RECOMMENDATION GENERATION BY INTEGRATING SEQUENTIAL PATTERN MINING AND SEMANTICS

Geethapriya Uvaraja

Computer Science and Engineering Anna University, Tamilnadu, India

Abstract

As the Internet usage keeps increasing, the number of web sites and hence the number of web pages also keeps increasing. A recommendation system can be used to provide personalized web service by suggesting the pages that are likely to be accessed in future. Most of the recommendation systems are based on association rule mining or based on keywords. Using the association rule mining the prediction rate is less as it doesn't take into account the order of access of the web pages by the users. The recommendation systems that are key-word based provides lesser relevant results. This paper proposes a recommendation system that uses the advantages of sequential pattern mining and semantics over the association rule mining and keyword based systems respectively.

Keywords: *Sequential Pattern Mining, Taxonomy, Apriori-All, CS-Mine, Semantic, Clustering*

----------***----------

1. INTRODUCTION

A recommender system understands the users' navigation pattern by exploiting web usage mining and provides personalization service based on the results of mining. The recommender system will propose links of possible interest to the user. Data mining and machine learning algorithms are used in developing a recommendation system. In general the recommendation systems take the users' navigations in the form of log file. Some kind of mining is performed over the data in the log file to find the usage patterns. When a new user arrives, the current access is matched with the patterns that are mined to generate recommendations.

The semantic information is used along with web usage data to get more relevant patterns. Since web logs lack semantic information about web pages visited by users, it is difficult to understand the preferences and intents of users. With the development of the Semantic Web, semantics in web content can be used for improving the relevancy of recommendation.
The objective of this paper is to provide recommendations by integrating sequential pattern mining and semantics. The sequential patterns obtained from two sequential pattern mining algorithms Apriori-All and Conditional Sequence Mine are compared. Taxonomy is constructed for the website for which the recommendations are to be generated. The keywords representing each web page of the website are found and they are mapped with the taxonomy to obtain the categories by using the similarity measure. The documents are then clustered based on the categories. The recommendation rules generated by sequential pattern mining algorithm and the clusters are used to generate recommendations.

The rest of the paper is organized as follows. Section 2 describes the literature survey and the related work. Section 3 presents the system design. Section 4 describes the results of the modules implemented and performance evaluation. Section 5 is about the conclusion.

2. LITERATURE SURVEY AND RELATED WORK

Web mining is the process of mining or extracting useful or valuable information on web. Web mining is an application of data mining techniques on World Wide Web to extract patterns from resources available on web. Web mining has three categories. They are

1. Web content mining
2. Web usage mining
3. Web structure mining

Web content mining is the discovery of patterns from contents of the web pages for improving the relevancy in the field of web mining. It involves information extraction, knowledge discovery and analysis of a collection of documents. Web usage mining is the discovery of the patterns from user navigation data in the form of log file. Web structure mining is the discovery of knowledge from hyperlinks and link structure between the web pages.

A recommendation system incorporates the web mining techniques in order to provide personalized suggestions to the users. There are various approaches through which recommendations can be generated. Most of the traditional recommendation systems are based on clustering, the user given rating and feedback provided by the users. Due to the

presence of a huge number of meaningful clusters and profiles for visitors of a typical highly rated Website, the model-based or distance-based method tend to make too strong and simplistic assumptions and becomes excessively complex and slow. Collaborative filtering systems predict a person's affinity for items or information by connecting that person's recorded interests with the recorded interests of a community of people and sharing ratings between likeminded persons. This approach is based on an assumption that those who agreed in the past tend to agree again in the future. It cannot acquire accurate recommendation results when user rating data are extremely sparse. This approach can make suggestions to a user that are outside the scope of previous selected items. But it not scalable when number of users and items increases. A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences.

In the personalized recommendation systems, which are based on web usage mining, association-mining technology was applied to predict user-browsing behavior. This method scales better with large datasets compared to systems based on clustering. Weighted association rule model was proposed by assigning a significant weight to the pages based on time spent by each user on each page and visiting frequency of each page. The weighting measure was used to judge the importance of a page to a user, and try to give more consideration to pages which are more useful to the user. This method is better in precision and coverage rates than the conventional association rule based recommendation [8]. The drawback with this approach is that the idle time of the user on a web page cannot be differentiated from the actual time spent by the user on a page. A recommendation method was proposed to be applied to web log mining by integrating user clustering and association rule mining techniques. The precision by this method was better compared to association rule based recommendation system. But there is no much improvement in coverage ratio and speed [7].

Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. Semantic web makes the web contents understandable not only by human but also by the machine. Machine processable information can provide more relevant results and can improve the precision and recall. To achieve this, the content of the semantic web are mostly represented by ontology, XML, RDF and meta-data. Ontology Learning is a method for extraction of semantics from the Web in order to create ontology. Machine learning techniques were used to improve the ontology engineering process. Semantic Web Mining improves the results of Web Mining by exploiting the new semantic structures in the web [10].

3. SYSTEM DESIGN

The overall system architecture of the proposed recommendation system is the Fig. 1.

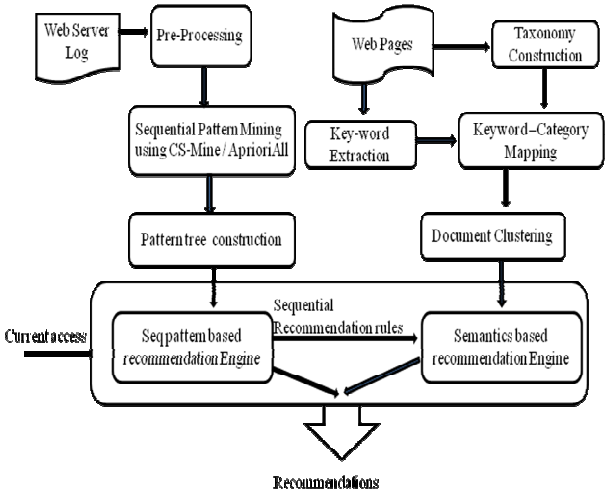

Fig. 1 System Design

System Description

The dataset used is the web server log file (web usage data) of the website http://cs.annauniv.edu and the web pages (web content data) of the same site. Each module in the designed recommendation system is explained as follows.

3.1 Pre-processing

The objective of pre-processing is to reduce the size of the web usage data (web server log) and to increase the quality of the data that will be suitable for mining. Web usage data contains lot of noise. The log file contains records for the requests to images, multimedia files or script files. Some records are due to the requests made by the web crawlers. Log file also contain the requests that are not processed. Those records are not necessary for analysis and hence they are eliminated. The un-processed requests can be identified using the http status code. If the requested page has images many records are created for a single web page request. Such repetitive records are pruned. A web server would receive request from multiple users at same time. The records for the request from different users will be interleaved. The session for each user has to be identified. A user can be identified as a combination of the IP address and the agent. The list of web pages requested by a user forms a web access list which is input to the sequential pattern mining algorithms.

3.2 Sequential Pattern Mining

The sequential pattern mining is the mining of frequently occurring patterns ordered by time. Using sequential pattern mining, one can identify the paths that users frequently follow

on a web site and hence it increases the prediction rate. Sequential pattern mining is well suited for log study due to the sequential nature of web users' activity. In this paper the frequent sequence patterns generated by two sequential pattern mining algorithms AprioriAll and CS-Mine are evaluated.

Apriori-All is similar to the well known Apriori algorithm but the difference being that the Apriori-All considers the ordering of the items in a transaction where as Apriori doesn't take into account the order of items in a transaction. Hence Apriori-All can be said as the sequential version of Apriori [3]. Conditional Sequence Mine algorithm works on the conditional sequence of each frequent pattern given a sequence database where each sequence is a list of transactions ordered by transaction time as an input to the algorithm [1].

3.3 Pattern Tree Construction

A pattern tree is used to store the sequential patterns compactly. A single scan of sequential access patterns generated by the AprioriAll or Conditional Sequence mine algorithms is necessary to construct a pattern tree. A pattern tree is based on trie data structure. Trie is used for storing strings to enable fast pattern matching. The root node of the pattern tree is a dummy node. All other nodes have a web access and its support. All the sequential access patterns in the pattern tree can be visited by following the path starting from the root node of the tree.

3.4 Sequence Pattern Based Recommendation Engine

This component searches for the best matching access path in the Pattern-tree for the given access sequence. The given access sequence is matched with the nodes in the pattern tree and the web accesses in the child nodes are generated as recommendations. The recommendations generated are in the descending order of the support of the web accesses i.e., the most frequent access is suggested first.

The suffix sequences of the current access sequence will be considered when the matching path of the whole access sequence cannot be found. Search will be performed on the matching path based on the same access sequence by removing the first item repeatedly until a matching path is found or when no more items can be removed from the access sequence. The length of the longest path in the Pattern-tree is the depth of the Pattern-tree. The matching path will not exist when the length of the current access sequence is longer than the depth of the Pattern-tree. Therefore, some initial items can be removed to make the current access sequence shorter than the depth of the Pattern-tree before the sequence matching process.

The recommendation rules generated by this engine are later used by semantic based recommendation engine to generate recommendations based on semantics of the web contents.

3.5 Keyword Extraction

The web pages are parsed, the tags are removed and the web contents are extracted. A text document is created corresponding to each page. Most common words (stop words) are removed from each document. The significant keywords for each document are identified using TF*IDF. TF*IDF is a statistical measure used to evaluate how important a word is to a document in a collection. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

3.6 Taxonomy Construction

The taxonomy is a concept hierarchy that provides a means for designing enhanced searching, browsing and filtering systems. A domain specific taxonomy is constructed using XML. The keywords extracted in the previous step have to be mapped with the elements in the taxonomy to get the categories. Hence the taxonomy has a greater influence on the outcome of the mapping process. For this purpose the taxonomy has to be semantically related to the contents of the website.

3.7 Mapping Keywords and the Categories

Keywords are the representatives of the contents of the web pages. These keywords are mapped with the categories in the taxonomy using the thesaurus (Wordnet). If a keyword exists in the taxonomy, it is included. If it doesn't exist, a closest category in the taxonomy is found by making use of the thesaurus. In this paper the closest category in the taxonomy is found using Jiang and Conrath similarity measure. Now the documents are represented as the categories in the taxonomy.

3.8 Document Clustering

As the web pages are now represented as the categories in the taxonomy, they are clustered based on the similarity between the categories in the taxonomy. Clustering based semantics would aid the recommendation generation easier. DBSCAN is a density based clustering algorithm used for document clustering.

3.9 Semantic Based Recommendation Engine

The recommendation rules generated by the sequential pattern based recommendation engine and the document clusters are provided as input to the semantic based recommendation engine. Every document is already assigned to a relevant cluster. Given a recommendation rule, all the URLs in the cluster in which the URL in the right hand side of the rule belongs are generated as recommendations [10].

4. RESULTS

Data set used is the log file of the website http://cs.annauniv.edu. The log file is assumed to be static for

experimental purpose. Number of records in the log file is 140283. After preprocessing the size of the records were reduced to 15% of the initial size. Web access lists are extracted from the preprocessed records. The web access lists generated are used by the sequential pattern mining algorithms Apriori-All and CS-Mine and their performance is evaluated.

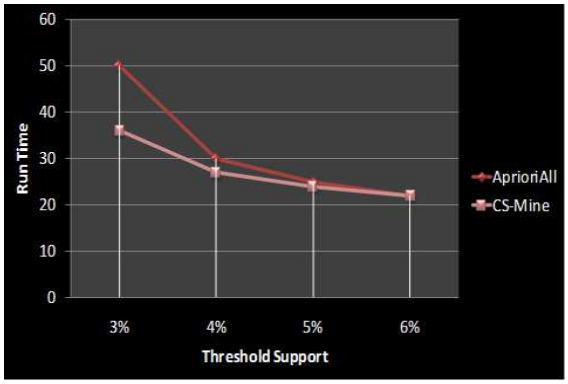

Fig. 2 Run time vs. Threshold Support

Fig 2 shows that the run time of CS-Mine is lesser compared to AprioriAll. This is because AprioriAll generates huge set of candidate sequences and needs many scans of the sequence database. After a certain support threshold, say 6% the run time of both the algorithms becomes equal. The run time of both the algorithms decreases as the support threshold increases.

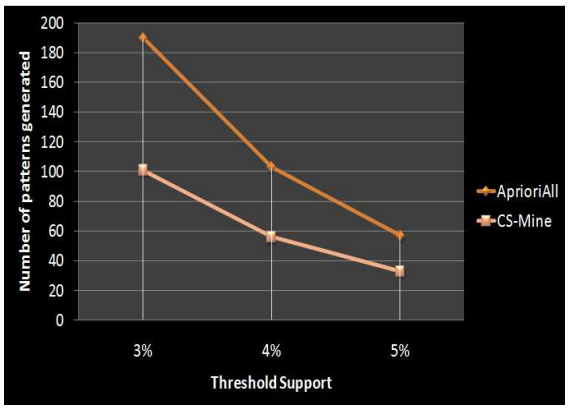

Fig. 3 Number of patterns generated vs. threshold support

Fig 3 shows that the number of patterns generated by AprioriAll is greater than that of the CS-Mine algorithm. Though the number of patterns generated by Apriori-All was greater than that of the CS-Mine, recommendation engine generated same recommendations using both the algorithms at all the threshold support. Hence CS-Mine which is lesser time consuming by generating fewer patterns is found to be efficient in terms of recommendation generation also.

Precision is an evaluation measure used to find how probable a user will access one of the recommended pages.

Precision = Number of correct recommendations / Total number of recommendations

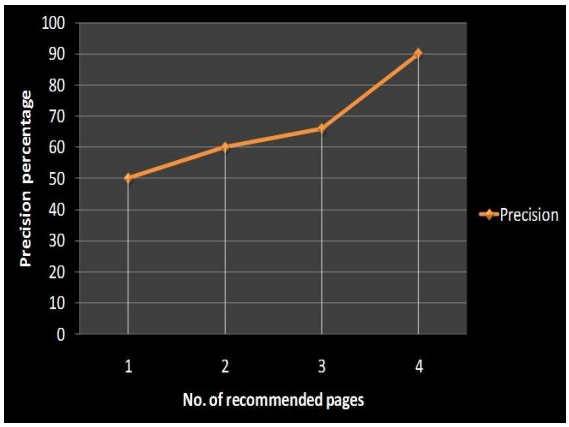

Fig. 4 Precision percentage vs. No. of pages recommended

Fig 4 shows that the precision of the recommendation system increases as the number of the recommendations generated increases. Since many recommendations will be generated at lower support threshold, the precision will be high at low threshold support.

CONCLUSIONS

In this paper the recommendations are provided by exploiting two sequential pattern mining algorithms – AprioriAll and CS-Mine and it has been shown that the CS-Mine algorithm is efficient in terms of time, generating less number of patterns. The recommendation obtained using both the algorithms are same. Rules generated by the sequential pattern based engine are then used to provide semantic based recommendations.

Sequential pattern based recommendations are generated by matching the current access of a user with paths in the pattern tree. During the pattern matching, only few paths in the pattern tree are matched. Hence there is a reduction in time as not all the paths in the tree are traversed. Also there is a reduction in space as all the sequences are compacted in this pattern tree and hence the sequences with same prefixes share common sub paths. The time consuming part of this approach is the pattern tree construction. In real time the pattern tree construction need not be performed for every request and hence it doesn't greatly affect the time when providing suggestions. One of the challenging problems in this recommendation system is when a page is newly visited and is not in the pattern tree, no recommendations are provided initially. When the pattern tree is updated with the new access sequences then the recommendations will be generated.

REFERENCES

[1]. Xiaogang Wang; Yan Bai; Yue Li: "An Information Retrieval Method Based On Sequential Access Patterns" Wearable Computing Systems (APWCS), 2010 Asia-Pacific Conference, April 2010, pp: 247 - 250

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

[3]. WANG Tong, HE Pi-lian: "Web Log Mining by an Improved AprioriAll Algorithm" World Academy of Science, Engineering and Technology, April 2005

[4]. Baoyao Zhou, Siu Cheung Hui, Kuiyu Chang: "An Intelligent Recommender System using Sequential Web access patterns" Cybernetics and Intelligent Systems, 2004 IEEE Conference, 2004, pp: 393-398

[5]. Yi Dong, Huiying Zhang, Linnan Jiao: "Research on Application of User Navigation Pattern mining recommendation" Intelligent Control and Automation, 2006. The Sixth World Congress, 2006, pp: 6106 - 6110

[6]. Forsati, R: Meybodi, M.R.; Rahbar, A: "An Efficient Algorithm for Web Recommendation Systems" Computer Systems and Applications, 2009.IEEE/ACS International Conference, 2009, pp: 579 - 586

[7]. Jaideep Srivastava , Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan: "Web Usage Mining: Discovery and Applications of Usage patterns from web data" ACM SIGKDD Volume 1, Issue 2, Jan 2000, pp: 12-23

[8]. Bettina Berendt, Andreas Hotho, and Gerd Stumme: "Towards Semantic Web Mining", Horrocks and J. Hendler (Eds.): ISWC 2002, LNCS 2342, 2002, pp. 264–278

[9]. Lappas, G: "An Overview of Web Mining in Societal Benefit Areas" 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services, 2007, pp: 683-690

[10]. M. Eirinaki, M. Vazirgiannis, I. Varlamis: "SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process" SIGKDD '03, August 24-27, 2003

BIOGRAPHIE:

Geethapriya Uvaraja, Post-Graduate from Anna University, Chennai