International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
International Conference & workshop on Advanced Computing 2013 (ICWAC 2013) – www.ijais.org

A Novel Approach for Web Recommendation System based on Sequential Access Patterns

Namdeo Badhe
Vidyalankar Institute of Technology
Wadala(E), Mumbai

Prof.Kavita Shirsat
Vidyalankar Institute of Technology
Wadala(E),Mumbai

ABSTRACT

In recent years, there is a tremendous growth of web applications in various fields. People are using World Wide Web for different needs such as to retrieve information for e-commerce, e-governance and many more. Day by Day information available on World Wide Web is increasing exponentially, it becomes much more difficult for web users to access relevant information efficiently. Hence, it is required to develop a good recommendation system which can used to reduce the workload of web users. Usually, most of the web users surf web sites in particular patterns. If the analysis of these usage patterns is done properly then it will be useful for recommendation system to determine which web pages are most likely to be accessed by the user in the future. All users' web access activities of a web site are stored as web log file into the web server. In the proposed system consists of sequential access pattern mining with modified CS-mine algorithm. The mined patterns are used for matching and generating web links for online recommendations.

Keywords

Web Access Pattern, Algorithm, Mining.

1. INTRODUCTION

As the scale of the Internet are getting larger and larger in recent years, it has become much more difficult to access relevant information from the Web .To solve this problem, many web recommender systems[1] are constructed which automatically selects and recommends web pages suitable for user's favor.

Various traditional techniques such as collaborative filtering and hybrid content-based collaborative filtering approaches have been developed for supporting web recommendations. However, such approaches suffer from a major drawback in which most users surf websites anonymously via a proxy, and their identities are hidden and difficult to get. More recent techniques are 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. Promising web usage mining techniques such as association rule mining and clustering have been applied for web recommendations.Usually, most of the web users surf web sites in particular patterns. If the analysis of these usage patterns is done properly then it will be useful for recommendation system to determine which web pages are most likely to be accessed by the user in the future. All users' web access activities of a web site are stored as web log file into the web server. Different from the majority of the existing web recommendation techniques, The proposed web recommendation system uses a sequential pattern mining technique. Unlike clustering and association rule mining, sequential pattern mining algorithms also consider the sequential characteristic of access patterns, which is very suitable for predicting the next web pages.The proposed algorithm known as MCS Mine which is modification of CS Mine algorithm [2].The rest of this paper is organized as follows. Section 2 consists of literature survey on various sequential access pattern mining techniques. Section 3 presents the system architecture of the proposed web recommendation system. Section 4 discusses MCS Mine algorithm and compared with CS Mine algorithm. Finally conclusion is drawn in section 5.

2. LITERATURE SURVEY

Web recommendation system predicts the information needs of users and provides them with recommendations to facilitate their navigation. Various techniques such as web content mining and web structure mining approaches have been developed for supporting web recommendation system. However, such approaches suffer from a major drawback in which most users surf websites anonymously via a proxy, and their identities are hidden and difficult to get. Web usage mining [1] is one of the main approaches to study the users' navigation patterns and their use of web resources such as web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions, cookies, user queries, and bookmark data, in order to understand and better serve the needs of Web-based applications.

Web recommendation systems based on Web usage mining techniques have strengths and weaknesses. Promising web usage mining techniques are association rule mining [2] and clustering [3].

Association rules [12] are introduced in 1993.Association rule mining is a 2-step process. In step 1,it finds set of largest no. of items which are frequent i.e above threshold value .and then in step 2,it finds association rules between frequent items. R.Agrawal has introduced basic Apriori algorithm for finding association rules. These rules can be used along with search engine for suggesting the items to new users who are likely to visit the shopping web sites.but, basic Apriori algorithm considers only non sequential kind of data.

Clustering techniques make groups of items based on similarity. Clustering techniques identify customer's preferences and make cluster of preferred items and is useful for recommend nation. This is also called as collaborative filtering based recommendation .clustering techniques are divided into various categories like partitioned based, Centroid based, hierarchical based etc.

The concept of sequence data mining was first introduced by Rakesh Agrawal and Ramakrishnan Srikant in the year 1995.Sequential pattern mining is the extension of Frequent Itemset mining which finds the relationships between occurrence of sequential events, to find if there exist any specific order of the occurrences. Sequential patterns indicate the correlation between transactions while association rule represents intra transaction relationships. From a book store's transaction database history, we can fine the frequent sequential purchasing patterns, for example 80% customers who brought the book Database Management typically bought the book Data Warehouse and then brought the book Web Information System with certain time gap. In the above example, all those books need not to be brought at the same time or consecutively, the most important thing is the order in which those books are brought and they are bought by the same customer.

Sequential pattern mining techniques can be used in personalization (like Personalized Customer Experience in B2C E-commerce-Amazon.Com, Personalized Portal for the Web MyYahoo), system improvement, and site modification and so on. In the past few decades, web recommendation systems(WRS) are highly utilized for information filtering, prediction, personalized service, and for customer service.

For WRS, web log file is used as a input. Whenever web user access a website, an entry is made automatically in the web logs by the web server. each line of a web log file consists of the following key information: IP address , user ID, date-timestamp, requested URL, and HTTP status code etc. There are various formats of web log file like NCSA, W3C Extended, WebSphere etc.

Sequential pattern mining can be broadly classified into two main categories

- Apriori Based
- Pattern Growth Based

The apriori property states that all nonempty subsets of a frequent itemset must also be frequent. Agrawal have proposed the Apriori[51],AprioriAll algorithms for finding frequent access patterns from sequential database. the Agrawal and Shrikant have proposed Generalized Sequential Pattern (GSP) algorithm [3][7] for web log analysis. For the longest length of frequent sequences in the database, GSP algorithm scans the original database multiple times,and it will likely generate a large number of candidate sequences. Basic Apriori[3] algorithm has been extended for finding frequent access patterns from sequential database.

Later Agrawal and Shrikant have introduced a new projection-based algorithm for mining sequential patterns called the PrefixSpan. It uses the frequent sequence trellis to partition the database. It traverses the frequent item lattice level-by-level in depth-first order.

Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu have proposed a algorithm known as WAP-Mine[5], is basically an extension of PSP algorithm for Sequential pattern mining. The algorithm has two steps, construction of a WAP tree and mining of constructed WAP-Tree. In first step, the preprocessed WASD is scanned twice, generates frequent web access sequences and stores in a compact structure called as WAP-Tree. In second step, it recursively mines a WAP-tree by constructing intermediate WAP-trees and generates sequential access patterns.

C.I. Ezeife, Yi Lu and Yi Liu [6] have proposed a PLWAP algorithm that uses a preorder linked, position coded version of WAP tree and eliminates the need to recursively re-construct intermediate WAP trees during sequential mining as done by WAP-Mine.

B.Y. Zhou and S.C. Hui have proposed a CS-Mine algorithm which is efficient than both WAP-Mine and PLWAP-Mine by eliminating the need for the reconstruction of intermediate conditional WAP-trees. However, it also needs to construct an initial WAP-tree from the web access sequence database firstly.

The Pattern Growth based algorithms use a compact data model, called WAP tree, which stores the web access patterns, and an efficient approach for finding frequent patterns. The existing web recommendation system uses conditional sequence mine algorithm in which WAP tree is constructed initially. Though, WAP tree is compressed data structure, time required for WAP tree creation increases proportionally if average length of web access sequences increases. The WAP tree structure explores maximal sharing of common prefix paths in the tree construction. If web access sequences have very less common prefix paths, thus the WAP tree generated will be huge. To find prefix sequence bases (PSB) of each frequent token, the whole WAP tree is traversed each time in bottom-up approach. Large number of frequent tokens increases the WAP tree traversal time.[

The proposed system is based on the Modified CS-Mine algorithm which eliminates the need for the construction of the initial WAP-tree. MCS-Mine algorithm considers suffix sequence bases (SSB) of each frequent token instead of prefix sequence.[8][9][10].

3. SYSTEM ARCHITECTURE

Figure 1 shows the block diagram of proposed web recommendation system

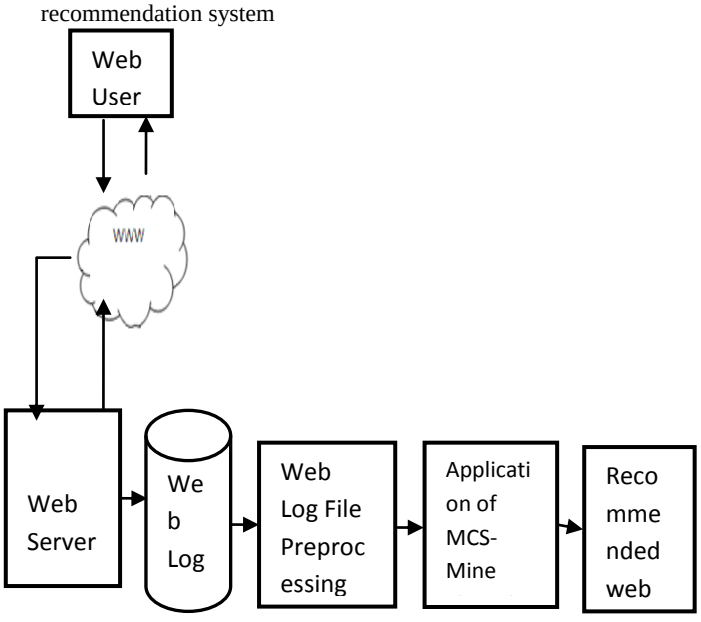

Figure 1 System Architecture

The proposed system offers a list of recommended web pages to the active user without the need for a accessing history or a minimal number of visited pages.Web user can be any one who is accessing online web site. Web user might be the naive user or the trained one. When a web user accesses a particular web page, an entry is made in the web log file by web server. Web log file consist of the information related to the user's web access such as the client IP address, request time, requested URL, userID, HTTP status code, etc.

The raw web log file contains many unwanted information for the mining process. Therefore, web log file has to be cleaned which is done through web log file preprocessing. The input to this block is raw web log file and output is web access sequence database (WASD).

After preprocessing ,MCS-Mine algorithm is applied for finding sequential patterns which are stored in text file and can be used as recommended list.

The figure 2 shows Web Log Preprocessing in detail.

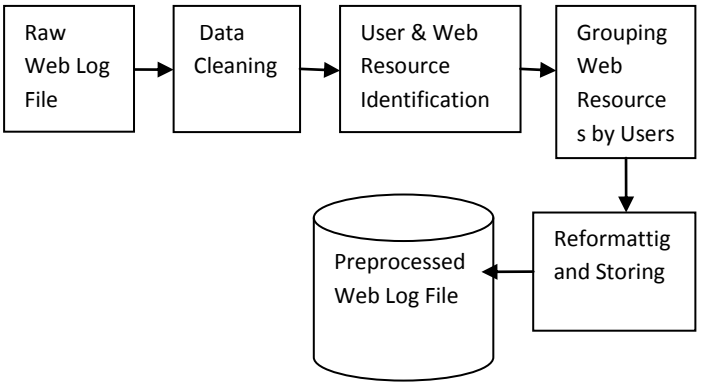

Figure 2 Web Log Pre-processing

The sample web log file consists of 5000 web log records are collected from NASA Kennedy Space Center from the IIS Web Server of July 1995. During data cleaning, irrelevant records are eliminated. The records having filenames suffixes of GIF, JPEG, and CSS etc. and status code over 299 and below 200 are removed and an intermediate web log file is created. In User & Web Resource identification step, unique IP addresses as different users and unique web pages are filtered out.
Then, users and their accessed web pages are combined together and final web access sequence database (WASD)is created by mapping web page as unique number.

The sample raw web log file is:

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GE
/history/apollo/ HTTP/1.0" 200 6245

unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GE
/shuttle/countdown/ HTTP/1.0" 200 3985

199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GE
/shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 20
4085

burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GE
/shuttle/countdown/liftoff.html HTTP/1.0" 304 0

199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GE
/shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0
200 4179
```

The combination of user & web resources of intermediate web log file1 is as follows:

```
199.120.110.2 /shuttle/missions/sts-73/mission-sts
73.html?/shuttle/missions/sts-73/mission-sts-73.html

burger.letters.com
shuttle/countdown/liftoff.html?/shuttle/countdown/liftoff.h
tml?/shuttle/count
```

The final WASD file is:

```
199.120.110.21 128 128

burger.letters.com 69 69 69 69 69 69 69 69 69 69 69 69 69
```

Pattern generation module uses MCS-Mine algorithm and generates sequential access patterns (SAP).

The Figure 3 shows general flow of the proposed Modified CS-Mine algorithm. It consists of following steps

Step1: Read WASD and find SSB of each token.
Step2: Construct Event queues for SSB (token).
Step3: Test Event queues of each frequent token.
Step4: construct Sub-SSB.
Step5: Recursive Mining for Sub-SSB.

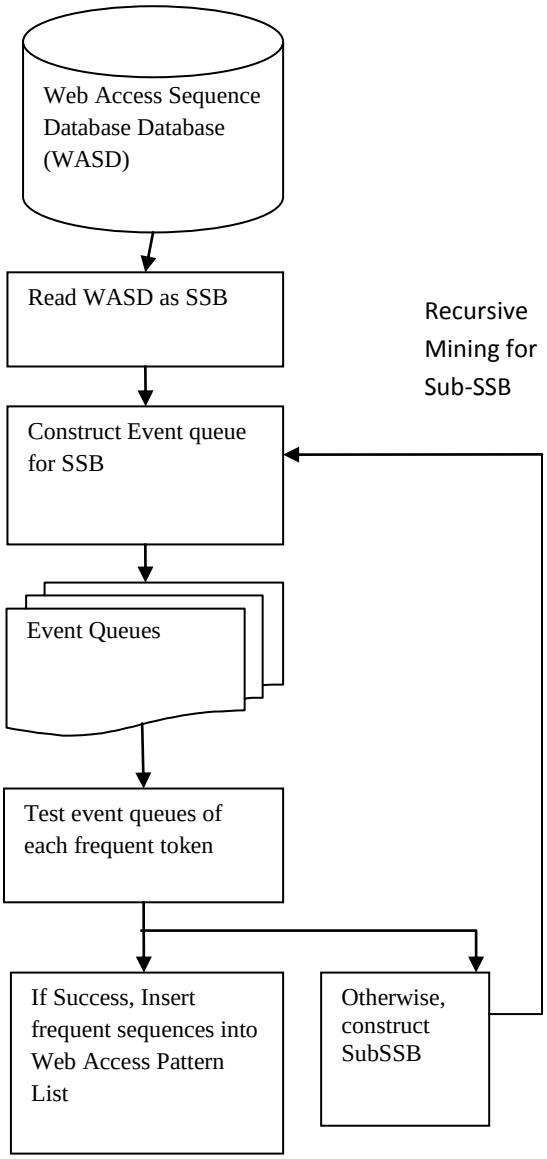

Figure 3: MCS-Mine Algorithm Flow

The Proposed MCS-Mine algorithm is directly based on the suffix sequence base (SSB) of each frequent token. It eliminates the need for the construction of the initial WAP-tree.

The detail explanation of the proposed MCS Mine algorithm is as following.

Algorithm MCS-Mine:
Step 1 Read WASD and for each frequent token, find suffix sequences base (SSB).
Step 2 Use ConstructEQ to construct event queues for SSB (token).
Step 3 Use TestSSB to test single sequence for final-SSB(token).
If test is successful,
Create frequent sequence FS = token+SingleSeq and insert into WAP List.
Else,
For each token t in Header Table of final-SSB, use ConstructSubSSB to construct final-SSB (token+t).
Set token = token+t and recursively mine final-SSB (token) from step 2.
Step 4 Return WAP List.

Algorithm ConstructEQ:
Step 1 Read SSB (token).
Step 2 Find conditional frequent token from SSB (token) and store in a header table.
Step 3 Create Event Queues for each token of a header table.
Step 4 Remove non-frequent tokens from event queues and make a final-SSB.

Algorithm TestSSB:
Step1 Initialize the single sequence of final-SSB (token),SingleSeq = 0.
Step 2 Read final-SSB.
Step 3 If final-SSB (token) is empty then test is successful and return SingleSeq = 0.
Step 4 for each sequences of final-SSB (token) do
(a) If all the ith items in each sequence are the same token t. And if total count of these tokens >= MinSup create a new item t with the count and insert it into SingleSeq.
(b) Otherwise, return false and SingleSeq = 0.
Step 5 Test is successful and returns SingleSeq.

Algorithm ConstructSubCSB:
Step 1 Initialize final-SSB(token+t) =0
Step 2 For each item in ei-queue of PSB(token+t), insert its suffix sequences into final-SSB(token+t).
Step 3 Return final-SSB (token+t).

After mining process, two text files are created named "Finalpattern.txt" for frequent access patterns and "URLOnly.txt" for Unique URLs. When a web user access a web page, if web page is present in "URLOnly.txt" then Recommendation module displays a list of recommended web page sequences. And, user will get knowledge about important web pages.

4. Experimental Results:

The Microsoft anonymous web dataset is used for the experimental performance. This dataset is cleaned and contains logs of visited pages by the web site users and has a total of 32,711 web access sequences, with each sequence containing from 1 upto 35 page references from a total of 294 pages. The data is in an ASCII-based sparse-data format called "DST". For finding sequential patterns, atleaset 2 length sequences are required. hence, the dataset is considered with sequences having length>=2.
To measure the performance, two experiments had been conducted. In the first experiment, the scalability of the two algorithms had measured with respect to different support thresholds. This experiment used different support thresholds from 5 to 100. The experimental result shows that the run time of the CS-mine algorithm slightly increases as compared to MCS-Mine algorithm, when the support threshold decreases.

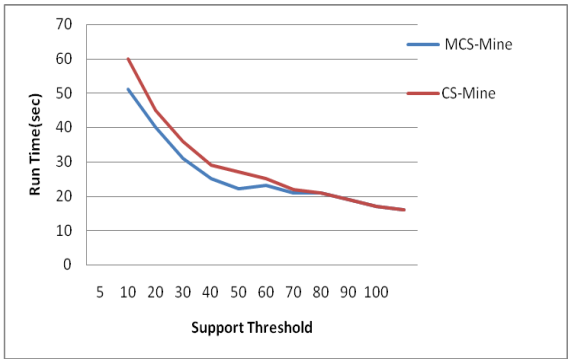

Figure 4 Scalability Measurement with different Support Thresholds

In the second experiment, the scalability of the two algorithms is compared with respect to different sizes of the web access sequence database. The experiment used a fixed support threshold i.e.10 with different numbers of web access sequences varied from 4,000 to 24000.

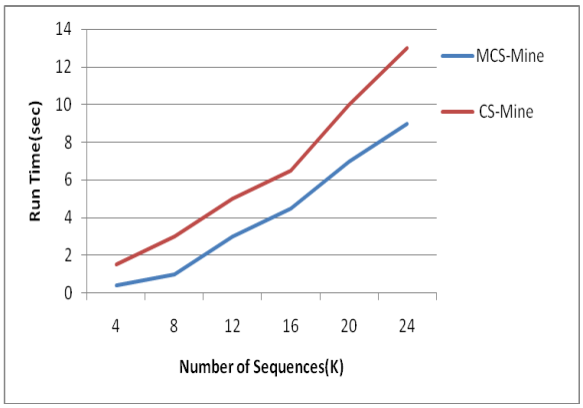

Figure 5 Scalability Measurement with number of input sequence

5. CONCLUSION AND FUTURE SCOPE:

The web recommendation systems recognizes interest of users by considering previous access record through web log file and provide an appropriate list of recommended web pages. To this end, the proposed system uses MCS-Mine algorithm for finding frequent patterns after preprocessing a web log file.

The performance of the proposed MCS-Mine algorithm has been evaluated with existing CS-Mine algorithm. Experimental results shown that the MCS-Mine algorithm performs better than CS-Mine algorithm, especially when the support threshold becomes small and the number of web access sequences gets larger.

The web recommendation system provides with web log file and then display the list of recommended web pages as output. The proposed MCS-Mine algorithm constructs SSBs as suffix sequences of every frequent token from header table and generates Event Queues for SSB.

Instead of performing physical projection, one can register the index (or identifier) of the corresponding sequence and the starting position of the suffix sequences from the web access sequence database (WASD).Then, a physical projection of a sequence is replaced by registering a sequence identifier and the projected position index point. This can avoids physically copying suffix sequences and can be efficient in terms of both running time and space.

The proposed Recommendation system can be extended to cope up with mobile devices as more users are surfing internet on mobiles.

6. REFERENCES

[1] Baoyao Zhou ,Siu Cheung Huia and Kuiyu Chang" An Intelligent Recommender System using Sequential Web Access Patterns" Proceedings of the 2004 lEEE Conference on Cybernetics and Intelligent Systems Singapore, 1-3 December, 2004.

[2] B.Y. Zhou, S.C. Hui, and A.C.M. Fong. "CS-mine: an efficient WAP tree Mining for Web Access Patterns mining for web access patterns", the 6th Asia Pacific Web Conference (APWEBW), Hangzhou, China, April 14/17/2004, p523-532.

[3] "Data Mining Concepts and Techniques" by Jiawei Han and Micheline Kamber ,second edition ,Morgan Kaufmann publisher.

[4] Nasser Ahmed Sajid,Salman Zafar "Sequential Pattern Finding :A Survey" 978-1-4244-8003 , 2010 IEEE

[5] Jian Pei,Jiawei Han, "Mining Access Patterns Efficiently from Web Logs" Code"–August 21,2005,Chicago, Illinois, USA.ACM.

[6] C.I.Ezeife "PLWAP Sequential Mining: Open SourceCode"–August 21,2005,Chicago, Illinois, USA.ACM.

[7] Xiaogang Wang,Yan Bai and Yue Li "A GSP-based Efficient Algorithm for Mining Frequent Sequences" Asia-Pacific Conference on Wearable Computing Systems 978-0-7695-4003-0/10 IEEE 2010.

[8] Mohammed Zaki "SPADE An Efficient Algorithm For Mining Frequent Sequences", Machine Learning, Volume 42, Issue: 1, Publisher: Springer, Pages: 31-60,2001.

[9] Ayres,J.,Flannick, J.,Gehrke, J., And Yiu. "Sequential Pattern Mining Using A Bitmap Representation", In Proceedings of the 8th Acm International Conference On Knowledge Discovery And Data Mining Transaction Data Analysis 2002.

[10] Jiawei Han, Jian Pei and Qiming Chen "FreeSpan: Frequent Pattern-Projected Sequential Access pattern Mining" In proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 355-359, 2000.

[11] C.I.Eeife "Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree" Data Mining and Knowledge Discovery, Business Media, Springer Science 10, pp5-38, 2005.

[12] R.Agrawal and A.swami "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD international Conference on Management of Data (ACM SIGMOD'93),Washington,USA,May 1993