WEB LOG MINING USING MULTIITEM SEQUNTIAL PATTERN BASED ON PLWAP

Jaymin Desai[1], Mrs. Risha Tiwari[2],
Post Graduate Student, 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 timerelatedbehaviors 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, PLWAP Algo.

I. INTRODUCTION

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, improvingsystem 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 Ramakrishnan Srikant 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 recordanalysis etc. Sequential pattern mining discovers frequent subsequences as patterns in asequence database. A sequence database stores a number of records, where all records aresequences of ordered events, with or without concrete notions of time. An examplesequence 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' behaviorin 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 ofuser. Using web log mining with SPM technique it helps to find frequent pattern and betterrecommendation. WLM is an important application of sequential pattern mining concernedwith finding user navigational patterns on the World Wide Web by extracting knowledgefrom 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 thanduring the others are loading in specific time interval. And it may or may not helpful forthe user. Existing systems do consider only single page at a given point in time with theassumption 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

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 ofvarious mining algorithms, it is clear that pattern growth based algorithms aremore efficient with respect to running time, space utilization and scalability.

Performance comparison of algorithms :
Comparative performance analysis of algorithms from each of the categories. Two datasets were used, a medium size data set described as C5T3S5N50D200K and a large-size data set described as C15T8S8N120D800K. These were run at different minimum support values: low minimum supports of between 0.1% and 0.9% and regular minimum supports of 1% to 10%.
It shows how slow the apriori-based SPAM algorithm could become as data set size growsfrom medium (| D |=200K) to large (| D |=800K), due to the increased number of AND operations and the traversal of the large lexicographical tree; although it is a little faster than PrefixSpan on large data sets due to the utilization of bitmaps as compared to the projected databases of PrefixSpan.

Pattern discovery using MPLWAP mine algorithm :
Input: MPLWAP tree T, header linkage table L,
Minimum support λ ($0 < \lambda \leq 1$), Frequent m-sequence F.
Suffix tree roots set R (R is root and F is empty first time)
Extendible set L (is frequent 1-sequence set the first time)
Output: Frequent (m+1)-sequence, F’.
Other Variables: S stores whether node is ancestor of the following nodes in the queue,
C stores the total number of events ei in the suffix trees.
□ □If R (suffix tree roots set) is empty, or the summation of R’s children is less than λ, return
□ □For each event, ei in L(header linkage table) , find the suffix tree of ei in T (i.e,ee | suffixtree), do
□ □Save first event in ei-queue to S.
□ □Following the ei-queue
□ □If event ei is the descendant of any event in R, and is not descendant of S, and Insert it
into suffix-tree-header set R’
- Add count of ei and ej to C.
- Replace the S with ei.
□ □If C is greater than λ (threshold)
- Append ei after F to F’ and output F’
- Call Algorithm MPLWAP-Mine (recursion)
□ □Else Remove ei from extendible set L

III. 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 theuser 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 patternsgenerated 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.

REFERENCE

Books
[1] Mining the web Discovering knowledge from hypertext data by Soumen Chakrabarti

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

Reference papers
[4] Web usage mining using improved frequent pattern tree algorithm Ashika Gupta , Rakhi arora, Ranjana sikarwar , 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. ManishaValera, Uttam Chauhan IEEE-2013
[8] PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-ProjectedPattern Growth .Jian Pei,Jiawei Han, Behzad Mortazavi-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. Surbhi Anand,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.Michal Munk, Jozef Kapusta, Peter Svec ELSEVIER - 2012
[13] A Hierarchical cluster based preprocessing methodology for web usage mining.Tasawar Hussain, Dr. Sohail Asghar, 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. Dheeraj kumar singh, Varsha Sharma ,Sanjeev Sharma IJCA-2012
[16] A Taxonomy of Sequential Pattern Mining AlgorithmsNizar R.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 Wenjia wang and Phuong thanh cao-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