RESEARCH AND APPLICATION IN WEB USAGE MINING OF THE INCREMENTAL MINING TECHNIQUE FOR ASSOCIATION RULE

Sulan Zhang and Zhongzhi Shi
Key Lab of Intelligent Information Processing, Institute of Computing Technology
Chinese Academy of Sciences, Beijing, 100080, China
Email: zhangsl@ics.ict.ac.cn

Abstract: The paper analyzes some existing incremental mining algorithms for association rule and presents an incremental mining algorithm for association rule fit for Web Usage Mining. Because there are some characteristics of web logs which are dynamic, attributed, smaller and updated frequently, the algorithm uses BORDERS algorithm when mining single log file, and takes advantage of partition algorithm when mining many log files simultaneously.

Key words: Web Usage Mining, association rule, access patterns, incremental mining

1. INTRODUCTION

Web Usage Mining is to mine the user access patterns on user's access records reserved on server. Main techniques include the path analysis technique peculiar to web mining and some traditional techniques in data mining such as association rule mining, Sequential patterns mining, Clustering, Classification and etc.

The rules mined out can only reflect the current status of the database. The alteration of data source can probably cause new rules and make some old rules invalid. To improve the rules' stability and reliability, the rules maintaining defined as incremental mining is necessary. A direct and simple solution is to run mining algorithm on the whole database after being altered,

but it has obvious defectiveness which can't take advantage of the anterior results and the efficiency is very low when data source is very big or the mining results are very dense. The research of incremental mining technique for association rule in web usage mining is very important. For one thing, association rule can be the reference for bettering web site, the evidence for market development on web and heuristic rule for prefetching the web page for remote customers. For another, server access logs, the object of data mining in web usage mining, are updating perpetually and the size of the union of several logs in some days on an usual web site is very big, so it is inefficient to mine on those logs all over again. In the paper, we will discuss and analyze the incremental mining technique for association rule, then present an incremental algorithm fit for Web Usage Ming according to the characteristics of web logs.

2. RESEARCH AND ANALYSIS OF RELATED THEORIES AND METHODS

The process of association rule mining can be divided into two sub-processes: finding frequent itemsets and producing association rule. Association rule Mining can be simplified as frequent itemsets mining because the second sub-process can be finished directly. Similarly, the essence of association rule incremental mining is to find new frequent itemsets of database after being altered. Cheung believes that maintain of frequent itemsets includes looking for the following two kinds of itemsets:

The defeater: the itemsets that are frequent before data source being altered but infrequent after data source being altered.

The winner: the itemsets that are infrequent before data source being altered but frequent after data source being altered.

The association rule maintaining was introduced in the reference [1] at first. The paper presented FUP algorithm that updates association rule when new transactions are added into database. A more general algorithm FUP2 was presented [2], the algorithm updates association rule when insertion, deletion or modification is applied on data source. The two algorithms are based on the principle of Apriori algorithm and need to scan database O(n) times (n is the number of items in the biggest frequent itemset). They take advantage of anterior mining results to decrease works and improve the efficiency of incremental mining when mining new rules, but the general and long-term price of them is not small. The DELI algorithm [3] uses sampling and statistics technique to estimate the distinction of association rule sets between before alteration and after alteration. The necessity of updating is judged from the distinction, then too frequent updating can be avoided and a

large amount of resource and long-term price can be saved. Thomas and Feldman presented the BORDERS algorithm at the same time [4,5]. The algorithm uses the negative border notion presented by Toivonen [6] to judge the necessity of checking every candidate itemset on database. The biggest contribution of the algorithm is that it needs only a time at most to scan database for updating frequent itemsets when the alteration of database gives rise to expansion of the negative border of frequent itemsets. The algorithm is fit for frequently-updated database, but it is not fit for very large database because the negative border reserved during computing process will occupy a large memory. It can be processed collaterally.

3. UPDATE_BP ALGORITHM PRESENTED

When a manager of web site wants to analyze the user access patterns in some days, the log files in these days should be mined. It is a most usual means to union all log file in these days and mine them simultaneously, but its efficiency is not high because data is too much and those results mined separately from these logs are not used. Now an incremental mining algorithm on web logs Update_BP is presented to solve the problem.

Because there are some characteristics of web log which are dynamic, attributed, smaller and updated frequently, the algorithm Update_BP uses BORDERS algorithm which has been introduced above when mining a single log file, and fulfill in virtue of partition algorithm[7] when mining many log files.

The data in web server access logs is different with market basket. It is changed into appropriate data format fit for association rule mining before mining. The following is the description of algorithm Update_BP:

The algorithm has the same description of terminology (p_i, C_k^p, L_k^p, L^p, C_k^G, C^G, L_k^G, L^G) as the reference [7] does.

(1) n= Number of Log files
(2) For i=1 to n begin
(3) L^i=gen_large_itemsets(p_i)
(4) End
(5) For(j=2; $L_1^j \neq \Phi$, j=1,2,…,n; i++) do
(6) $C_i^G = \cup_{j=1,2,\ldots,n} L_1^j$ // union
(7) For i=1 to n begin
(8) For all candidates $c \in C^G$ gen_count(c,p_i)
(9) End
(10)$L^G=\{c \in C^G | c.count \geq minsup\}$

Figure 1 Update_BP

The function gen_large_itemsets is achieved with the BORDERS algorithm whose input is single log file and output is local frequent itemsets with a variety of length of the log file. All local frequent itemsets with same length of every log file are merged into the whole candidate itemset with the same length in the step (6). In the step (7) to step (9), the whole support of every candidate is computed on all log files. The whole frequent itemset is built in step (10). The function gen_count was in the reference [7]. If some log files have already been mined separately, the step (3) that builds local frequent itemsets can be omitted and the time for mining can be reduced.

All-round analysis of the performance of BORDERS algorithm and Partition algorithm has been made in the original papers and will not be iterated in this paper. The simulation of the algorithms in lab verified that the two algorithms can be used in web usage mining and can improve largely the efficiency of log incremental mining.

4. CONCLUSIONS

The paper actualizes the incremental mining of web log in the light of BORDERS algorithm and Partition algorithm. The experiment has proved that the algorithm can improve largely the efficiency of log mining. In the following research, more incremental mining algorithms for other patterns such as Sequential patterns, Clustering, Classification and etc. in web mining should be developed.

REFERENCES

1. D.Cheung, J.Han, V.Ng and C.Y.Wong. Maintenance of discovered association rules in large databases: An incremental updating technique. In ICDE'96, New Orleans, Louisiana, USA, Feb.1996
2. D.Cheung, S.Lee and B.Kao. A general incremental technique for maintaining discovered association rules. In Proc. Of the 5th International Conference on Database Systems for Advanced Applications, Melbourne, Australia, April 1-4,1997
3. S.Lee and D.Cheung. maintenance of discovered association rules: When to update? In DMKD'97, Tucson, Arizona, May.1997
4. S.Thomas, S.Bodagala, K.Alsabti and S.Ranka. An Efficient Algorithm for the Incremental Updation of Association Rules in Large Databases. In KDD'97, New Port Beach, California, Aug.1997
5. R.Feldman, Y.Aumann, A.Amir and H.Mannila. Efficient Algorithms for Discovering Frequent Sets in Incremental Databases. In DMKD'97, Tucson, Arizona, May.1997
6. H.Toivonen. Sampling Large Databases for Association Rules. In VLDB'96, pp.134-145
7. A.Savasere, E.Omiecinski, and S.Navathe. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very large Database,1995, pp.432-444