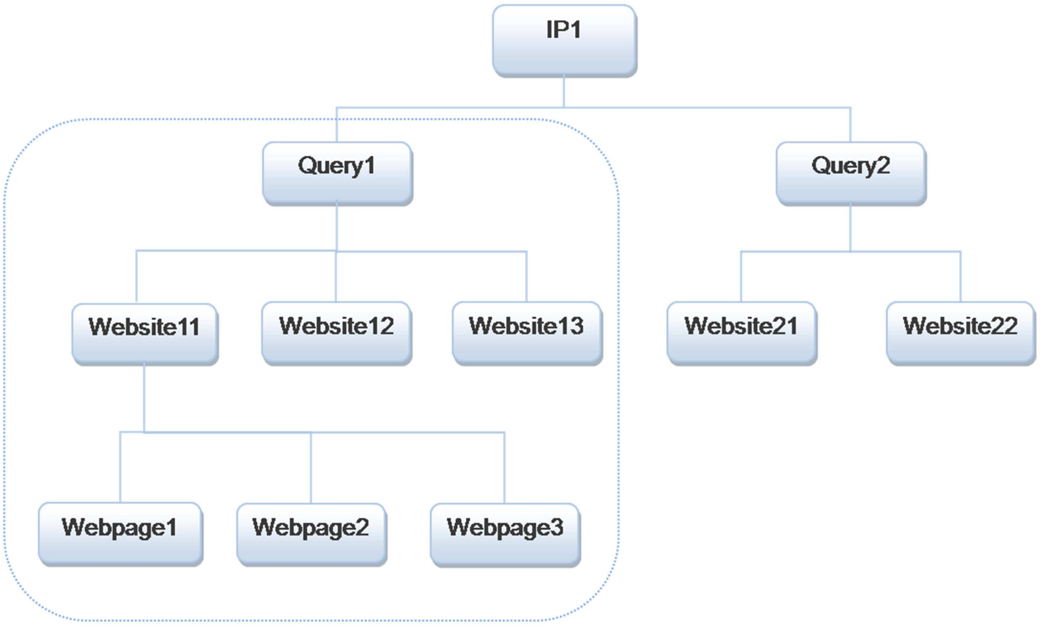

Figure 3: Session Creation

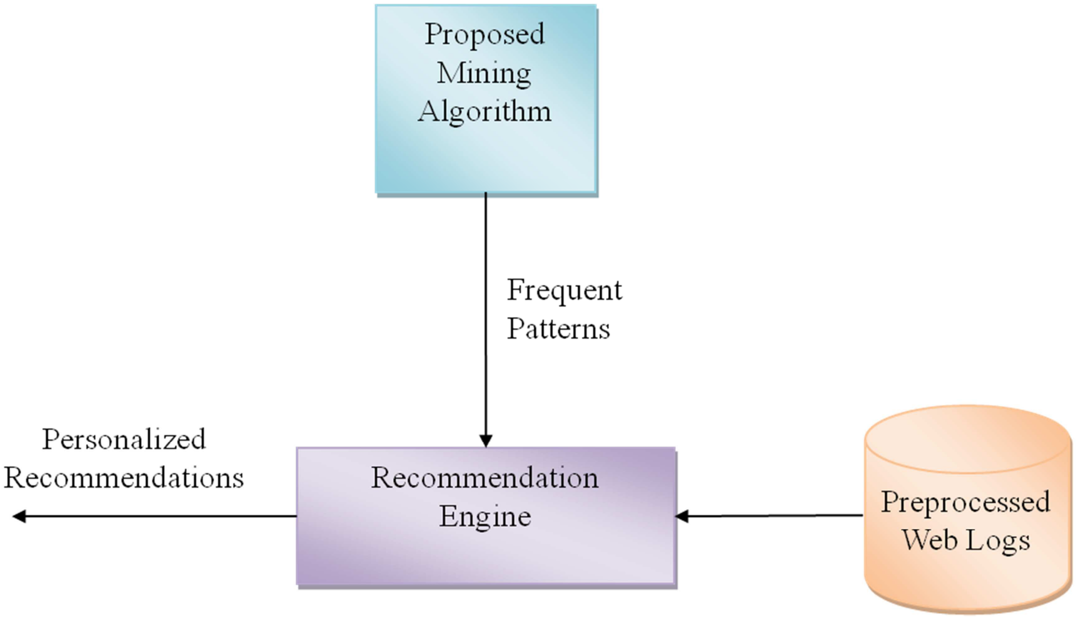

Figure 4: Recommendation Engine

Table 1: Results of CSB-mine algorithm

Support in %	Number of Sequential Patterns generated	Memory Consumption (in KB)	Execution Time (in Seconds)
1.20	1570	20203	687
1.40	1013	20139	361
1.60	771	20117	223
1.80	620	20098	153
2.00	533	20086	116

Table 2: Results of proposed mining algorithm

Support in %	Number of Sequential Patterns generated	Memory Consumption (in KB)	Execution Time (in Seconds)
1.20	1570	10167	636
1.40	1013	9901	351
1.60	771	9693	205
1.80	620	9548	121
2.00	533	9386	87

Table 3

Number of Days	02
Total Web Log Entries	39623
MinSupport	10 %
Cleaned Web Log Entries	2943
Total Unique Queries	14
Number of Useful Log Entries related to Unique Queries	111
Number of Unique Pages related to Unique Queries	55
IP Address	* 192.168.80.26
Queries Issued	Mouse What is Mouse
Frequent Sequential Access Patterns generated by Proposed Mining Algorithm	9 7 20 12 43 7 20 10 31 33 31 8 8
Recommended pages as per above SAP	platform.twitter.com/widgets/follow_button.html? odb.outbrain.com/utils/ping.html? platform.twitter.com/widgets/tweet_button.html? platform.twitter.com/widgets/tweet_button. ff7d9077a26377d36b6a53b1a95be617.en.html computer.howstuffworks.com/mouse2.htm odb.outbrain.com/utils/ping.html? platform.twitter.com/widgets/tweet_button.html? widgets.outbrain.com/nanoWidget/3rd/comScore/comScore.htm odb.outbrain.com/utils/get? www.engineersgarage.com/insight/how-ball-mouse-works odb.outbrain.com/utils/get? clients1.google.co.in/complete/search? clients1.google.co.in/complete/search?

Note: * Denotes the representative IP address and has been changed due to privacy reason.

Table 4

Number of Days	03
Total Web Log Entries	54027
MinSupport	10 %
Cleaned Web Log Entries	4458
Total Unique Queries	21
Number of Useful Log Entries related to Unique Queries	214
Number of Unique Pages related to Unique Queries	70
IP Address	* 192.168.80.26
Queries Issued	What is Mouse
Frequent Sequential Access Patterns generated by Proposed Mining Algorithm	43 10 12 7 19 43 12 33 7 31 32 8 8
Recommended pages as per above SAP	computer.howstuffworks.com/mouse2.htm widgets.outbrain.com/nanoWidget/3rd/comScore/comScore.htm platform.twitter.com/widgets/tweet_button.ff7d9077a26377d36b6a53b1a95be617.en.html odb.outbrain.com/utils/ping.html? t3.gstatic.com/shopping? computer.howstuffworks.com/mouse2.htm platform.twitter.com/widgets/tweet_button.ff7d9077a26377d36b6a53b1a95be617.en.html www.engineersgarage.com/insight/how-ball-mouse-works odb.outbrain.com/utils/ping.html? odb.outbrain.com/utils/get? static.ak.fbcdn.net/connect.php/js/FB.Share? clients1.google.co.in/complete/search? clients1.google.co.in/complete/search?

Table 5

Number of Days	04
Total Web Log Entries	98007
MinSupport	10 %
Cleaned Web Log Entries	8640
Total Unique Queries	33
Number of Useful Log Entries related to Unique Queries	427
Number of Unique Pages related to Unique Queries	86
IP Address	* 192.168.80.27
Queries Issued	Mouse
Frequent Sequential Access Patterns generated by Proposed Mining Algorithm	26 89 90 12 91
Recommended pages as per above SAP	fonts.googleapis.com/css? aax.amazon-adsystem.com/e/dtb/bid? www.orkin.com/rodents/house-mouse/ platform.twitter.com/widgets/tweet_button. ff7d9077a26377d36b6a53b1a95be617.en.html www.informatics.jax.org/