Finding of Weighted Sequential Web Access Patterns for Effective Web Page Recommendations

[1]K. Suneetha, [2]Dr. M. Usha Rani
[1]Dept. of MCA, SVEC, Tirupati, Andhra Pradesh, India
[2]Dept. of Computer Science, SPMVV, Tirupati, Andhra Pradesh, India

Abstract
Recommender systems aim at directing users through this information space, toward the resources that best meet their needs and interests by extracting knowledge from the previous users' interactions. Currently much research is focus on web page recommendations using sequential pattern mining techniques. Sequential access pattern mining discovers interesting and frequent user access patterns from web logs. Most of the previous studies have adopted Apriori-like sequential pattern mining techniques, which faced the problem on requiring expensive multiple scans of databases. In this paper a traditional sequential pattern mining algorithm called prefixspan is modified by incorporating two measures such as, spending time and recent view. Then, the weighted sequential patterns are utilized to construct the recommendation model using the Patricia trie-based tree structure. Finally, the recommendation of the current users is done with the help of markov model.

Keywords
Pattern Growth Approach, Frequent Patterns, Projection Database, Web Page Recommendations, Particia-Trie

I. Introduction

Sequential pattern mining has been intensively studied during recent years, there exists a great diversity of algorithms for sequential pattern mining. Sequential pattern mining was first introduced by Agrawal and Srikant [1995]. Sequential mining is the process of applying data mining techniques to a sequential database for the purposes of discovering the correlation relationships that exist among an ordered list of events. It is the process of extracting certain sequential patterns whose support exceeds a predefined minimal support threshold. Sequential pattern mining is an important data mining problem with broad applications including customer purchase behavior analysis, web-log analysis, medical treatments, natural disasters, science and engineering processes, stocks and markets, DNA sequences and gene structures.

Web Mining [7] deals with the extraction of interesting knowledge from the World Wide Web. Web Content Mining focuses on the raw information available in web pages; source data mainly consist of textual data in web pages. Web Structure Mining focuses on the structure of web sites; source data mainly consist of the structural information in web pages (e.g., links to other pages). Web Usage Mining deals with the extraction of knowledge from server log files.

With the massive amount of information available on the World Wide Web, attracts the users to seek and retrieve relevant information from the internet. But, it becomes very difficult for the users to access right or interesting information from the Web. A solution to this problem is web personalization [1]. Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web context, namely, structure, content and user profile data.

In order to provide customized information to the user, many Web-based recommender systems are applied to various web-based applications. One of web personalization system is Web recommender system, which provides substantial user values by personalizing number of sites on the web and provides relevant web pages more efficiently and effectively. Recommender systems aim at directing users through this information space, toward the resources that best meet their needs and interests by extracting knowledge from the previous users' interactions.

The goal of the intelligent recommender system is to determine which web pages are more likely to be accessed next by the current user in the near future. Various traditional techniques such as collaborative filtering [2-3] and hybrid content-based collaborative filtering approaches [4-5] have been developed for supporting web recommendations.

Recently, there has been a substantial interest in using sequential mining approaches to construct web page recommendation systems based on web usage mining, which aims to discover interesting usage patterns derived from the data stored in web server logs or web browser logs.

The paper is organized as follows: Section 2 reviews the recent works available in the literature. Section 3 provides the proposed technique of web page recommendation and section 4 presents experimentation and the results obtained. Section 5 concludes the paper.

II. Background Work

The term 'Web Usage Mining' [12] was introduced by Cooley et al, is used to extract the knowledge hidden in the log files of a web server, interesting patterns concerning the users' navigational behavior can be identified, as well as possible correlations between Web pages and user groups. It consists of three phases they are Data collection and preprocessing, pattern discovery and pattern analysis. In the first phase data is preprocessed in order to identify users, sessions and so on. In the second phase various data mining and statistical methods are applied to find interesting patterns and in the third phase these interested patterns are stored and can be analyzed further using recommender systems. Web usage mining has gained much attention in the literature as a potential approach to fulfill the requirement of web personalization [6, 12-16].

An extensive overview of intelligent methods for Web Personalization has been presented by Sarabjot Singh Anand and Bamshad Mobasher [19]. They have studied the state-of-the-art in Web personalization. Initially, a depiction of the personalization process and a classification of the current techniques to Web personalization have been presented. Also, they have discussed the different sources of data available to personalization systems, the modeling techniques utilized, and the current techniques to analyze these systems. Numerous challenges faced by the researchers in developing these systems and also the solutions to these challenges proposed in literature have been described. They have concluded with a discussion on the open challenges that must be addressed by the research community if this technology is to create a positive impact on user satisfaction with the Web.

A numerous approaches for web page recommendations based

on web usage mining can be categorized into two major groups, such are content-based filtering and collaborative filtering. Content-based filtering [8] systems are solely based on individual users' preferences. The system tracks each user's behavior and recommends items to them that are similar to items the user liked in the past. The users profile are then used to predict the rating for previously unseen items and those deemed as being potentially interesting are presented to the user. Collaborative filtering is traditionally a memory-based approach to recommendation generation, though model-based approaches which does not use any item content descriptions. Collaborative filtering systems invite users to rate objects or divulge their preferences and interests and then return information that is predicted to be of interest to them. This is based on the assumption that users with similar behavior have analogous interests.

In [9], Mobasher et al use statistical significance testing to judge whether a page is interesting to a user. Its main idea is: A duration threshold is calculated for each page using the average duration and standard deviation of the visits to the page; if the duration of a page is longer than the threshold, that page is considered interesting to the user and vice versa. The drawback of such an approach is that it simply divides pages into interesting and uninteresting groups, and neglects the difference in the degrees of interest.

An approach for recommendations of unvisited pages has been presented by Forsati, R et al [18]. They have focused on the recommender systems based on the user's navigational patterns and provided proper recommendations to cater to the current needs of the user. The group of users with analogous browsing patterns has been identified by employing an offline data preprocessing and clustering technique. The experiments conducted on real usage data from a commercial web site have demonstrated a considerable enhancement in the recommendation efficiency of the proposed system.

Web page recommendation based on weighted association rules was proposed by R. Forsati, M. R. Meybodi [10]. Here, they have proposed three algorithms to solve the web page recommendation problems. In the first algorithm, a distributed learning machine has been employed to study the behavior of previous users' and to recommend pages to the current user based on the learned patterns. In the second algorithm, Weighted Association Rule mining algorithm has been applied for recommendation purposes. One of the challenging problems in recommendation systems is dealing with unvisited or newly added pages. Finally, in the third algorithm, the above two algorithms have been combined to enhance the competence of web page recommendation.

III Finding of Weighted Sequential Web Access Patterns for Web Page Recommendations

Different from the majority of the existing web recommendation techniques, we propose an efficient web recommender system that uses weighted sequential pattern mining technique. Here, Prefixspan sequential pattern mining algorithm is modified by incorporating two measures such as spending time and recent view, to find relevant sequential patterns. Then, the markov model described in[22] is used to recommend the web pages. The major steps in generating recommendations of web pages are defined as follows,

- Data Preprocessing
- Finding of weighted sequential web access patterns by using W-Prefixspan algorithm
- Construction of a Pattern tree
- Generation of web page recommendations

A. Data Preprocessing

Data preprocessing is a pre-requisite phase before the data can be mined to obtain useful and interesting patterns. Initially all users 'web access activities of a website are recorded by the WWW server of the website and stored into the Web Server Logs and the web log file consists of following key information: IP address, access time, HTTP request method used, URL of the referring page and browser name. After obtaining the Web Server Logs, sequential pattern mining process is applied in order to convert the web log file into the sequential database which should be in a proper format to mine the weighted sequential patterns.

1. User Identification

Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. A user session is fixed for certain period of time and if the user is having unique IP address then the user is said to be a new user. If the user session is reached to he particular period with the same IP address then the user will be acted as a new user.

2. Construction of Weighted Sequential Database

A weighted sequential database is generated with sequence of web pages visited by the user, time spent on corresponding web page and its recent information. This step consists of, (i) identifying the different users' sessions from the usually very poor information available in web log files and (ii) reconstructing the users' navigation path within the identified sessions.

B. Finding of Weighted Sequential Web Access Patterns by Using W-Prefixspan Algorithm

1. Prefixspan Algorithm

The proposed recommendation approach makes use of existing Prefixspan algorithm by incorporating two important measures such as spending time and recent view. Prefixspan is based on the idea of database projection and sequential pattern growth. This algorithm examines only the prefix subsequences after scanning the sequence database once and then projects their corresponding postfix subsequences into projected database. Prefixspan algorithm doesn't generate and tests candidate sequences, non existent in a projection database. Projected database keeps on shrinking because only the suffix subsequences of a frequent prefix are projected into a projected database. But the major cost of prefixspan is the construction of projected databases. The procedure of prefixspan algorithm is defined as follows.

Input : A sequence database S, and the minimum support threshold min_sup

Output : The complete set of sequential patterns

Method: Call $PrefixSpan(\langle\ \rangle, 0, S)$.

Subroutine: $PrefixSpan(\alpha, l, S|_\alpha)$

Parameters: α : a sequential pattern; l : the length of α ; $S|_\alpha$: the α-projected database, if $\alpha \neq \langle\ \rangle$; otherwise, the sequence database S.

Method:

1. Scan SI_α once, find the set of frequent items b such that
   a) b can be assembled to the last element of α to form a sequential pattern; or
   b) $\langle b \rangle$ can be appended to α to form a sequential pattern.
2. For each frequent item b, append it to α to form a sequential pattern α', and output α';
3. For each α', construct α'-projected database $SI_{\alpha'}$, and call *PrefixSpan* $(\alpha', l+1, SI_{\alpha'})$.

2. Mining of Weighted Sequential Patterns

The traditional sequential pattern mining problem is extended by allowing a weight to be associated with each page in a user session to reflect interest of each page within the user session. In turn, this provides with an opportunity to associate a weight parameter with each page in a resulting sequential pattern, which called a Weighted Sequential Pattern (WSP). Here, PrefixSpan [17] is modified by incorporating weightage measures such as spending time and recent view into the mining procedure.

Spending time: We propose a weighting measure which is calculated from web logs to extract the interest of web page for the user. The importance of each page can be identified based on how much time the user is spending on a more useful page, because if a user is not interested in a page, he/she do not spend much time on viewing the page and usually jumps to another page quickly. Hence to identify the interest of the users, Spending time is one of an important measure. So, by using this measure one can find out the interesting relationships then from this we can get good number of web page recommendations. Recent view: Generally a web page which is accessed recently is having more importance than the older one. So, a significant measure taken for sequential pattern mining is recent view that describes whether the page is accessed recently or not.

3. W-Prefixspan Algorithm

For finding frequent sequential patterns, an eminent W-PrefixSpan algorithm [23] is developed by modifying the existing traditional PrefixSpan algorithm, which uses the pattern growth methodology. Let us consider a weighted database W_{ij} . A sequence Ws is said to be a sub sequence of W_{ij} only if,

(1) W_s is a subsequence of W_{ij} , $W_{ij} \in W_s$

(2) $t_1 < t_2 < \cdots < t_m$ where, t_1 is the time at which p_{ij} occurred in W_s , $1 \le r \le m$.

A sequence is said to be SR sequence W_{ij} if and only if, (1) W_s is a subsequence of W_{ij} , (2) the W-support should be satisfied. The following is the procedure for W-Prefixspan algorithm:

- By scanning the weighted sequential database once it finds 1-length weighted sequential patterns.
- It finds 1-SR (Spending time and Recent view) patterns which satisfies the predefined W-Support threshold value.
- Later projection database is formed by projecting the collection of postfixes of mined 1-SR sequence.
- Then, the 2-length SR-patterns are mined from the projected database by computing the weighted support on the projected database.
- This process is repeated recursively until all SR sequential patterns are mined.

C. Construction of Pattern Tree Model

In this section a pattern tree is constructed using the procedure defined in [15, 11]. A pattern tree is generated using Patricia tree data structure for effective web page recommendations. Patricia trie is used to store a set of strings but in regular trie, single character is stored in each node. By using Patricia trie, the tree can be even more compacted when compare with the regular trie.

The procedure for constructing the pattern tree in the proposed system is as follows:

1. Create an empty root node.
2. Add the most sub pattern in the SR-sequential pattern set into a node next to the root node
3. Insert the postfixes of pattern into child node only if the current pattern to be inserted is a super pattern of inserted patterns
4. Otherwise, current pattern is inserted into the node next to the root node
5. Step 3 and step 4 is repeated for every pattern in the mined SR-pattern set.

D. Generation of Web Page Recommendations

After mining the weighted sequential patterns, the Patricia tree is constructed. From the Patricia tree, a recommendation model is developed based on Markov model for predictions of users to find web pages they want to visit. Since the recommendation process is based on the behavior of previous users access pages.

Based on the sequence of pages accessed previously by the user helps to find out the next pages that are to be accessed by the same user using Markov model [12].

If a new user wants to get recommendations then his sequence path is matched with the Patricia-trie structure to find out whether it is from the same node or from its child node. By using the probability definitions defined in Markov model, are used to find the accurate recommendations.

IV Results and Discussion

This section presents the results obtained from the experimentation and its detailed discussion about the results. The proposed approach of web page recommendation is experimented with the synthetic dataset and the result is evaluated with the precision, applicability and hit ratio.

A. Experimental Set Up and Dataset Description

The proposed web page recommendation approach is implemented in Java (jdk 1.6) with I3 processor of 2GB RAM. Here, the synthetic dataset is generated as like the same format of real datasets and the performance of the proposed approach is evaluated with the evaluation metrics. The generated synthetic dataset is divided into two parts such as, Training dataset and test dataset. The training data set is used for building the pattern tree and test data set is used for testing the web page recommendations.

B. Comparison of Prefixspan with W-Prefixspan

Here the performance of W-Prefixspan algorithm is compared with Prefixspan algorithm and the results shown that the performance of proposed algorithm is increased in finding the number of patterns which helps us in finding good number of web page recommendations. The values are plotted in a graph and the results are shown in the fig. 1.

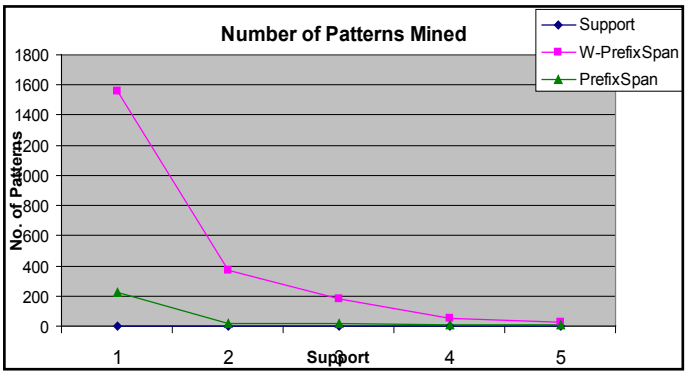

Fig. 1: Number of Patterns Mined

V. Conclusion

In this paper a traditional sequential pattern mining algorithm called prefixspan is modified by incorporating two measures such as, spending time and recent view. Then, the weighted sequential web access patterns are mined from weighted sequential database to construct the recommendation model.

After mining the sequential patterns, Patricia trie-based tree structure is constructed. Finally, the recommendation of the current users is done with the help of markov model.

References

[1] M. Eirinaki, M. Vazirgiannis,"Web mining for web personalization", ACM Transactions on Internet Technology, Vol. 3, No. 1, 2003, pp. 1-27.

[2] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, J. Riedl,"GroupLens: applying collaborative filtering to usenet news", Communications of the ACM, 40(3), 1997, pp. 77-87.

[3] T.W. Yan, H. Garcia-Molina,"The SIFT information dissemination system", ACM Transactions on Database Systems, Vol. 24, No. 4, 1999, pp.529-565.

[4] T. Joachims, D. Freitag, T. Mitchel,"WebWatcher: a tour guide for the World Wide Web", Proc. of the 5th International Joint Conference on AI, Japan, 1997, pp. 770-775.

[5] C. Shahabi, F. Banaei-Kashani, Y. Chen, D. McLeod, "Yoda: an accurate and scalable web-based recommendation system", Proc. of the 6th International Conference on Cooperative Information Systems (CoopIS 2001), Trento, Italy, September 2001.

[6] B. Mobasher, R. Cooley, J. Srivastava,"Automatic Personalization based on Web Usage Mining", Communications of the ACM, Vol. 43 No. 8, 2000, pp. 142-151.

[7] Oren Etzioni,"The world-wide web: Quagmire or gold mine?", Communications of the ACM, 39(11), pp. 65–68, 1996.

[8] P. Resnick, N. Iacovou, M. SUSHAK, P. Bergstrom, J. Riedl,"Grouplens: An Open Architecture for Collaborative Filtering of Netnews", Proceedings of the 1994 Computer Supported Collaborative Work Conference, 1994.

[9] B. Mobasher, H. Dai, T. Luo, M. Nakagawa,"Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data", Proceedings of the IJCAI 2001Workshop on Intelligent Techniques for Web Personalization (ITWP01), August 2001.

[10] R. Forsati, M. R. Meybodi,"Effective Page Recommendation Algorithms Based on Distributed Learning Automata and Weighted Association Rules".

[11] Utpala Niranjan, R. B. V. Subramanyam, V. Khanaa, "Developing a Web Recommendation System Based on Closed Sequential Patterns", Communications in Computer and Information Science, Vol. 101, No. 1, pp. 171-179, 2010.

[12] R. Cooley, B. Mobasher, J. Srivastava,"Web Mining: Information and Pattern Discovery on The World Wide Web", Proceedings of IEEE International Conference Tools With AI, 1997, pp. 558–567.

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

[14] X. Fu, J. Budzik, K. Hammond,"Mining Navigation History for Recommendation", In Proceedings of The Fifth International Conference on Intelligent User interfaces, 2000, pp. 106–112.

[15] M. Gery, H. Haddad,"Evaluation of Web Usage Mining Approaches For User's Next Request Prediction", Proceedings of The Fifth ACM International Workshop on Web Information and Data Management, 2003, pp. 74–81.

[16] M. D. Mulvenna, S. S. Anand, A. G. Buchner, "Personalization On the Net Using Web Mining", Communications of the ACM, Vol. 43, No. 8, 2000, pp.123–125.

[17] Jian Pei; Jiawei Han; Mortazavi-Asl, B.; Pinto, H.; Qiming Chen; Dayal, U.;Mei-Chun Hsu,"PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth", in Proceedings of 17th International Conference on Data Engineering, 2001.

[18] C.P. Sumathi,R. Padmaja Valli,T. Santhanam,"Automatic Recommendation of Web Pages in Web Usage Mining",(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 09, pp. 3046-3052, 2010.

[19] M. Deshpande, G. Karypis,"Item-Based Top-N Recommendation Algorithms", ACM Transactions on Information Systems (TOIS), 2004.

[20] Utpala Niranjan, R.B.V. Subramanyam, V.Khana,"An Efficient System Based On Closed Sequential Patterns for Web Recommendations", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No. 4, May 2010.

[21] Faten Khalil, Jiuyong Li, Hua Wang,"Integrating Recommendation Models for Improved Web Page Prediction Accuracy", Proceedings of the thirty-first Australasian conference on Computer science, Vol. 74, 2008.

[22] Faten Khalil, Jiuyong Li, Hua Wang,"Integrating Recommendation Models for Improved Web Page Prediction Accuracy", Proceedings of the thirty-first Australasian conference on Computer science, Vol. 74, 2008.

[23] K. Suneetha, Dr.M.Usha Rani,"Web Page Recommendation Approach Using Weighted Sequential Patterns and Markov Model", in Global Journal of Computer Science and Technology, Vol. 12, Issue 9, Version 1.0, April 2012.



Ms. K. Suneetha obtained Bachelor's degree in Sciences from S.V.University, Tirupathi. Then she obtained her Master's degree in Computer Applications from S.V.University. She is working as Assistant Professor in the Department of Master of Computer Applications at Sree Vidyanikethan Engineering College, A.Rangampet, Tirupati. She is pursuing her Ph.D. in Computer Science in the area of Data Warehousing and Data Mining. She is in teaching since 2001. She presented many papers at National and Internal Conferences and published articles in National & International journals.



Dr. M. Usha Rani is an Associate Professor in the Department of Computer Science and HOD for MCA, Sri Padmavati Mahila Viswavidyalayam (SPMVV Woman's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence and Expert Systems. She is in teaching since 1992. She presented many papers at National and Internal Conferences and published articles in national & international journals. She also has written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in the areas like Artificial Intelligence, Data Warehousing and Data Mining, Computer Networks and Network Security etc.