RESEARCH PAPER | Engineering | Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

A review on hybrid algorithm of Sequence Pattern Mining

KEYWORDS

Vishal Barot	Harshita Kanani
ME scholar , Deprtment of Computer Engg, KSV University,Gandhinagar,Gujarat	Asst. Professor , Deprtment of Computer Engg, KSV University,Gandhinagar,Gujarat

ABSTRACT As the use of websites are increasing and from the website owner's view it is much better to understand the user's behavior so they get the idea to improve the service as well as quality .When user surf the websites their activities store in weblog files on the server. To understand user's behavior first we can use the web log files. To extract the users' behavior mining of web log is needed. Sequence Pattern Mining (SPM) is one of the technique to mine web logs to find the sequence of user for web pages.There are various algorithm for SPM. The This paper presents survey of SPM algorithms which can be used in web log mining.

1. Introduction

Web usage mining (WUM), WUM 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 [1].Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site. With using weblogs we can find out frequent patterns to improve the website and it can be useful for the web recommendation. Web log files contain the information about user like Date, Time , Site Name , IP Address , URI ,User-Agent , Status , Access time etc..

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 where the values are delivered in a sequence [2].The concept of sequence Data Mining was first introduced by Rakesh Agrawal and Ramakrishna Srikant in the year 1995 [3]. 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. 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. SPM algorithms are broadly categorized into three approaches: Apriori based Pattern Growth and Early pruning [9]. These techniques are for finding sequence pattern mining using different algorithms. Basically it provides the sequence patterns not the frequent pattern, web recommendation and also from application point of view it is helpful. Apriori algorithm generates very large amount of candidate sequence and perform too many database scan. Pattern growth based algorithm can solve above problem like Prefix span. It works on projected database but it fails if we consider the time to generate to projected database. Early pruning is a way to reduce search space and processing time for mining. In comparison of these algorithms, the hybrid algorithm of pattern growth and early pruning is more efficient for Sequence pattern mining.

2. Categorization of SPM algorithms.

SPM algorithms are mainly categorize in three terms namely apriori based , pattern growth and early pruning [9].

Key features of 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 :

Disadvantage

- It is a very undesirable characteristic of most apriori-based algorithms.
- Requires a lot of processing time and I/O cost.

Key features of pattern growth based methods

Sampling /Compression : Compression is used in the data structure that holds the candidate sequences, usually a tree.

Sampling : 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 [6] - Uses direct antimonotic app of apriori property to prune candidate sequence alongwith projected database.

PLWAP [10] - It also 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.

Search Space Partitioning : It allows partitioning of the generated search space of large candidate sequences for efficient memory management.

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

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

frequent sequences.

Depth first Traversal :

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 : 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.

Key features of early pruning based methods

Support couting avoidance : 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 amount of computation incurred by bitwise (usually AND) operations used to count the support for each candidate sequence.

Position Coded : It enables an algorithm to look-ahead to avoid generating infrequent candidate sequences.

3. Comparitive Study of Sequence Pattern Mining algorithms.

Comparative analysis of sequential pattern mining algorithm is done on the basis of their various important features.Apriori based algorithms has disadvantage of repeated scanning of database and huge sequence of candidate generation, which decreases the efficiency. The performance of apriori based algorithm goes down because of these reasons. Pattern growth-method is the solution method of limitations of the Apriori-based methods. It comes up with solution of the problem of generate-and-test. The features of this method that mentioned as above. Early pruning is a way to reduce search space and processing time for mining. It also has a key features like support counting avoidance and position coded.

The analysis of apriori based algorithm and pattern growth based algorithm shows that prefix span , a pattern growth algorithm is efficient than others. Time taken for lower support is almost half to double for SPAM and PrefixSpan as compare to Apriori. Gradually time taken by SPAM and PrefixSpan are decreased as compare to Apriori. PerfixSpan really perform better in case of execution time of algorithm[7].

GSP[3]	SPIRIT[13]	SPADE[11]	FREESPAN[12]	PREFIXSPAN[6]
Apriori Based	Apriori Based	Apriori Based	Pattern Growth Based	Pattern Growth Based
Uses Candidate generation and test approach	Uses Regular Expressions (REs) as a flexible Constraint	Uses vertical format sequential pattern mining method	Uses divide-and- conquer approach	Uses Projected database concept
Requires Multiple database scan.	Requires Multiple scans	Requires only three database scans	Reduces the cost of scanning multiple projected databases	Requires single database scan
Generates long sequential pattern, large number of candidates	Generates long sequential pattern that satisfies userspecified RE constraints	Generates large number of patterns, many of them are trivial or useless	Projects a large sequence database recursively into a set of small projected sequence databases	Generates long sequential pattern and less number of projected databases
Generates some candidates which doesn't have any existence in sequence database	Generates fewer candidates which have the potential to be frequent for higher values of minimum support.	Using equivalence classes on frequent sequences, the original problem Decomposes into smaller sub-problems	Recursively project a sequence database into a set of smaller databases based on the current set of frequent patterns	Never generates any prefix which is not present in sequence database
Not good for those applications where low support thresholds are used	Performs well, even if it contains a large number of cycles of moderate length	Good for fast mining of sequential patterns in large databases	Good for large set of sequential patterns	Good for those applications where low support thresholds are used
Performance is poor than Prefixspan algorithm	Performance is poor than Prefixspan algorithm	Performance is better than GSP and poor than prefixspan	Performance is better than GSP and poor than prefixspan	Performance is better than GSP algorithm

Table 1: A Comparative Study of Apriori and pattern growth based algorithms.[5]

It is true that in comparison of apriori based algo and patten growth based algo prefix span is more efficient algorithm compare to others.But another algorithm that is PLWAP[10],a hybrid algorithm of Pattern growth and early pruning is more efficient for Sequence pattern mining.

The table shows the comparison of prefixspan and PLWAP algorithm which is hybrid algorithm of pattern growth and early pruning.Comparative performance analysis of algorithms from each of the categories. Two data sets 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% . PrefixSpan needs memory space to hold the sequence database plus a set of header tables and pseudoprojection tables[8], PLWAP enjoys the fastest execution times, as it clearly separates itself from WAP-mine and Prefix Span , especially at low minimum support values when more frequent patterns are found and with large data sets[9].

Volume - 5 | Issue - 1 | Jan Special Issue - 2015 | ISSN - 2249-555X

Prefixspan[6]	PLWAP[10]
Pattern Growth based	Hybrid algorithm
Requires single database scan	Requires two database scan
It is pure Pattern growth based algorithm	It is hybrid algorithm of early pruning and pattern growth
It does not using DFS approach	It uses DFS approach
It uses more memory gradually	It uses less memory in comparison of Prefixspan
Uses projected database	Uses tree projection
It has features like Candidate sequence pruning, Search space partitioning and memory only.	It has features like sapmling, tree projection,DFS and position coded.

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

Comparative analysis of algorithm performance [9].

4. Conclusion

With the help of analysis of different algorithms of SPM and theoretical study, we can say that PLWAP ,a hybrid algorithm with outperforms pattern growth algorithms like prefixspan[6] and aprio based algorithm like GSP[3].It is clear that PLWAP Algorithm is more efficient with respect to running time, space utilization and scalability then other algorithms.

REFERENCE

Web References|[1]http://en.wikipedia.org/wiki/Web_mining | [2]http://en.wikipedia.org/wiki/Sequential_Pattern_Mining|Reference papers |[3]Srikant R. and Agrawal R., Mining sequential patterns: Generalizations and performance improvements , Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17.. | [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] PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth .Jian Pei,Jiawei Han, Behzad Mortazavi-Asl,Helen Pinto.IEEE-2013. | [7] Sequential Pattern Mining Methods: A Snap Shot.Niti Desai, Amit Ganatra IOSR-JCE-2013 | [8]Sequence Pattern Mining:Surveyand current research challenges.Chetna Chand ,Amit Thakkar ,Amit Ganatra.IJSCE-2012 | [9]A Taxonomy of Sequential Pattern Mining Algorithms.Nizar R.Mabroukeh , C.I. Ezeife.ACM-2010. | [10] Position Coded Pre-order Linked WAP-Tree for Web Log Sequential Pattern Mining Yi Lu and C.I. Ezeife2003. | [11] M. Zaki, „SPADE: An Efficient Algorithm for Mining Frequent Sequences , Machine Learning, vol. 40, pp. 31-60, 2001. | [12] Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C., Freespan: Frequent pattern-projected sequential pattern mining , 2000, pp. 355-359. | [13] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints" , VLDB'99, 1999