WEB LOG MINING USING MULTI ITEM SEQUNTIAL PATTERN BASED ON PLWAP

Jaymin Desai[1], Mrs. Risha Tiwari[2]
[1]Post Graduate Student, [2]Professor,
Dept. of Computer Engg., Hasmukh Goswami Collage of Engineering, Ahmedabad, Gujarat, India.

Abstract: Web Log Mining (WLM) is the process to extract information from the Web Log data. Web logs records user activities and website resources usage when user browses the website. Sequential pattern mining (SPM) is an important data mining task of discovering timerelated behaviors in sequence databases. SPM technology has been applied in many domains, like web-log analysis, the analyses of customer purchase behavior, process analysis of scientific experiments, medical record analysis etc. Using SPM methods for web log mining we can propose a good recommendation for web. It can be more beneficial to find the sequence of users' behavior in web usage mining. System generates pattern by assuming that user access only one page at a given point in time. In actual system when user searches for any item he may load multiple pages for the same at a given point in time. By considering all the pages for the same parent page we can generate more useful patterns.

Keywords: Sequential pattern mining, PrefixSpan, PLWAPAlgo.

I. INTRODUCTION

Web Log Mining (WLM) is the process of extracting useful information from server logs. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.Web Log Mining (WLM) is the process to extract information from the Web Log data. Web logs records user activities and website resources usage when user browses the website. They are one of the primary sources that can be analyzed to mine valuable knowledge. Web log mining may reveal interesting and unknown knowledge about both the user and website. Such knowledge can be used by different special purpose to perform task such as analyzing system performance, understanding internet traffic, improving system design, modeling user behavior and business intelligence.Sequential Pattern Mining (SPM) is an important data mining task of discovering Time - related behaviors in sequence databases. Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples wherethe values are delivered in a sequence .The concept of sequence Data Mining was firstintroduced by Rakesh Agrawal and RamakrishnanSrikant in the year 1995. SPMtechnology has been applied in many domains, like web-log analysis, the analyses of customer purchase behavior, process analysis of scientific experiments, medical record analysis etc. Sequential pattern mining discovers frequent subsequences as patterns in a sequence database. A sequence database stores a number of records, where all records are sequences of ordered events, with or without concrete notions of time. An example sequence database is retail customer transactions or purchase sequences in a grocery store showing, for each customer, the collection of store items they purchased every week for one month.With using SPM methods for web log mining we can propose a good recommendation for web. It can be more beneficial to find the sequence of users' behavior in web usage mining. In sequential pattern mining for web WLM technique is very useful. By extracting the information from the web logs which are nothing but the activities of user. Using web log mining with SPM technique it helps to find frequent pattern and better recommendation. WLM is an important application of sequential pattern mining concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items.In reality when user search for particular keyword or system he may load more than during the others are loading in specific time interval. And it may or may not helpful for the user. Existing systems do consider only single page at a given point in time with the assumption that a web user can physically access only one web page at any given point in time. When user searches for any content he may load other pages while other is loading which may be useful. We propose a system in which we take multiple web pages into account for recommendation. We consider those pages which were surfed together by same user for the same purpose. So we may provide better recommendation with this approach.

II. LITERATURE SURVEY

Web usage mining is an important application of sequential pattern mining concerned with finding user navigational patterns on the world wide web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items, with the assumption that a web user can physically access only one web page at any given point in time. If a time window for access is considered that may allow a web user to browse acollection of web pages over a specified period of time, it then reverts back to a general sequence database.Sequential pattern mining can be classified into three main categories, namely, apriori-based, pattern-growth, and early-pruning with a fourth category as a hybrid of the main three. That investigation of sequential pattern-mining algorithms in the literature shows that the important heuristics employed include the following: using optimally sized data structure representations of the sequence database; early pruning of candidate sequences; mechanisms to reduce support counting; and maintaining a narrow search space. The quest for finding a reliable sequential pattern-mining algorithm should take these points into consideration.Improving the efficiency and representation or managing the database, so based on these criteria’s sequential pattern mining is classified into two major groups, Apriori Based and Pattern Growth based algorithms. Comparative analysis of various mining algorithms, it is clear that pattern growth based algorithms are more efficient with respect to running time, space utilization and scalability.

III. ANALYSIS

Web Log Mining (WLM) is the process to extract information from the Web Log data. Web logs records user activities and website resources usage when user browses the website. Sequential Pattern Mining (SPM) is an important data mining task of discovering time -related behaviors in sequence databases.SPM technology has been applied in many domains like web-log analysis, the analyses of customer purchase behavior, process analysis of scientific experiments, medical record analysis etc. There are so many algorithms of Sequence pattern Mining and some of the algorithms are specially for web log mining or single item set sequence or a variation of web log mining namely WAP-Mine algorithm, PLWAP algorithm and LAPIN algorithm.

Key features of different techniques of sequential pattern mining –

Apriori based methods :

- Breadth first search :Apriori-based algorithms are described as breath-first (level-wise) search algorithms because they construct all k-sequences together in each kth iteration of the algorithm as they traverse the search space.
- Generate and test : Algorithms that depend on this feature only display an inefficient pruning method and generate an explosive number of candidate sequences, consuming a lot of memory in the early stages of mining.
- Multiple database scan : It is a very undesirable characteristic of most apriori-based algorithms. Requires a lot of processing time and I/O cost. A solution to this limitation is to scan the database only once or twice to create a temporary data structure, which holds support information used during mining.

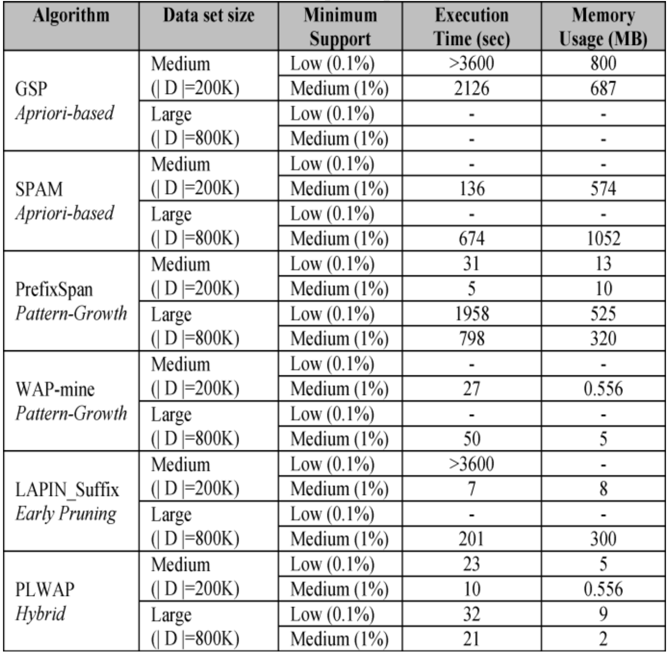
Pattern growth based methods :

- Sampling / Compression : Compression is used in the data structure that holds the candidate sequences, usually a tree. Shared prefixes in sequences are represented in the tree by one branch; each node represents an item in the sequence alongside with its support count.
- The problem with sampling is that the support threshold must be kept small, which causes a combinatorial explosion in the number of candidate patterns.
- Candidate Sequence Pruning : Pattern-growth algorithms that can prune candidate sequences early display a smaller search space and maintain a more directed and narrower search procedure. Prefix span uses direct antimonotic app of apriori property to prune candidate sequence along with projected database. PLWAP has a position-coded feature that enables it to identify locations of nodes relevant to each other as a look-ahead capability and to prune candidate sequences early in the mining process. WAP-Mine - Claims to b a memory only algorithm.
- Search Space Partitioning : It allows partitioning of the generated search space of large candidate sequences for efficient memory management. WAP-mine and PLWAP handle sub trees of a tree-projection structure recursively. Once the search space is partitioned, smaller partitions can be mined in parallel. Free span uses projected database to generate database annotations that guide the minimum process to find frequent pattern faster.
- Tree Projection : Here algorithms implement a physical tree data structure representation of the search space, which is then traversed breadth-first or depth-first in search of frequent sequences. WAP-mine uses the WAP-tree which is generated in only two scans of the sequence database and is used instead of the database to store candidate sequences. FS-Miner [El-Sayed et al. 2004] uses the FS-tree, which allows the mining process to start with 2-sequences immediately from the second scan of the database.
- Depth first Traversal : Reason for including this as a feature on its own because it is very important to have in any algorithm that uses a tree model. It has been stressed a lot and made very clear in several works that depth-first search of the search space makes a big difference in performance, and also helps in the early pruning of candidate sequences as well as mining of closed sequences.
- Suffix / Prefix growth : Algorithms that depend on projected databases and conditional search in trees first find the frequent 1-sequences, and hold each frequent item as either prefix or suffix, then start building candidate sequences around these items and mine them recursively. This greatly reduces the amount of memory required to store all the different candidate sequences that share the same prefix/suffix.
- Memory only : This feature targets algorithms that do not spawn an explosive number of candidate sequences, which enables them to have minimum I/O cost.

Early pruning based methods :

- Support counting avoidance : Several recent algorithms have found a way to compute support of candidate sequences without carrying a count throughout the mining process, and without scanning the sequence database iteratively. It is very important for an efficient algorithm not to scan the sequence database each time to compute support. A sequence database can be removed from memory and no longer be used once the algorithm finds a way to store candidate sequences along with support counts in a tree structure, or any other representation for that matter.

- Vertical projection of db : The mining process uses only the vertical layout tables to generate candidate sequences and counts support in different ways. SPADE uses less memory because the sequence database is no longer required during mining. The amount of computation incurred by bitwise (usually AND) operations used to count the support for each candidate sequence.
- Position Coded : Key idea for early pruning methods. It enables an algorithm to look-ahead to avoid generating infrequent candidate sequences. This feature also plays a major role in PLWAP, making it a hybrid pattern-growth\early-pruning algorithm that outperforms WAP-mine and Prefix Span with low minimum support when there is large amount of mined frequent patterns.



Algorithm	Data set size	Minimum Support	Execution Time (sec)	Memory Usage (MB)
GSP *Apriori-based*	Medium (| D |=200K)	Low (0.1%)	>3600	800
		Medium (1%)	2126	687
	Large (| D |=800K)	Low (0.1%)	-	-
		Medium (1%)	-	-
SPAM *Apriori-based*	Medium (| D |=200K)	Low (0.1%)	-	-
		Medium (1%)	136	574
	Large (| D |=800K)	Low (0.1%)	-	-
		Medium (1%)	674	1052
PrefixSpan *Pattern-Growth*	Medium (| D |=200K)	Low (0.1%)	31	13
		Medium (1%)	5	10
	Large (| D |=800K)	Low (0.1%)	1958	525
		Medium (1%)	798	320
WAP-mine *Pattern-Growth*	Medium (| D |=200K)	Low (0.1%)	-	-
		Medium (1%)	27	0.556
	Large (| D |=800K)	Low (0.1%)	-	-
		Medium (1%)	50	5
LAPIN_Suffix *Early Pruning*	Medium (| D |=200K)	Low (0.1%)	>3600	-
		Medium (1%)	7	8
	Large (| D |=800K)	Low (0.1%)	-	-
		Medium (1%)	201	300
PLWAP *Hybrid*	Medium (| D |=200K)	Low (0.1%)	23	5
		Medium (1%)	10	0.556
	Large (| D |=800K)	Low (0.1%)	32	9
		Medium (1%)	21	2

IV. PROPOSED WORK

We have considered two approaches namely:

- Closed sequence pattern mining
- An improved approach to PLWAP without the assumption that web user can physically access only one web page at given point in time.

We have preferred second approach because in first approach which algorithm we are using it already using the technique of closed sequence pruning.Sequential pattern mining concerned with finding user navigational patterns on the world wide web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items, with the assumption that a web user can physically access only one web page at any given point in time. The propose method is to assign multiple pages which could be accessed in a specific time interval from the same parent node but till now we could not see this consideration, so the probable approach is basically to perform sequential pattern mining using this approach.Web usage mining is an important application of sequential pattern mining concerned with finding user navigational patterns on the world wide web by extracting knowledge from web logs, where ordered sequences of events in the sequence database are composed of single items and not sets of items, with the assumption that a web user can physically access only one web page at any given point in time. We surf internet for any information and mostly we get the information, but when we are searching for any term we usually open multiple pages while others are loading to find the result. So basically the idea is to consider multiple pages instead of single page while mining the weblog data. Instead of considering single web page as a single node we can store multiple pages in a single node which were surfed by user in a specific time interval and to find the same information. By doing this we can find the frequent pattern and also recommend more pages.

Algorithm of MPLWAP tree :
Input: Web Access Sequence
Output: MPLWAP Tree

- It scans the access sequence database first time to obtain all events. Events have support greater than or equal to minimum support.
- Identify the number of events in a single node.
- Each node in a tree registers three information: number of items, pointer to the item , position code. And Item registers information: count.
- Scan the database second time to obtain frequent sequence S.
- Build tree data structure.
- Considering the first event ,increment the count of the same if exist, otherwise
- Check parent page of current node and for the first event is same
- If yes same then put event into current node and increase number of items to the node and also assign count 1 to that event
- Otherwise create new child node and set count of that to one for event and make that event as current node. Also assign position code for that.
- Add current node to the sequence.

This is an existing algorithm in which they mine using the header table. In our proposed algorithm we have two option to generate header table, first one is, with one of the item from the node to the items in the tree and second one is, consider the whole node to generate header table for items.

Implementation-

- Data set Details
- Synthetic Data

Synthetic datasets are generated using the publicly available synthetic data generation program of the IBM Quest data mining project at http://www.almaden.ibm.com/cs/quest/, which has been used in most sequential pattern mining studies. The parameters shown below are used to generate the data sets.

|D| Number of sequences in the database
|C| : Average length of the sequences
|S|: Average length of maximal potentially frequent sequence
|N|: number of events

For example, C10.S5.N2000.D60k means that |C| = 10, |S| = 5, |N|= 2000, and |D| = 60k. It represents a group of data with average length of the sequences as 10, the average length of maximal potentially frequent sequence is 5, the number of individual events in the database is 2000, and the total number of sequences in database is 60 thousand.

Real time data

Real time data obtained from the Website URL: http://www.audiorec.co.uk/

It is basically E-Commerce website. There are total 22 fields and some of the relevant fields are: Date, Time, Serve Name, Server IP, CS-method, CS-uri-query, S-port, C-IP, CS-version, CS-user agent, CS-cookie, CS-Refer and Time taken.We have implemented a proposed algorithm to generate MPLWAP-Tree. The MPLWAP algorithm scans the access sequence database first time to obtain the support of all events in the event set, E. All events that have a support of greater than or equal to the minimum support are frequent. Each node in a MPLWAP-tree registers three information: number of items, pointer to the item, position code. And Item registers information: count. The root of the tree is a special virtual node with an empty label and count 0.

V. CONCLUSION

The algorithm MPLWAP proposed in this thesis, improves on mining efficiency by accommodating multiple pages in a single node instead of single page in single node as done by PLWAP mining algorithm. MPLWAP accommodates multiple web pages in a single node. By considering that the user can surf more than one page in a specific time interval we accommodate multiple web pages in a single node by checking the referred url of the respective web pages. MPLWAP provides multi-item support.Even though the execution time of MPLWAP is higher than PLWAP, the patterns generated from MPLWAP are more than PLWAP mining algorithm. Experiments show that mining of MPLWAP tree gives more patterns than PLWAP tree. Thus if we consider multi-item sequence, we can extract useful patterns from the web log data and it can be useful for web recommendation and personalization.

Future Work

Future work should consider

Applying MPLWAP-tree mining techniques to distributed mining as well as to incremental mining of web logs and sequential patterns.

By using the previous records (patterns) of the user, pattern discovery for particular person can be done.

According to the application, customized recommendation can be done.

REFERENCE

Books

[1] Mining the web Discovering knowledge from hypertext data by SoumenChakrabarti

Web Site

[2] http://en.wikipedia.org/wiki/
[3] http://en.wikipedia.org/wiki/Sequential_Pattern_Mining

Papers

[4] Web usage mining using improved frequent pattern tree algorithm Ashika Gupta , Rakhiarora, Ranjanasikarwar , Neha Saxena.IEEE-2014
[5] A survey on improving the efficiency of prefix span sequential pattern mining algorithm. K Suneetha, Dr. M Usha Rani. IJCCIT - 2014
[6] A Complete PreProcessing Method for Web Usage Mining algorithm. Ankit R Kharwar, Chandani A Naik, Niyanta K Desai IJETAE-2013
[7] An Efficient web Recommender System based on approach of mining frequent sequential pattern from customized web log processing. Manisha Valera, Uttam Chauhan IEEE-2013
[8] PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth .JianPei,Jiawei Han, BehzadMortazavi-Asl,Helen Pinto IEEE-2013
[9] Sequential Pattern Mining Methods: A Snap Shot Niti Desai, Amit Ganatra IOSRJCE- 2013
[10] An Efficient algorithm for data cleaning of log file using file extension. SurbhiAnand,Rinkle Rani Aggarwal IJCA-2012
[11] Efficient preprocessing technique using web log mining Sheetal A. Raiyani, Shailendra Jain IJART-2012
[12] Data Preprocessing Evaluation for Web Log Mining: Reconstruction of activities of a web visitor.MichalMunk, JozefKapusta, Peter Svec ELSEVIER - 2012
[13] A Hierarchical cluster based preprocessing methodology for web usage mining.TasawarHussain, Dr. SohailAsghar, Simon Fong IEEE-2012
[14] Sequence Pattern Mining:Survey and current research challenges. Chetna Chand, Amit Thakkar ,Amit Ganatra IJSCE-2012
[15] Graph based approach for mining frequent sequential access patterns of web pages. Dheerajkumarsingh, Varsha Sharma ,Sanjeev Sharma IJCA-2012
[16] A Taxonomy of Sequential Pattern Mining AlgorithmsNizarR.Mabroukeh, C.I. Ezeife ACM-2010

[17] Fast incremental mining of web sequential patterns with PLWAP tree Yi Lu and C.I. Ezeife Springer-2009

[18] Novel position coded methods for mining web access patterns Wenjiawang and Phuong thanhcao-thai IEEE-2008

[19] Incremental Mining of Web Sequential Patterns Using PLWAP Tree Min Chen and C.I. Ezeife Springer-2005

[20] Position coded preorder Linked WAP-Tree for web log Sequential Pattern Mining Yi Lu and C.I. Ezeife Springer-2003