Survey: Efficient Approach for large Database Compressed in Association Mining

Sunichchha Chauhan[1], Vimal Tiwari[2]

[1]*M.Tech Scholar , Department of computer science & engineering, Bhopal institute of technology & science, Bhopal, India*

[2]*HOD & Professor, Department of computer science & engineering, Bhopal institute of technology & science, Bhopal, India*

ABSTRACT

In an era of knowledge explosion, the growth of data increases rapidly day by day. Since data storage is a limited resource, how to reduce the data space in the process becomes a challenge issue. Data compression provides a good solution which can lower the required space. Data mining has many useful applications in recent years because it can help users discover interesting knowledge in large databases. However, existing compression algorithms are not appropriate for data mining. In this research a new approach called Mining Merged Transactions with the Quantification Table was proposed to solve these problems. Mining Merged Transactions with the Quantification Table uses the relationship of transactions to merge related transactions and builds a quantification table to prune the candidate item sets which are impossible to become frequent in order to improve the performance of mining association rules. The experiments show that Mining Merged Transactions with the Quantification Table perform better then existing approaches.

Keywords: *Association rule, Apriori Algorithm, merged transaction, quantification table.*

1. INTRODUCTION

Data compression is one of good solutions to reduce data size that can save the time of discovering useful knowledge by using appropriate methods, for example, data mining. Data mining is used to help users discover interesting and useful knowledge more easily. It is more and more popular to apply the association rule mining in recent years because of its wide applications in many fields such as stock analysis, web log mining, medical diagnosis, customer market analysis, and bioinformatics. In this research, the main focus is on association rule mining and data pre-process with data compression. Proposed a knowledge discovery process from compressed databases in which can be decomposed into the following two steps:

1.1 Data pre-process step:

Data pre-process transforms the original database into a new data representation where several transactions are merged to become a new transaction. Eventually, it generates a new transaction database at the end of the data pre-process step [4].

1.2 Data mining step:

It uses an Apriori-like algorithm of association rule mining to find useful information. There are some problems in this approach. First, the compressed database is not reversible after the original database is transformed by the data pre-process step. It is very difficult to maintain this database in the future. Second, although some rules can be mined from the new transactions, it still needs to scan the database again to verify the result. This is because the data mining step produces potentially ambiguous results. It is a serious problem to scan the database multiple times because of the high cost of re-checking the frequent item sets[6].

It is even a bigger challenge to maintain the compressed database in the future. In addition, it spends too much time to check candidate itemsets in the data mining step. In this research, a more efficient approach, called Mining Merged Transactions with the Quantification Table is proposed, which can compress the original database into a smaller one and perform the data mining process without the above problems. Our approaches have the following characteristics:

(a) The compressed database can be decompressed to the original form.
(b) Reduce the process time of association rule mining by using a quantification table.
(c) Reduce I/O time by using only the compressed database to do data mining.
(d) Allow incremental data mining.

2. PROBLEM DOMAIN

In an era of knowledge explosion, the growth of data increases rapidly day by day. Since data storage is a limited resource, how to reduce the data space in the process becomes a challenge issue. Data compression provides a good solution which can lower the required space. Data mining has many useful applications in recent years because it can help users discover interesting knowledge in large databases. However, existing compression algorithms are not appropriate for data mining. In two different approaches were proposed to compress databases and then perform the data mining process. However, they all lack the ability to decompress the data to their original state and improve the data mining performance [10].

3. SOLUTION DOMAIN

The description of the proposed algorithm focuses on compressing related transactions and building a quantification table for pruning candidate itemsets that are impossible to become frequent itemsets. Finally, an example is provided to show the processes of our method. To simplify the description, it assumes the items in each transaction are presented in a lexicographical order. Algorithms like compress transactions to reduce the size of a transaction database. Then, they use Apriori-like algorithms to mine the compressed database [7,8].

Since both need to scan the database more than once, they have a much higher process cost. The first problem is due to the lack of rule or constraint in the process of merging transactions in the data compression phase. Therefore, the compressed database can not be decompressed to its original form In addition, they don’t use user-defined threshold to filter infrequent 1-itemsets from the original database.

Another problem is that Apriori-like algorithms generate a lot of candidate itemsets and need to check the candidate itemsets by scanning the database. It is very time-consuming. Our goal is to take the advantages of and Apriori algorithm without suffering from the problem of checking candidate itemsets and recovering the database for new data. In order to provide a better performance, we limit the number of database scan to be one in the data compression phase and build a quantification table. In the data mining phase, we use the same approach of Apriori algorithm to generate candidate itemsets and reduce the number of candidate itemsets by using the quantification table. We also reduce the time of scanning the database.

4. REFERENCES

[1]. Jia -Yu Dai, Don – Lin Yang, Jungpin Wu and Ming-Chuan Hung, “ Data Mining Approach on Compressed Transactions Database” in PWASET Volume 30, pp 522-529, 2010.
[2]. M. C. Hung, S. Q. Weng, J. Wu, and D. L. Yang, "Efficient Mining of Association Rules Using Merged Transactions," in WSEAS Transactions on Computers, Issue 5, Vol. 5, pp. 916-923, 2008.
[3]. D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "MAFIA: A maximal frequent itemset algorithm," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, pp. 1490-1504, 2008.
[4]. D. Xin, J. Han, X. Yan, and H. Cheng, "Mining Compressed Frequent-Pattern Sets," in Proceedings of the 31st international conference on Very Large Data Bases, pp. 709-720, 2007.
[5]. G. Grahne and J. Zhu, "Fast algorithms for frequent itemset mining using FP-trees," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, pp. 1347-1362, 2005.
[6]. M. Z. Ashrafi, D. Taniar, and K. Smith, "A Compress-Based Association Mining Algorithm for Large Dataset," in Proceedings of International Conference on Computational Science, pp. 978-987, 2003.
[7]. Ashrafi and K. Smith, “Data Compression-Based Mining Algorithm for Large Dataset," in Proceedings of International Conference on Computational Science, 2003.

[8]. D. I. Lin and Z. M. Kedem, "Pincer-search: an efficient algorithm for discovering the maximum frequent set," IEEE Transactions on Knowledge and Data Engineering, Vol. 14, pp. 553-566, 2002.

[9]. E. Hullermeier, "Possibilistic Induction in Decision-Tree Learning," in Proceedings of the 13th European Conference on Machine Learning, pp. 173-184, 2002.

[10]. Cheung, W., "Frequent Pattern Mining without Candidate generation or Support Constraint." Master's Th esis, University of Alberta, 2002.

[11]. Huang, H., Wu, X., and Relue, R. Association Analysis with One Scan of Databases. Proceedings of the 2002 IEEE International Conference on Data Mining. 2002.

[12]. Beil, F., Ester, M., Xu, X., Frequent Term-Based Text Clustering, ACM SIGKDD, 2002

[13]. Zaki, M. J., Parthsarathy, S., Ogihara, M., and Li, W. New Algorithms for Fast Discovery of Association Rules. KDD, 283-286. 1997. Agarwal, R., Aggarwal, C., and Prasad, V.V.V. 2001.

[14]. Goulbourne, G., Coenen, F., and Leng, P. H. Computing association rule using partial totals. In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases, 54-66. 2001.

[15]. Pei, J., Han, J., Nishio, S., Tang, S., and Yang, D. H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases. Proc.2001 Int.Conf.on Data Mining. 2001.

[16]. Orlando, S., Palmerini, P., and Perego, R. Enhancing the Apriori Algorithm for Frequent Set Counting. Proceedings of 3rd International Conference on Data Warehousing and Knowledge Discovery. 2001.

[17]. Grahne, G., Lakshmanan, L., and Wang, X. 2000. Efficient mining of constrained correlated sets. In Proc. 2000. Int. Conf. Data Engineering (ICDE'00), San Diego, CA, pp. 512–521.