Enhancing the Performance of Association Rule Mining Using Inverted Index Compression

Esa Mohammed A, Hariharan N, Mohan R, Arul jothi K
Student, Information Technology, New Prince Shri Bhavani college of Engg and Tech.
Student, Information Technology, New Prince Shri Bhavani college of Engg and Tech.
Student, Information Technology, New Prince Shri Bhavani college of Engg and Tech.
Assistant Professor, Information Technology, New Prince Shri Bhavani college of Engg and Tech.

ABSTRACT

Extracting and Mining information from massive data collection is to be a great challenge nowadays. The interest in huge volume of data collection has increased exponentially due to their needs. So the size of the data as well as the computational complexity also increased. In order to reduce the space and time complexity, we are going to implement shuffling and sorting strategies in inverted index compression and also association rule mining. Thus the above strategies propose a new framework which improves the efficiency and solve the problem of space and time complexity.

Keyword:- *Association Rule Mining , Index compression*.

1. INTRODUCTION

Data mining is mainly used for analyzing data from huge volume of data collection. It is an analytical tool for analyzing data. It allows users to identify the relationships between data in different perspective. It finds out the frequent pattern form large scale databases. Data mining is mainly used for many organization who focusing the customer strongly in nature. It is used to determine the impact on sales with customer satisfaction, and corporate profits. Data mining concepts used to find out the sales performance in an accurate way. Based on that many organizations can improve their sales in marketing.

2. EXISTING WORK

The algorithm was proposed by Agrawal and srikant in 1994. It is one of the algorithms which is used to find the frequent itemset. Aprori is implemented by using bottom up approach where frequent subset is extended one at a time. BFS and hash tree structure are used to count the candidate itemset. For the length of K itemset the length of k-1 candidate itemset will be generated. Then it will delete the candidate which have infrequent pattern the drawback of apriori is that it consumes more space in the database and to generate the candidate item set it requires more computation to scan the database

3. PROPOSED WORK

As we mentioned earlier in this paper, we are not going to implement new algorithm. We are going to implement the existing data structure in the association rule mining which is uses the combination of following.

1) Shuffling Strategy
2) Mapping Strategy

3.1 Shuffling Strategy

The concept of shuffling strategy is used for grouping the similar data items. The data which contains the similar value stored under a common name. Here we are using Inverted Index mapping to find the index value for the similar data items.

3.1.1 Inverted Index Mapping

Inverted Index Mapping is also called as centralized data structure in information retrieval. The Inverted Index Mapping is an efficient technique to retrieve the data from database. There are two major benefits of Inverted Index Mapping they are given below

1. **Increasing the efficiency of usage of cache memory:**

In our project we consider the super market inventory system each similar product should have a unique index value. The frequent items are stored in database by using an index value. If we need to retrieve some data from data base, we need not to scan the entire data base and we can get the required information with the help of that index value. Because of this scheme we can consume lesser time rather than normal search and retrieval of data or information. As a result, we can reduce the response time of information retrieval.

2. **Transfer data from disk to memory:**

With the help of inverted index compression, the size of the data also reduced. The time taken for transfer the uncompressed data from disk to memory is higher than the transferring compressed data from disk to memory.

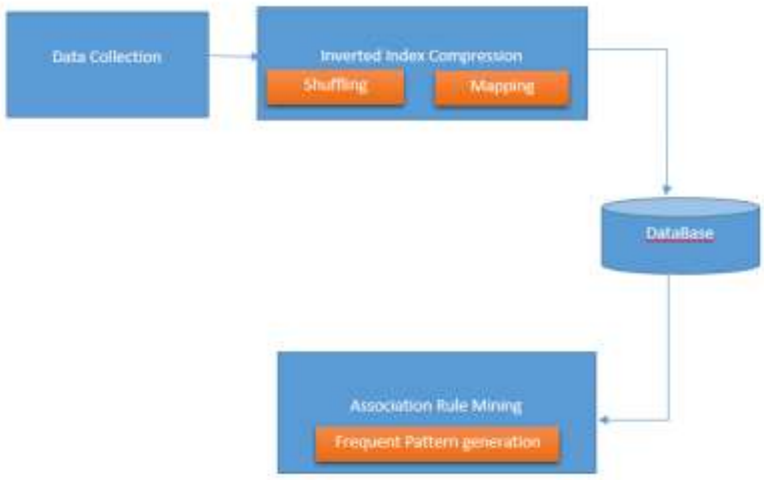
3.2 Mapping Strategy

The concept of sorting is to store the unique value which contains the similar data and it can be done using RLE compression.

3.2.1 RLE compression

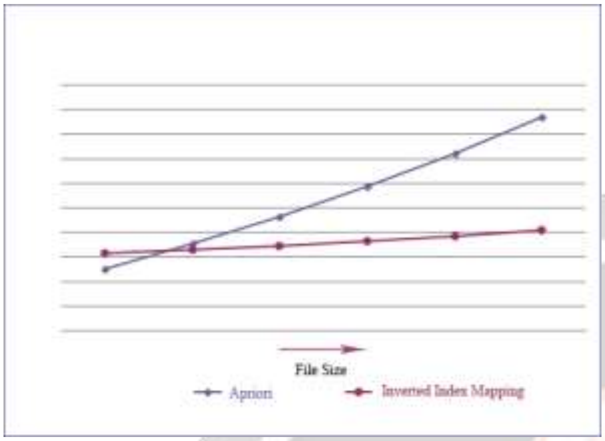
The compression procedure that we are using is RLE compression. The abbreviation of RLE compression is Run Length Encoding. It is a lossless compression algorithm which compresses the value of similar data. The purpose of RLE compression in our project is to reduce the size of the data cells in the data base. The similar values are stored under a common name so that the usage of data cells is reduced in our project. The simplification of the data is done using the RLE compression. The advantage of RLE compression is easy to implement and it reduces the repetitive data in the database.

4. ARCHITECTURE DIAGRAM

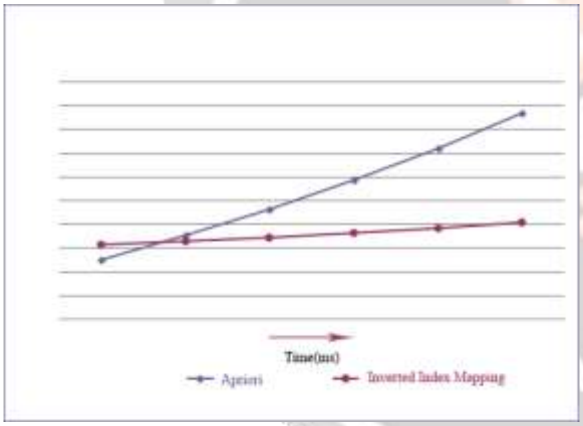


5. EXPERIMENTAL RESULTS

When compared to previous algorithm, the inverted index compression and association rule mining improves the efficiency through solve the problem of space and time complexity. The space complexity is reduced by mapping the similar value and the time complexity is reduce by assigning unique index value for each data.by assigning unique index value the value can be retrieved easily.



The above graph shows the difference between the space allocation between two algorithms. Inverted index compression occupies less space compared to the apriori algorithm.



The above given graph used to display the difference between the computational time of the both algorithm. The graph shows that Inverted Index Mapping requires less computational time compared to the Aprori algorithm**.**

6. CONCLUSIONS

In this paper we proposed a data structure that simplifies and reorganize the data which utilizes the data storage effectively. In this algorithm instead of storing all the data into a single table we are maintaining separate table for each product and assign the unique index value. The space complexity is reduced by avoiding the unwanted entries in the database. The mapping of data is based on the unique index value. By assigning the index value the data can be retrieved easily so the computational complexity also reduced here.

7. REFERENCES

[1] V. Marx, “Biology: The big challenges of big data,” Nature, vol. 498, no. 7453, pp. 255–260, 2013.

[2] S. Bechikh, A. Chaabani, and L. Ben Said, “An efficient chemical reaction optimization algorithm for multiobjective optimization,” IEEE Trans. Cybern., vol. 45, no. 10, pp. 2051–2064, Oct. 2015.

[3] H. Gao, S. Shiji, J. N. D. Gupta, and W. Cheng, “Semi-supervised and unsupervised extreme learning machines,” IEEE Trans. Cybern., vol. 44, no. 12, pp. 2405–2417, Dec. 2014.

[4] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, “An efficient accelerator for attribute reduction from incomplete data in rough set framework,” Pattern Recognit., vol. 44, no. 8, pp. 1658–1670, 2011.

[5] E. Lo, N. Cheng, W. W. K. Lin, W.-K. Hon, and B. Choi, “MyBenchmark: Generating databases for query workloads,” Int. J. Very Large Data Bases, vol. 23, no. 6, pp. 895–913, 2014.

[6] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, “Large-scale unsupervised hashing with shared structure learning,” IEEE Trans. Cybern., vol. 45, no. 9, pp. 1811–1822, Sep. 2015.

[7] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, “Toolkit-based high-performance data mining of large data on MapReduce clusters,” in Proc. IEEE Int. Conf. Data Min., Miami, FL, USA, 2009, pp. 296–301.

[8] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in Proc. 20th Int. Conf. Very Large Data Bases (VLDB), Santiago, Chile, 1994, pp. 487–499.

[9] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, “Positive approximation: An accelerator for attribute reduction in rough set theory,” Artif. Intell., vol. 174, nos. 9–10, pp. 597–618, 2010.

[10] B. Goethals and M. J. Zaki, “Advances in frequent itemset mining implementations: Report on FIMI’03,” ACM SIGKDD Explor. Newslett., vol. 6, no. 1, pp. 109–117, 2004.