Biomedical Research 2016; Special Issue: S92-S98 *ISSN 0970-938X*
www.biomedres.info

A novel framework for an efficient online recommendation system using constraint based web usage mining techniques.

S Prince Mary[1*], E Baburaj[2]

[1]Department of Computer Science Engineering, Sathyabama University, Chennai, India

[2]Department of Computer Science Engineering, Sun Engineering College, Nagercoil, Tamil Nadu, India

Abstract

With the fleetly development of the internet, discovering useful knowledge from the World Wide Web became a censorious issue. With the huge volume of information present in the internet, user needs a help via recommendation system. From the user's log data lot of recommender systems developed to predict the user's next request when they view the web pages. However, each recommender system has its own advantages and drawbacks. A novel hybrid recommender system is proposed using a modified DBSCAN, modified prefix span algorithm and Genetic algorithm to mine the user's sequential navigational patterns, and then a hybrid recommendation model is proposed. The proposed Hybrid recommender system produces better prediction accuracy than the existing single recommender systems. Testing results recommend that the hybrid system is better in predicting the next request page of the web user.

Keywords: Web usage mining, DBSCAN, Modified prefix span algorithm, Genetic algorithm.

Accepted on August 03, 2016

Introduction

A well-known popular user interactive media is World Wide Web. WWW becomes an important information surfing source and provide various services to the users. Extracting knowledge from web data become more interesting research in the field of web mining. Web is the huge dynamic data source for mining knowledge form in it. Mining data using from web data is known as web mining Tug et al. [1]. Web mining is divided into three categories such as web structure mining, web content mining and web usage mining Kim et al. [2]. Mining knowledge from web usage data is known as web usage mining. There are different source available to provide web usage data. Web usage data is stored at server level, proxy level and cookies. The primitive data source used for web usage mining is server level stored data which is known as server access logs. Web site visitor activity is stored in a web log file. Web log files at server are automatically created. Different types of server log files are available. They are like access log, referrer log and Agent log file. Various formats of web log files such as W3C Extended log file format, NCSA common log file format, and IIS log file format. The data logged for each request is fixed in NCSA and IIS log file formats. In W3C format user has to log each request.

The problems that are faced by the data miner with the data collected at server side to discover sequential navigational patterns, first one is to identify the visitors and distinguish them. The complication is that some visitors when they use proxy servers or share the same system to browse the website. Therefore IP address is not alone sufficient to identify the visitor. Second, when the user presses the forward and backward buttons are not registered in the log file. Therefore, missing information must be considered. Visitor request for a page, more than one entry also registered in the server log. Need of duplicate data must be removed from the log file. Third, sessions are to be identified for each user within a period of time. Fourth, during a particular session time spent by the visitor must be calculated. Aforementioned problems are related to web usage mining preprocessing itself; there are problems when it comes to the mining navigational web pages and its applications like navigational pattern prediction process. To get the best accuracy in prediction, prediction should be done in a timely manner.

Knowledge from web log data can be mined using various techniques. Techniques like clustering and sequential pattern mining are used. Clustering is of two types like clustering pages and clustering users. When the pages are clustered it results the group of web pages which are visited many times by different users. Clustering the user results the group of visitors visiting the similar pages. DBSCAN algorithm used to generate the web page clusters. Mining successive subsequence's as web pages from a web log database is known as Sequential Pattern mining in web usage mining area. First sequential pattern mining problem was introduced by Agrawal and Srikant in [3]. Mining sequential patterns in web usage mining are to discover sequential navigation web pages that present in visitors sessions file frequently. The normal

sequential patterns like Wong and Pal [4]: the 80 % of the visitors, who first visited P1.html, then visited P2.html and P3.html in one session. Like association rules, sequential patterns are present syntactically. However, mined sequential patterns depend on the time order of the sequence. An efficient algorithm prefix span is used for mining frequent web pages Han et al. [5]. By an intermediate database generation, Prefix Span mines the frequent sequences. But traditional approaches generate candidate sequences. Prefix Span algorithm memory consumption is less than the traditional approaches [5].

Web users most of the complaint is about finding right information on web sites. To assist user in finding the useful information a recommender system is used. Recommender system helps the user to determine the next web page to be accessed in a web site. Here the goal to find out the next web page to accessed. Many recommender systems are available to assist the user to navigate. Web usage mining and web content mining techniques are used by the traditional recommender systems [6-8]. Currently web content mining techniques are applied for recommender system. But web content techniques are not able to handle the dynamic web pages, and constant changes in news web sites. Thus recommender model based on the web content leads model update frequently. Due to the above reason, here a web usage mining technique is applied for the construction of the navigational pattern prediction.

Recommender model performance depends on the structure of the web site; hence it could be complex to use a best single model for recommendation. Each model has its own merits and demerits. To improve the performance of a recommendation system, a hybrid model should is proposed. Burke [9] proposed a hybrid models based on the collaborative filtering techniques. The information from various resources is used in collaborative filtering. But collaborative filtering used the web site content based approach. Due to this approach to increase the accuracy of prediction model, a hybrid recommender system is generated by combining the recommendation models based on the web usage mining [10,11]. Proposed a navigational pattern prediction using KNN based algorithm and association rule mining. This paper explains the process of navigational pattern prediction via three steps. Initially data is collected from the server access log file and preprocessed using the preprocessing steps. Second, a modified density based clustering algorithm, a modified prefix span algorithm and genetic algorithm are applied on the preprocessed data to find navigational patterns. And finally, a hybrid prediction model proposed for an online recommender system.

Related Work

Identification of visitors and discrimination between each visitor is one of the tasks of pre-processing in WUM. If the web sites allows user to login before browsing starts it's easy to identify the visitors and easily visitors can be discriminated. The problem comes when the web sites permit the visitor to browse anonymously. In such case visitor discrimination is the challenging problem because server records the visitors request as a log file. If the server stores the visitors request as a common log format then user discrimination becomes more challenging problem. Pabarskaite and Raudys [10] suggested various ways to distinguish visitors are listed with its strong and weakness points. Visitors can be identified with cookies. When the user first time request a web sites, the cookie is send to the user along with the request by the server. Next time the same visitor request for the another resource form the same website, then the browser send the request to the server along with the cookie saved in the user computer. If the server receives request with the cookie then the server will identify the user. Server always cannot use this technique because the user may delete the cookies and send the request. Deleting the request with proxy and shared IP then also visitor can be identified from the same IP. But this technique will reduce the log size due to this analysis will not be performed efficiently.

To find the navigational patterns it's compulsory to know what user has viewed at each time when they have visited the website. Each user visits is known as session. It is a web page related series during a particular period. Session identification becomes the problem of navigational pattern mining. Dividing the sequence of user request into subsequence is the session identification. Pabarskaite and Raudys [10] identified the sessions based on the time gap between the consecutive requests. Catlege and Pitkow [12] used 30 min threshold in many commercial products. Many [5-7] researches used different threshold values from 10 min to 2 h. Pabarskaite and Raudys [10] used the threshold time duration is 25.5 m in Catlege and Pitkow [12] calculated the mean inactivity duration is 9.3 min in a website, and added 1.5 SD (Standard deviation) to the mean and got 25.5 min was obtained. The standard threshold used as a cutoff threshold value to identify the sessions. Pabarskaite and Raudys [10] mentioned many free and commercial websites used 30 min threshold value as a time gap for session identification. Chen et al. [13] Identified the sessions based on the forward sequences. Session file does not contain the backward sequences. A session sequence with no repeated pages in each sequence is known as forward sequence. A session sequence which containing the repeated web pages in each sequence is known as a backward sequence. Session identification from forward sequence is not a suitable way, because website visitors can press a back button as well as without visitors are forced to go back to the previous page if the page is a hub page of the website. Berendt et al. [14] have identified sessions by using referrer, but it gives poor performance.

Different types of analysis can be performed on web log data once it is pre-processed. With raw web log itself statistical analysis is done using log analysis tools, it is useful to the web server administrator for identifying the problems. It will also support for decision making Pabarskaite and Raudys [10]. However Li and Zhong [15] done more complex observations to extract knowledge using other mining techniques. In the data pre-processing step to identify the user and sessions, a time based heuristic is applied and technique is framed based on the shared patterns exists among the same user sessions. This new approach produces a best session sequence for the next step. Second, navigational pattern mining still divided into three

phases, such as identifying the pattern using a modified density based clustering approach, a modified prefix span algorithm and a genetic algorithm. In modified density based clustering approach page clustering is applied. It mines patterns without any user support. In association rule discovery and sequential pattern mining technique need input parameter such as minimum support. Density based clustering is used because it detects original patterns unlike other clustering algorithms. In WUM to find navigational patterns two types of clustering are applied. One is user based clustering and the other one is page based clustering. Perkowitz and Etzioni [9] proposed a PageGather clustering technique to cluster the pages based on cliques, but clustering pages of website into one page there is a loss in page data. Mobasher et al. [11] used k-means clustering algorithm to cluster users [3] used a density based clustering algorithm to cluster users in significant group. In the proposed navigational pattern mining approach density based clustering algorithm is used to cluster the pages visited by the users. Clustering approach is used because beforehand number of clusters needs not to be known. The advantage of this density based clustering algorithms is it can filter the noise.

In the second phase of mining navigational patterns a modified prefix span algorithm used. This algorithm mines the inter-session patterns in a time ordered sessions. These patterns can be used to find the relationship between the sequential visits in order to predict the future pages. Then last phase is applying the genetic algorithm to find navigational patterns. Tug et al. [1] Genetic algorithm based on the functionality of natural selection and genetics and GA is a probabilistic search method. Kim and Zhang [2] GA is a very effective global search to find solutions in non- deterministic problems. To find sequential access pages Genetic algorithm is used. After discovering visitor's sequential navigational web page sequence from the preprocessed session file, a recommendation models constructed using a combined single models. Different techniques of web usage mining applied to build an effective recommender system. One of the successful models is collaborative filtering based recommender model Resnick and Varian [16]. Collaborative filtering predicted the utility of the product of an active visitor by comparing the records of the active visitor against the similar records of the other users. The demerit of these techniques is prediction accuracy cannot be maintained when the number of items is more. The next hybrid techniques is content based, it works based on the item description. Balabonovic and Shoham [17] proposed a model to help the user to find a web site of their interest. Pazzani [18] proposed a system by combing the collaborative filtering, content based (product attribute) and customer attribute (demographic filtering). Abdelghani et al. [9] achieved an accurate prediction the pages users visited and the order of pages visited is used for this a generalized suffix trees used, then a binary matrix is constructed with the pages as column and sessions as rows.

System Design

The proposed system explain the methods for finding a quality sessions from the web log database, then to find the sequential navigational patterns using web page clustering algorithm, a modified prefix span algorithm and genetic algorithm is used based on the time and space constraints and then a recommendation model is proposed using the combination of various models. The proposed methods are tested using various datasets and the results are evaluated and discussed.

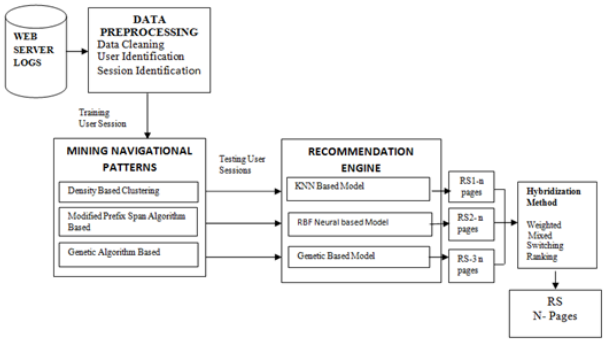

Figure 1. *Architecture of the proposed hybrid recommender system.*

Figure 1 Illustrate the proposed hybrid recommender system architecture. There are is an offline component and online component. Offline component performs the task of data pre-processing and mining the navigational patterns in three different ways and online components of recommender engine with various models and hybridization.

Off-line Component

Data pre-processing

The primary task of the preprocessing is to arrange the log data for retrieving the usage patterns. It is a time consuming task. It has three works; they are data cleaning, user identification, and session identification. These preprocessing steps are common for web usage mining techniques. In the first step irrelevant entries logged with the filename with the suffixes such as gif, GIF, jpeg, jpg, are removed. Similarly sound and video files and robot file also removed. Next, the log file is normalized by determining the different URLs which are syntactically similar. The time difference between the consecutive pages is known as the visiting time of the web pages. Based on the returning visitor request the sessions are identified. If any returning visitor visits a new page always, then those requests are not useful in finding navigational patterns.

Algorithm to find sessions:

begin

Arrange web pages according to user id and timestamp

for each web page sequence do

split the web page sequence by threshold

if the web page sequence is single

add the web page sequence to session list

else

if there is a shared pattern between the subsequence

add subsequence to the session list

else

skip the web page sequence

end if

end for

end

Navigational pattern mining: The identified sessions are used as input to the navigational pattern mining algorithms. A database of visitors transactions are used to find navigational patterns of each user. From the database among all the webpage sequences maximal occurrence of sequential patterns are mined. This is useful for website administrators to find the path and to insert the advertisement banners.

Density based clustering: Density based clustering algorithm is used to frame clusters of web pages and errors in the database. Two attributes to be known for every cluster. That is Eps and MinPts among these two attributes at least one point to be known initially. From the known point, find nearest points that are density wise nearer. Definition 1 is used to discover clusters [9].

Definition 1: (cluster) Let L be the set of web pages. A cluster CL is a non- empty subset of L satisfying the following conditions:

- a, b : if a ϵ C and b is density- reachable from a wrt. Eps and MinPts, then b ϵ C (Max)
- a,b ϵ C : a is density –connected to b with respect to Eps and MinPts.(Connectivity)

Once the found sessions are grouped, then the input attribute found for density based clustering. The input attributes is found via OPTICS [11]. OPTICS scans the found sessions as input and produces the output with respect to the web page to other web pages. The found sessions are framed as a matrix form; in the matrix columns are the number of web pages in the website and rows is the number of sessions. There are two types of matrix representations used. They are binary and non-binary matrix. Table 1 shows the binary matrix, each cell says that during a specific session, whether a specific web page is visited or not, 0 for NO and 1 for YES. In a non-binary matrix, each cell says how duration of a particular page is visited in a specific session shown in Table 2.

Table 1. *Binary matrix.*

VP1	VP2	VP3
0	1	1
1	0	0
0	0	1

Table 2. *Non-binary matrix.*

VP1	VP2	VP3
2	5	10
6	7	3
9	8	1

Sequential pattern mining

Modified p-prefix span approach: The most useful pattern growth algorithm P-Prefix Span algorithm is based on a probability value of inter-arrival time where frequent patterns are found. Nevertheless, frequent incidence of sequential pattern increases the complexity to predict the navigational patterns.

The proposed research work introduces a novel approach, namely Modified Prefix Span pattern mining algorithm to determine the issues in mining. Based on various user specified constraints, the proposed mining algorithm extracts the feasible patterns from the sequential web log file storage. At first, based on user specified maxima and minima, space constraints, the hybrid pattern mining algorithm will divide the sequential log file storage. Once the sequential log file storage is divided, the hybrid pattern mining algorithm can discover frequent sequential patterns with inter-arrival time probability of consecutive items. With the measured inter-arrival time probability, the proposed technique can be employed to recalculate the frequent patterns. Given a group of patterns, the goal of web pattern mining will discover each frequent pattern with the probability of a predefined time-period length greater than smallest probability threshold value. For searching a sequential pattern, the probability value should be checked when a frequent query is to be appended to a sequential pattern.

A sequence is expressed as $[s_1,t_1]$, $[s_2,t_2]$, …, $[s_n, t_n]$, where s_j stands for an item and t_j represents the requested time period at which s_j occurs, $t_1 \leq t_2 \leq \cdots \leq t_n$ and $1 \leq j \leq n$ (ω)=the sequential pattern(ω_1, ω_2, … ω_n). The following terms are defined based on the above representation.

Definition 1: A sequence S={$[s_1, t_{s1}]$, $[s_2, t_{s2}]$… $[s_n, t_{sn}]$}, and a page pattern A={a_1,a_2,....a_m), where 'a' is the time-rest prefix of S and $m \leq n$.

Definition 2: A web log storage WS={S_1, S_2, …, S_∞} i.e. a sequential web log storage having infinite sequence patterns because of it every time a set of queries occurs.

Definition 3: A time-rest subsequence (A′) of sequence (A) with respect to the R. The Time-rest prefix is described. If the sequence A is called a projection of A with respect to the time-rest prefix R if A′ has time-rest prefix R and there exists no proper time-rest super sequence A″ of A′ such that A′ is a time-rest subsequence of A and has a time- rest prefix R. If one removes directly the time- rest prefix R from the projection A′,

the new sequence obtained is called the postfix of A with respect to time-rest prefix R.

Definition 4: Let TLast denote the occurrence time of the last query in a web log storage WS. S={[s_1, t_{s1}], [s_2, t_{s2}] . . . [s_n, t_{sn}]} is a sequence in WS. {S(t_1,t_2)} represents all the queries that exist in S between transaction time t_1 and t_2. Assume that A = {[a_1, t_{a1}], [a_2, t_{a2}], ..., [a_m, t_{am}]} is a time-rest subsequence of S, and χ is a query which does not belong to {S (t_{am},t_n)}. According to the current sequence S, query occur time of χ is not known. The consecutive queries probability of time interval and for for query χ the potential censoring time with respect to A is defined as T^{Last}-t_{am}.

Steps of modified p-prefix span approach

Step 1: Divide the persistent web log storage using min and max threshold values.

Step 2: Read the persistent web log storage to extract repeated queries

Step 3: Calculate arrival rate among each and every query

Step 4: Calculate inter-arrival time probability for repeated queries at time t. (how will you calculate)

Step 5: If inter-arrival probability is higher than the threshold, update the repeated query into a pattern.

Step 6: Repeat steps-2 to 5, till obtaining the specified reliable patterns

Steps of arrival rate function

Arrival Rate(<ω>, WS|<ω>,χ)

Input: <ω>, WS|<ω>,χ

Output: δ-arrival rate of query χ occurrence after last query in <ω>

tq → the transaction time of last query in <ω>, α1=0; α2=0; r=0;

r-total number of subsequence

α_1- Difference among query time of last time in (ω) and first query χ,

α_2- the time of input query χ

for every pattern in WS|<ω>

if χ ∧ queries in pattern

$\alpha_2 \rightarrow \alpha_2$+ (Query time of last occurrence query-Query time of first occurrence query)

Otherwise r → r+1; t_χ →query time of query χ

$\alpha_1 \rightarrow \alpha 1$+($t_\chi$- query time of initial occurrence of next query)

end for

$\delta \rightarrow r/(\alpha_1+\alpha_2)$

return δ;

The sequential patterns are mined using a harmonized prefix span algorithm, and then these patterns are applied to train the network for navigational pattern prediction.

Genetic algorithm: Genetic uses three operators such as data selection, crossover and mutation [1]. The successor of the GA is generated using the operators. Crossover use two web users pages and a new offspring is generated for mutation operation based on a single point crossover.

Algorithm:

Begin

Initialize population

Compute fitness

Compute Threshold

Repeat

Perform Crossover

Compute fitness

Apply Mutation

Compute fitness

Until Constant Tree is obtained

End

Fitness calculation: Fitness function is calculated for every user in the population.

Fitness function F is the sum of three factors.

F=f1+f2+f3

f1, f2, f3 are the proposed fitness factors. Figure 2 shows the sample fitness function evaluation tree structure.

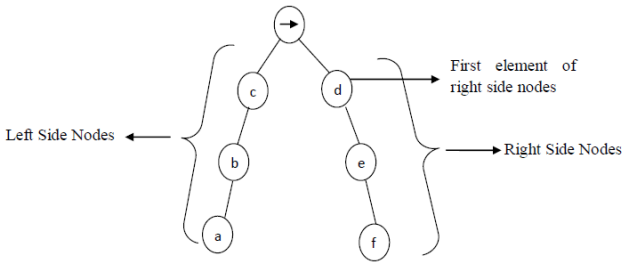

Figure 2. *Initial population.*

$$f_1 = \frac{f(LN+RN)}{TD},\ f_2 = \frac{f(LN+RN)}{f(LN)},\ f_3 = \frac{f(LN)}{f(LN+first\ element\ of\ RN)}$$

Experiments and Results

The proposed navigational patterns are mined using two techniques such as clustering and sequential pattern mining. The DBSCAN algorithm generates the page clusters. Each page cluster gives the most frequently accessed pages by each user. Modified prefix span algorithm generate the frequently

accessed patterns and genetic algorithm also generate the sequential web pages. These techniques are tested using a web log data set collected from NASA web site for a month of August 1999.

Preprocessing

Once the data set is cleaned by removing image files, robots and sound files, then user identification and heuristic session identification is done based on the threshold time. Table 3 shows the sample data set after performing data cleaning. Table 4 shows the number of record identified before data cleaning and after data cleaning.

Table 3. *Sample data set after data cleaning.*

153.19.130.21	23/Aug/1995:04:35:52 -0400	HEAD /shuttle/missions/missions.html HTTP/1.0	200 0
128.39.105.38	23/Aug/1995:04:37:28 -0400	GET /shuttle/missions/missions.html HTTP/1.0	200 8677
168.126.93.101	23/Aug/1995:04:38:32 -0400	GET /shuttle/missions/sts-68/mission-sts-68.html HTTP/1.0	200 49705
160.29.73.222	23/Aug/1995:04:38:46 -0400	GET /history/apollo/apollo-13/apollo-13.html HTTP/1.0	304 0
160.29.73.222	23/Aug/1995:04:39:58 -0400	GET /shuttle/missions/sts-71/movies/movies.html HTTP/1.0	304 0

Table 4. *User identification.*

Data Set	Period	Number Of entries	Number of Data After Pre-Processing	Number of users Identified
NASA Web Log	Month of August-1995	10,48,576	93980	56385

Session identification

Table 5 shows the number of sessions identified for classical time out sessions. Table 6 shows the shared time out based sessions with 30 mins duration. Table 7 shows the shared time out based sessions with 10 mins duration. Figure 3 shows the comparison between proposed algorithm results. It shows that the number of identified navigational web pages is more in DBSCAN algorithm than the other two techniques.

Table 5. *Classical time out based.*

Threshold	Precisely Identified Sessions
30 mins	17,350
10 mins	19,220

Table 6. *Proposed session identification: time-out =30 mins.*

%	Shared pattern ≥ 3	Shared pattern ≥ 2	Shared pattern ≥ 1
100	7420	7708	10,984
90	7420	7735	11,345
80	7420	7942	12,342
70	7425	8237	13,045
60	7530	8901	13,872
50	7545	9601	14,189
40	7654	10,142	14,324
30	7800	10,650	14,472
20	8210	11,150	14,493
10	8800	11,311	14,500

Comparison between the proposed navigational pattern mining techniques

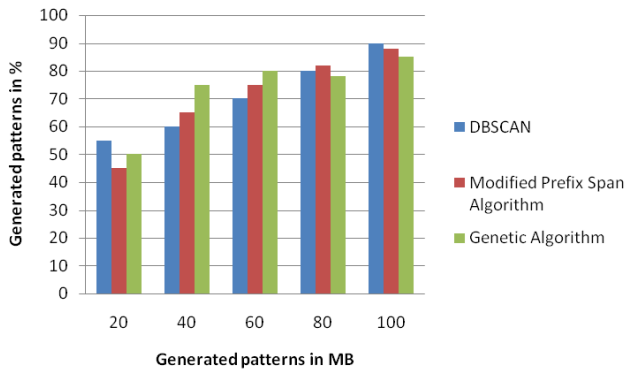

Figure 3. *Comparative analysis.*

Table 7. *Proposed session identification result: time-out=10 mins.*

X (%)	Shared pattern ≥ 3	Shared pattern ≥ 2	Shared pattern ≥ 1
100	6994	7309	10,502
90	6995	7347	10,899
80	6995	7488	12,400
70	6996	7834	13,342
60	7090	8403	14,237
50	7103	9404	14,780
40	7275	9888	15,060

30	7445	10,515	15,182
20	7803	11,456	15,201
10	8307	11,569	15,209

Conclusion

An efficient web page recommender system is proposed using web usage mining techniques. Initially the shared patterns are identified under session identification step of preprocessing. Secondly navigational patterns are identified using DBSCAN algorithm, modified prefix span algorithm and genetic algorithm applied. These techniques are compared in terms of the number of pattern identified for each generated patterns in MB size. The proposed techniques are evaluated using the data set and the best patterns are generated using clustering and sequential pattern mining techniques. Generated patterns will be applied to the online component to find the recommendations of web pages in future.

References

1. Tug E, Merve S, Arslan A. Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. Knowledge Based Sys 2006; 19: 180-186.
2. Kim S, Zhang B. Genetic mining of HTML structures for effective web document retrieval. Appl Intel 2003; 18: 243-256.
3. Agrawal R, Srikant R. Mining Sequential Patterns Proc. Int Conf Data Eng 1995.
4. Wong SSC, Pal S. Mining fuzzy association rules for web access case adaptation, in: Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case-Based Reasoning (ICCBR_Ol), 2001.
5. Han J, Pel J, Yin Y. Mining Frequent Pattems without Candidate Generation, in proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD' 00), Dallas, TX.
6. Agarwal R, Srikant R. Mining sequential Patterns. In proceeding of the International Conference on Data Engineering 1995.
7. Cadez D, Heckerman C, Meek P, Smyth S. White Model-based clustering and visualization of navigation patterns on a web site. Data Mining Knowledge Disc 2003; 7: 399-424.
8. Deshpande M, Karypis G. Selective Markov Models For Predicting Web Page Accesses. Acm Transact Internet Technol 2004; 4: 163-184.
9. Abdelghani G, Adam O, Zaarour O, Nagi M, Elhajj A, Ridley M, Alhaji R. Effective web log mining and online navigational pattern prediction. Knowledge Based Sys 2013; 49: 50-62.
10. Pabarskaite Z, Raudys A. "A process of knowledge discovery from web log data", Systematization and critical review. J Intel Informa Sys 2007.
11. Cooley R, Mobasher B, Srivastava J. Data Preparation for mining World Wide Web browsing patterns. Knowledge Informa Sys 1999.
12. Catlege L, Pitkow J. Characterizing browsing strategies in the world-wide-web. Comput Networks ISDN Sys 1995.
13. Chen M, Park J, Yu P. Data Mining for path traversal patterns in a web environment. Proceedings of the 16th International Conference on Distributed Computing Systems, 1996, 385-392.
14. Berendt B, Mobasher B, Nakagawa M, Spiliopoulou M. The impact of site of structure and user environment on Session reconstruction in web usage analysis. Webkdd2002– Mining web data for Discoveriing Usage Patterns and Profiles 2003; 159-179.
15. Li Y, Zhong N. Web mining Model and its applications for information gathering. Knowledge Based Sys 2004; 17: 207-217.
16. Resnick P, Varian RH. Recommender Systems. Commun Acm 1997; 40: 56-58.
17. Balabonović M, Shoham Y. Learning Information Retrieval Agents: Experiments with Automated Web Browsing. In Proceedings Of The AAAI Spring Symposium On Information Gathering From Heterogenous, Distributed Resources 1995.
18. Pazzani MJ. A Framework For Collaborative, Content-Based And Demographic Filtering. Arti Intel Rev 1999; 13: 393-408.

***Correspondence to**

S Prince Mary

Department of Computer Science Engineering

Sathyabama University

India