Asian Journal of Information Technology 11 (6-12): 270-275, 2012
ISSN: 1682-3915
© Medwell Journals, 2012

Performance Analysis of Web Page Recommendation Algorithm Based on Weighted Sequential Patterns and Markov Model

K. Suneetha
Department of MCA, Sri Vidyanekethan Engineering College, Andhra Pradesh, India

Abstract: Web usage mining techniques helps the users to predict the required web page recommendations. In recent times there has been a considerable significance given to sequential mining approaches to construct Web Page Recommendation Systems. This study focuses on developing a web page recommendation approach for accessing related web pages more efficiently and effectively using weighted sequential pattern mining and Markov Model. Here, researchers have developed an algorithm called, W-PrefixSpan that is the modification of traditional Prefixspan algorithm including the constraints of spending time and recent visiting to extract weighted sequential patterns. Then, by utilizing Weighted Sequential Patterns Recommendation Model is constructed based on Patricia-trie data structure. Later the web page recommendation of the current users is done with the help of Markov Model. Experimentation is done with the help of synthetic dataset and we present the performance report of web page recommendation algorithm in terms of precision, applicability and hit ratio. The results have shown that the precision of the algorithm is improved by 5% than the earlier algorithm. Also, researchers have achieved high applicability in the support of 50% and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

Key words: Prefixspan, web page recommendation, weighted sequential pattern, patricia-trie, Markov Model

INTRODUCTION

With the explosive growth of information available on the World Wide Web it has become much more difficult to access relevant information from the web. Web mining is one of the most propitious fields of Data Mining which deals with the extraction of meaningful or relevant knowledge from the World Wide Web (Etzioni, 1996). More specifically, web content mining is the branch of web mining that focuses on extracting the raw information available in web pages. Web usage mining is also the branch of web mining which deals with the extraction of relevant information from server log files. Here, the source data is mainly composed of the (textual) logs that are gathered when the users access the web servers and might be depicted in standard formats and classic applications are those based on user modeling approaches, namely web personalization, adaptive web sites and user modeling (Mulvenna *et al*., 2000).

Web personalization (Anand and Mobasher, 2003) refers to any action that adapts the information or services provided by a web site to the needs of a particular user or a set of users by using the knowledge procured from the navigational activities and individual interests of users recorded in the web usage logs in conjunction with the content and the structure of the web site (Eirinaki and Vazirgiannis, 2003). The role of Web Personalization System is to provide the users with an information they desire or need without expecting from them to inquire for it explicitly (Mulvenna *et al*., 2000). Web Recommender System is one type of personalized web application which provides substantial user value by personalizing numerous sites on the web (Schafer *et al*., 2006). Recently, Web-based Recommender Systems (RS) are widely applied to provide diverse type of customized information to the users. Generally, there are many data mining techniques such as association rule mining, sequential pattern discovery, clustering and classification. Among them, sequential pattern-mining method is an extensively used data analysis technique in web usage mining (Wang and Shao, 2001). Sequential pattern mining (Zhao and Bhowmick, 2003), an advance of association rule mining is an imperative subject of data mining, often applied for extracting the useful information (Hou and Zhang, 2008). In recent times, there has been a considerable significance given to sequential mining approaches to construct web Page Recommendation Systems. This study focuses on developing a web page recommendation approach for accessing related web pages more efficiently and effectively. The main goal of this approach is to determine which web pages are more likely to be accessed next by the current user in the near future.

MOTIVATING ALGORITHMS

This study describes the motivating algorithms of the proposed web recommendation approach. Here, researchers have mentioned three different algorithms that are based on weighted association rules, Markov Model and closed sequential patterns.

Weighted association rule-based web page recommendation algorithm: Web page recommendation based on weighted association rules was proposed by Forsati and Meybodi (2010). Here, they have proposed three algorithms to clear up 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. Finally in the third algorithm, the earlier two algorithms have been combined to enhance the competence of web page recommendation. The general block diagram of the hybrid algorithm based on distributed learning automata and weighted association rule mining algorithm is given in Fig. 1.

Markov Model-based web page recommendation algorithm: The probability theory-based Markov Model is effectively utilized by Khalil *et al.* (2008) for web page recommendation. Here, the web page access prediction accuracy has been enhanced by including three prediction models such as Markov Model, Clustering and association rules according to certain constraints. They have integrated these three models using 2-Markov Model computed on clusters achieved by means of k-means clustering algorithm and cosine distance measures for states that belong to the majority class and performing association rule mining on the rest. The algorithmic procedure is described as follows:

Training:

- Combine functionally related web pages according to services requested
- Group user sessions into l-clusters
- Construct a k-Markov Model for each cluster
- For Markov Model states where the majority is not clear
- Mine association rules for each state
- End For

Prediction:

- For each coming session
- Find its closest cluster
- Use relevant Markov Model to make prediction
- If the predictions are made by states that do not belong to a majority class
- Use association rules to make a revised prediction
- End If
- End For

Closed sequential pattern-based web page recommendation algorithm: Closed sequential pattern, one of the variants of sequential pattern is used by the Niranjan *et al.* (2010a, b) for web page recommendation. The proposed system was mainly based on discovering the closed sequential web access patterns. Firstly, the PrefixSpan algorithm has been applied on the preprocessed web server log data for extracting the sequential web access patterns. Then, the closed sequential web access patterns have been mined from the

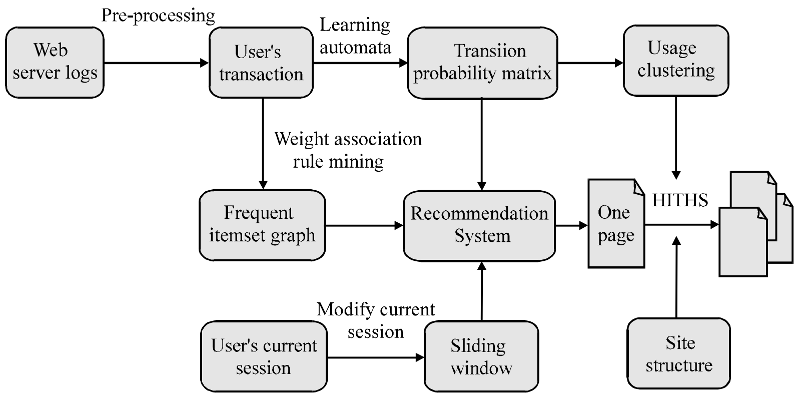

Fig. 1: Weighted association rule-based web page recommendation algorithm

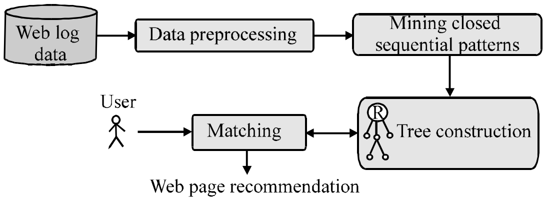

Fig. 2: Closed sequential pattern-based web page recommendation algorithm

complete set of sequential web access patterns via post-pruning approach. Subsequently, a pattern tree, a compact representation of closed sequential patterns has been build from the mined closed sequential web access patterns. Moreover, the patricia trie based data structure has been employed in the construction of the pattern tree. Based on the constructed pattern tree, the proposed system has provided recommendations for a given user's web access sequence. The general block diagram of the recommendation algorithm based on closed sequential pattern mining algorithm is given in Fig. 2.

CONTRIBUTIONS OF THE STUDY

The main contributions of the research are given as follows:

- Researchers have presented an algorithm for web page recommendation by combining the weighted sequential pattern and Markov Model
- Researchers have developed an algorithm called W-prefixSpan that is the modification of traditional Prefixspan algorithm including the constraints of spending time and recent visiting
- Researchers analyze the performance of W-prefixspan algorithm with the Prefixspan algorithm in terms of computation time and memory usage
- Researchers present the performance report of web page recommendation algorithm in terms of precision, applicability and hit ratio

WEB PAGE RECOMMENDATION ALGORITHM BASED ON WEIGHTED SEQUENTIAL ACCESS PATTERNS

In recent years there have been increasing interests in applying web usage mining techniques to build web page recommendations. With the intention of real world applicability, researchers have developed an approach for web page recommendation using weighted sequential pattern and Markov Model. Here, the traditional sequential pattern mining algorithm called Prefixspan is modified significantly by incorporating the significant measures such as spending time and recent view to mine more useful patterns. Then, the Markov Model (Khalil *et al.*, 2008) is used to recommend the web pages. The steps in the algorithm for generating recommendations to the user could be briefly summarized as follows (Suneetha and Rani, 2012):

Step 1 (Data preprocessing): This step is used to extract the useful and relevant information from raw web logs. This raw web logs need to be processed analyzed and converted into proper format of sequential database to mine the weighted sequential patterns.

Step 2 (W-PrefixSpan for mining of weighted sequential web access patterns): To identify the interesting sequential patterns from a weighted sequential database, the proposed recommendation system utilizes a traditional Prefixspan (Pei *et al.*, 2001) which is a well known pattern-growth algorithm by incorporating two measures spending time and recent view into the mining procedure.

Step 3 (Building pattern Tree Model): Once weighted sequential patterns are mined, a pattern tree is constructed using the procedure defined by Niranjan *et al.* (2010a, b) is applied to the proposed approach for constructing tree structure. Patricia-based data structure is used for web page recommendation due to the advantages of particia structure over the trie structure.

Step 4 (Generation of recommendations using Markov Model): Here, researchers make use of the Markov Model described by Khalil *et al.* (2008) that is used in the identification of the next page to be accessed by the web site user based on the sequence of earlier accessed pages. The accurate recommendations can be found using the definitions of the probability described by (Suneetha and Rani, 2012).

Data preprocessing: Due to large amount of irrelevant information available in the web log, the original log data cannot be directly used in the web mining procedure. A web log file consists of IP address, access time, HTTP request method used URL of the referring page and browser name. To mine the required sequential patterns, it is very difficult to directly use the web log data. Hence, the following preprocessing techniques can be used to convert the data into proper format.

User identification: In this step a sequential database is constructed by identifying each user accessing web pages. Users may be tracked based on IP address and user session. A new IP address is used to identify the new user but at the same time, the user session should be fixed for a particular time period.

Weighted sequential database generation: The weighted sequential database is generated including the sequence of web pages visited by the user, time spent by the user on corresponding web page and its recent information.

W-PrefixSpan for mining of weighted sequential web access pattern: To identify the interesting sequential patterns from a weighted sequential database, the proposed recommendation system utilizes a traditional Prefixspan (Pei *et al*., 2001) which is a well known pattern-growth algorithm by incorporating two measures spending time and recent view into the mining procedure.

Spending time is an important measure for the researchers who are attempting to identify the interest of the users. Time spent by the user within a particular page is necessary to identify the importance of web pages.

Recent view is another important measure to find whether the page is accessed recently or not. More importance should be given for the web pages which are accessed recently because the behavior of the user surely varies depend on the time so the recent behavior of the user is significant for finding the sequence analysis.

W-PrefixSpan algorithm: An efficient sequential pattern mining algorithm called W-Prefixspan (Suneetha and Rani, 2012) is developed by modifying traditional sequential pattern mining algorithm Prefixspan for finding frequent sequential patterns.

Initially, the weighted sequential database W_{ij} is given as an input to the proposed W-PrefixSpan algorithm that discovers the 1-length weighted sequential patterns from the weighted sequential database by scanning the database once. The 1-SR patterns (spending time with recent view) which satisfy the predefined support threshold are mined from the sequential database by simply scanning the database. The W-support for the 1-length pattern is computed as follows:

$$W\_\ sup(p)=\frac{1}{N}\frac{\sum_{i=1}^{N_T\in p} I_s(i)\times R(i)}{\sum_{i=1}^{N_T\in p} R(i)}$$

Where:

N = Number of user transaction in the weighted sequential database

N_T = Number of transaction that contains the web page p

R (i) = Recent information

$$I_s(i)=\sum_{i=1}^{N_T}\left(\frac{s_i}{\sum_{i=1}^{M_T} s_i}\right)$$

Where:

M_T = Total number of web pages in one transaction

S_i = Spending time

Then, the projection database is formed by projecting the collection of postfixes of mined 1-SR sequence. In projection database, n disjoint subsets are generated if the mined 1-SR patterns contain n number of sequence. Then, the 2-length SR-patterns are mined from the projected database by computing the weighted support on the projected database. Again, the projected database is formed with the help of mined 2-SR patterns and this process is repeated recursively until all SR sequential patterns are mined.

Building of pattern tree model: In this study, a pattern tree is constructed using the procedure defined by Niranjan *et al*. (2010a, b) is applied to the proposed approach for constructing tree structure. The constructed pattern tree is used in making the web page recommendations for users. The constructed pattern tree is based on Patricia-trie data structure. The procedure for constructing a pattern tree defined by Suneetha and Rani (2010) is applied to the proposed approach.

Generation of recommendations using Markov Model: Markov Models are the most effective techniques for web page access prediction and to improve the web server access efficiency. The Markov Model used for the identification of next page to be accessed by the user based on the sequence of earlier accessed pages. Whenever, a new user comes to get the recommendation, the sequence path of the new user is matched with the Patricia-trie structure. Then, the subsequent web page whether it may be from same node or from its child node is retrieved. Now, the sequence path of the new user is used to find the accurate recommendation using the probability definition used in the earlier research (Khalil *et al*., 2008). The probability of computation is carried out to find the most important sequence for the user (Suneetha and Rani, 2012). The probability, pro $(s_{n+1}|s)$ is estimated by using all sequences of all users in tree structure constructed from the weighted sequential database:

$$P\,(s_{n+1} = (s \in s_1, s_2, s_n, s_{n+1}, \ldots, s_m) | s_1, s_2, \ldots, s_n)$$
$$= \frac{W\_sup(s_1, s_2, \ldots, s_n)}{W\_sup\,(s_1, s_2, s_n, s_{n+1}, \ldots, s_m)}$$

Then, the final recommendation is based on the following equation:

$$s_{n+1} = \arg\ sort\ \{s_{n+1}^{(1)}, s_{n+1}^{(2)}, s_{n+1}^{(3)}\}$$

RESULTS AND DISCUSSION

This study presents the detailed discussion about the results which are obtained from the experimentation. The experimentation is done on the proposed approach using synthetic dataset and the results are evaluated with the precision, applicability and hit ratio.

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 (It is used for building the pattern tree model and test dataset (It is used for testing the web recommendation approach).

Evaluation metrics: For evaluating the proposed approach, researchers have used three measures such as precision, applicability and hit ratio (Niranjan *et al*., 2010a, b). The formal definition of these three measures are given as:

$$\text{Precision} = \frac{C^+}{C^+ + I^-}$$

Where:
C^+ = Number of correct recommendations
I^- = Number of incorrect recommendations

Definition: Let S = $s_1s_2 \ldots s_js_{j+1} \ldots s_n$ be a web access sequence of test dataset R = $\{r_1, r_2, \ldots r_k\}$. The recommendation is generated by using the constructed pattern tree for the subsequence S_{sub} = $s\ s_{1..2}s_j$ (minlen≤j≤maxlen). The recommendation R is said to be correct if it contains s_{j+1} ($s_{j+1} \in R$). Otherwise, R is said to be incorrect recommendation:

$$\text{Applicability} = \frac{C^+ + I^-}{|N|}$$

Where:
|N| = Total number of given requests

$$\text{Hit ratio} = \text{Precision} \times \text{Applicability} = \frac{C^+}{|N|}$$

Performance of the web page recommendation algorithm: The proposed web page recommendation approach is analyzed with the help of precision, applicability and hit ratio. Here, the testing dataset is given to the Tree Model constructed with the help of training data. Subsequently, the precision is computed based on the result obtained for the test dataset. Here, the results are taken for the proposed approach, PrefixSpan algorithm-based approach and the earlier algorithm (Niranjan *et al*., 2010a, b) and the graphs are plotted for the taken results which have shown in Fig. 3-5.

In Fig. 3, the W-prefixSpan algorithm has achieved the precision of about 70% where the Niranjan algorithm has achieved only 65%. In the Fig. 4, researchers have achieved high applicability in the support of 50% and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

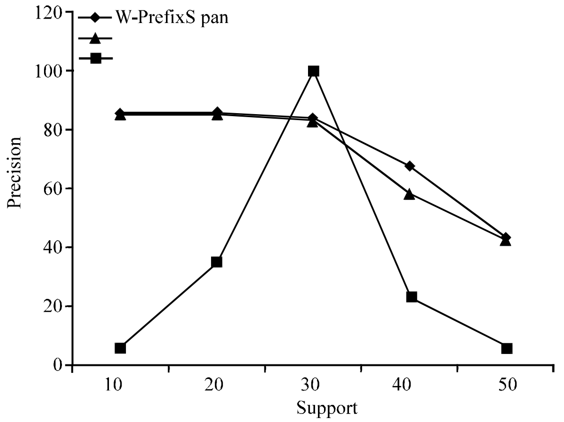

Fig. 3: Precision

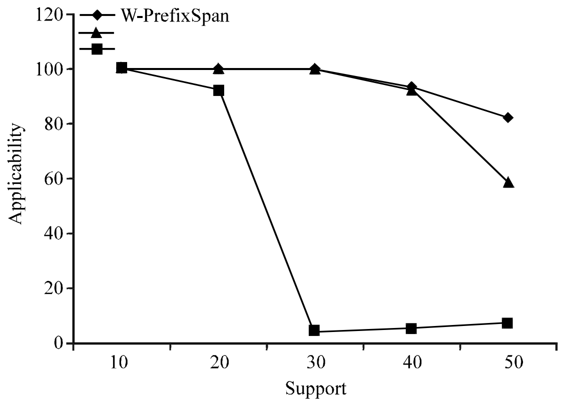

Fig. 4: Applicability

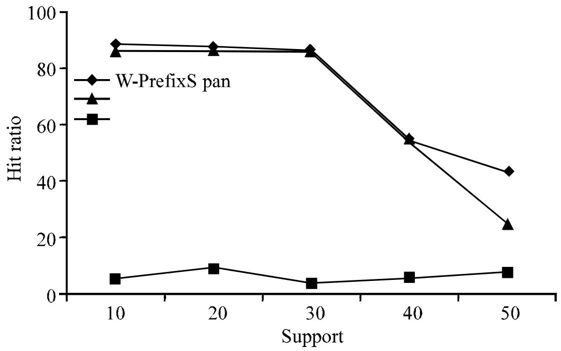

Fig. 5: Hit ratio

CONCLUSION

Researchers have proposed a web page recommendation algorithm using weighted sequential patterns and Markov Model. Here, researchers have presented W-PrefixSpan algorithm that is the developed by incorporating the weightage constraints such as spending time and recent visiting with the prefixspan algorithm. The mined weighted sequential patterns are then utilized to construct the recommendation model using the Patricia-trie based tree structure. At last, Markov Model-based recommendation is carried out for the current users by matching the visiting path with the tree and Markov Model. The experimentation is done with the help of synthetic dataset and the performance of W-Prefixspan algorithm as well as web page recommendation algorithm is analyzed. From the results, the precision of the algorithm is improved by 5% than the earlier algorithm. Also achieved high applicability in the support of 50% and in terms of hit ratio, the proposed algorithm ensured that the performance is considerably improved for various support values.

REFERENCES

Anand, S.S. and B. Mobasher, 2003. Intelligent techniques for web personalization. Proceedings of the 2003 International Joint Conference on Artificial Intelligence, Volume 3169, August 11, 2003, Springer-Verlag, Berlin Heidelberg, pp: 1-36.

Eirinaki, M. and M. Vazirgiannis, 2003. Web mining for web personalization. ACM Trans. Internet Technol., 3: 1-27.

Etzioni, O., 1996. The world-wide web: Quagmire or gold mine? Commun. ACM., 39: 65-68.

Facca, F.M. and P.L. Lanzi, 2003. Recent developments in web usage mining research. Lect. Notes. Comput. Sci., 2737: 140-150.

Forsati, R. and M.R. Meybodi, 2010. Effective page recommendation algorithms based on distributed learning automata and weighted association rules. Exp. Syst. Applic., 37: 1316-1330.

Hou, S. and X. Zhang, 2008. Alarms association rules based on sequential pattern mining algorithm. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, Volume 2, October 18-20, 2008, Shandong, pp: 556-560.

Khalil, F., J. Li and H. Wang, 2008. Integrating recommendation models for improved web page prediction accuracy. Proceedings of the 31th Australasian Computer Science Conference, (ACSC`08), Wollongong, NSW, pp: 91-100.

Mulvenna. M.D., S.S. Anand and A.G. Buchner, 2000. Personalization on the net using web mining. Commun. ACM., 43: 123-125.

Niranjan, U., R.B.V. Subramanyam and V. Khana, 2010a. An efficient system based on closed sequential patterns for web recommendations. Int. J. Comput. Sci., 7: 26-26.

Niranjan, U., R.B.V. Subramanyam and V. Khanaa, 2010b. Developing a web recommendation system based on closed sequential patterns. Commun. Comput. Infor. Sci., 101: 171-179.

Pei, J., J. Han, M.A. Behzad, P. Helen, Q. Chen and M.C. Hsu, 2001. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany, pp: 215-224.

Schafer, J.B., J.A. Konstan and J.T. Riedl, 2006. Recommender Systems for the Web. In: In: Visualizing the Semantic Web, Geroimenko, V. and C. Chen (Eds.)., Springer Verlag, 2nd Edn., pp: 102-123..

Suneetha, K. and M.U. Rani, 2012. Web page recommendation approach using weighted sequential patterns and markov model. Global J. Comput. Sci. Technol., Vol. 12.

Wang, F.H. and H.M. Shao, 2004. Effective personalized recommendation based on time-framed navigation clustering and association mining. Exp. Syst. Applic., 27: 365-377.

Zhao, Q. and S.S. Bhowmick, 2003. Sequential pattern mining: A survey. Technical Report, CAIS, Nanyang Technological University, Singapore.