Applied Research of an improved Apriori algorithm in the logistics industry

Jufang Li [1, a] , Yuan tang [2,b] , Xiao xiao[3,c] and Yu gao[4,c]

[1]City College of Science and Technology. Chongqing University, Chongqing, China

[2]City College of Science and Technology. Chongqing University, Chongqing, China

[3]City College of Science and Technology. Chongqing University, Chongqing, China

[a]821851471@qq.com,[b]352627730@qq.com,[c] 147745193@qq.com

Keywords: Apriori algorithm, database scanning, ABVO algorithm, Logistics management system

Abstract. Modern logistics is a huge and complex system and the flow of information is large. We must deal with the huge amounts of data accurately in time, and data mining technology is appropriate to solve it. We improved the typical association rule-Apriori algorithm which is often used in data mining technology and got the new algorithm-ABVO algorithm. It is applied in the logistics management system in a company and proved working effective in reducing time of algorithm running and database scanning.

Background information

Like other industries, the information logistics network system of logistics industry has a huge data flow, and one of difficult problems is how to accurately and timely collect and analyze the information[1]. With the accumulation of the time, the logistics company accumulated a lot of historical data information, traditional database management system cannot do a good job in analysis to historical data of database, but data mining and data warehouse technology can well solve this problem. Association rule mining is an important research branch of data mining and applied in many fields. So this paper will focus on the introduction of application of improved association rule Apriori algorithm in logistics industry.

Apriori algorithm description and shortcomings

Apriori algorithm uses prior knowledge of frequent itemsets and a search iterative method which called layer-by-layer, namely K itemsets used to explore itemsets (K+1) [2].The shortcoming of the traditional Apriori algorithm is the high frequency of scanning the transaction database and the connection into a high-dimensional candidate item-sets[3]. We can know from Apriori algorithm that, the generation of each candidate itemset in the algorithm scans the database for one time, and the generation of candidate itemset by each iteration and the support statistics are the waste of the time, in fact, some transactions in candidate itemset have not played any role to the generation of frequent itemset, so we find that the reduction of transactions without role in the database is very necessary to the improvement of algorithm efficiency. In order to improve the algorithm's efficiency, the classical Apriori algorithm will be optimized mainly according to the following two aspects: (1) Reduce database scanning's frequency; (2) Reduce the number of generated candidate itemsets [4].

This paper presents an improved algorithm of Apriori algorithm baseing on vector operation, which referred to the ABVO algorithm. It guarantees to scan the database only once and have a higher efficiency. In ABVO, the database is visited for once first, then create a vector of the

present-status which appears in the database records, and the count of the itemsets which appear later is computing by the operation among the vectors. This method can reduce the time of reading large amounts of data from the external memory into internal storage repeatedly. In Table1, T_1, T_2, T_3,…,T_m represents a transaction record, and I_1, I_2, I_3,…,I_n represents the various items which contained in all records in the database. Each item in the table represents the vector of the occurrence of a itemset in the record as shown in Equation 1:

$$a_{ij}=\begin{cases} 0 \text{ (when the item } I_j \text{ doesn't appear in } T_j) \\ 1 \text{ (when the item } I_j \text{ appears in } T_j) \end{cases} \quad (i=1,2,3,...,m;j=1,2,3,...,n)\ . \tag{1}$$

The column vector means that all items are contained in the record. $(a_{1j},a_{2j},a_{3j},\ldots,a_{mj})$ $(j=1,2,\ldots,n)$

Table1 Vector schematic sheet

F_1-I / CT	I_1	I_2	…	I_n
T_1	a_{11}	a_{12}	…	a_{1n}
T_2	a_{21}	a_{11}	…	a_{2n}
…	…	…	…	…
T_m	a_{m1}	a_{m2}	…	a_{mn}

The occurrence number of all the 1- candidate can be find by using the vectors in table1, the calculating as shown in Equation 2：

$$N_j^1=\sum_{i=1}^{m} a_{ij}(j=1,2,...,n)\ . \tag{2}$$

Compare the occurrence number of all the 1- candidate itemsets with Minsupport, and to get all the 1- candidate itemsets which are larger than Minsupport, namely, frequent 1-itemsets. The second step is to find frequent 2- itemsets. The first is to compress the record for once by using the transaction compression, according to the principle “The record which contains only k items has no possibility to contain k+1 items in collection.”, therefore, the frequent 1- itemsets can be removed from the vector table when finding frequent 2- itemsets. The calculating for the items in record as shown in Equation 3:

$$M_i=\sum_{j=1}^{n} a_{ij}(j=1,2,...,m)\ . \tag{3}$$

So the vector form of candidates of frequent 1- itemsets for generates the frequent 2- itemsets has found, as shown in table2:

table2 The vector table of the frequent 1-itemsets for generate candidates of frequent 2-itemsets

F2-I / CT	I1	I2	…	In
T_1	a_{11}	a_{12}	…	$a_{1n'}$
T_2	a_{21}	a_{22}	…	$a_{2n'}$
…	…	…	…	…
T_m	$a_{m'1}$	$a_{m'2}$	…	$a_{m'n'}$

According to the connection step of Apriori algorithm, by frequent 1-itemsets to generate 2-itemsets, and in accordance with Apriori property to generate all candidate 2-itemsets, the

obtained vectors are shown in Table 3:

Table3 The vector table of candidates of 2-itemsets

CT \ C_2-I	I_1I_2	I_1I_3	…	I_1I_n	I_2I_3	…	$I_2I_{n'}$	…
T_1	b_{11}	b_{12}	…	$b_{1(n'-1)}$	$b_{1n'}$	…	$b_{1(2n'-3)}$	…
T_2	b_{21}	b_{22}	…	$b_{2(n'-1)}$	$b_{2n'}$	…	$b_{2(2n'-3)}$	…
…	…	…	…	…	…	…	…	…
$T_{m'}$	$b_{m'1}$	$b_{m'2}$	…	$b_{m'(n'-1)}$	$b_{m'n'}$	…	$b_{m(2n'-3)}$	…

In which, I_i and I_j are obtained by column vectors Ii and I_j in Table2 through Equation 4 operation:

$$I_i I_j(b_{1k}, b_{2k}, ..., b_{ik}) = I_i(a_{1i}, a_{2i}, ..., a_{m'i}) \text{Ä} I_j(a_{1j}, a_{2j}, ..., a_{n'j})$$
$$(i = 1, 2, ..., m'; j = 1, 2, ..., n'; k = 1, 2, ..., n'(n'-1)/2). \quad (4)$$

At this time, each vector of occurrence in the table is as shown in Equation 5:

$$b_{ik} = \begin{cases} 1 & \text{if } a_{m'i}=1 \text{ and } a_{m'j}=1 \\ 0 & \text{if } a_{m'i}=0 \text{ or } a_{m'j}=0 \end{cases}. \quad (5)$$

Column vectors (b_{1j}, b_{2j}, b_{3j}, …,b_{mj}) (j=1,2,…,n) mean the situation of all 2-itemsets in the record, then by Equation 6 to calculate the frequency of 2-itemsets:

$$N_j^2 = \sum_{i=1}^{m'} b_{ij} (j = 1, 2, ..., n'). \quad (6)$$

Compared all candidate 2-itemsets' frequency with the minimum support degree to obtain all 2-itemsets of greater than Minsupport, namely, frequent 2-itemsets. Then, again use transaction compression method to delete the record of only including two items from the vector representation, to obtain frequent 2-itemsets vector representation, which is used to generate candidate frequent 3-itemsets. And so on, use the same method to generate frequent 3-itemsets, frequent 4-itemsets and …, until generate frequent n-itemsets.

Results analysis: It can be drawn from the description of ABVO algorithm that, the algorithm only browses the database once and then establishes the vectors of representation items in the database records, the count of later itemsets' frequency will be realized through the operation between the vectors, thus greatly reduced the frequency of database scanning, and saved the operation time. In addition, in operation process, ABVO algorithm uses transaction compression method, which will directly delete the transaction items without any role in generated candidate itemsets to the generation of frequent itemsets, thus improved the algorithm efficiency.

Analysis of experimental results

Randomly sampled 30000 data records of a logistics company and set the minimum support degree and minimum confidence. With the continuous increase of Minsupport and the increase of database records, compared ABVO algorithm with Apriori algorithm, the change curves of time are as shown in Figure1 and Figure2:

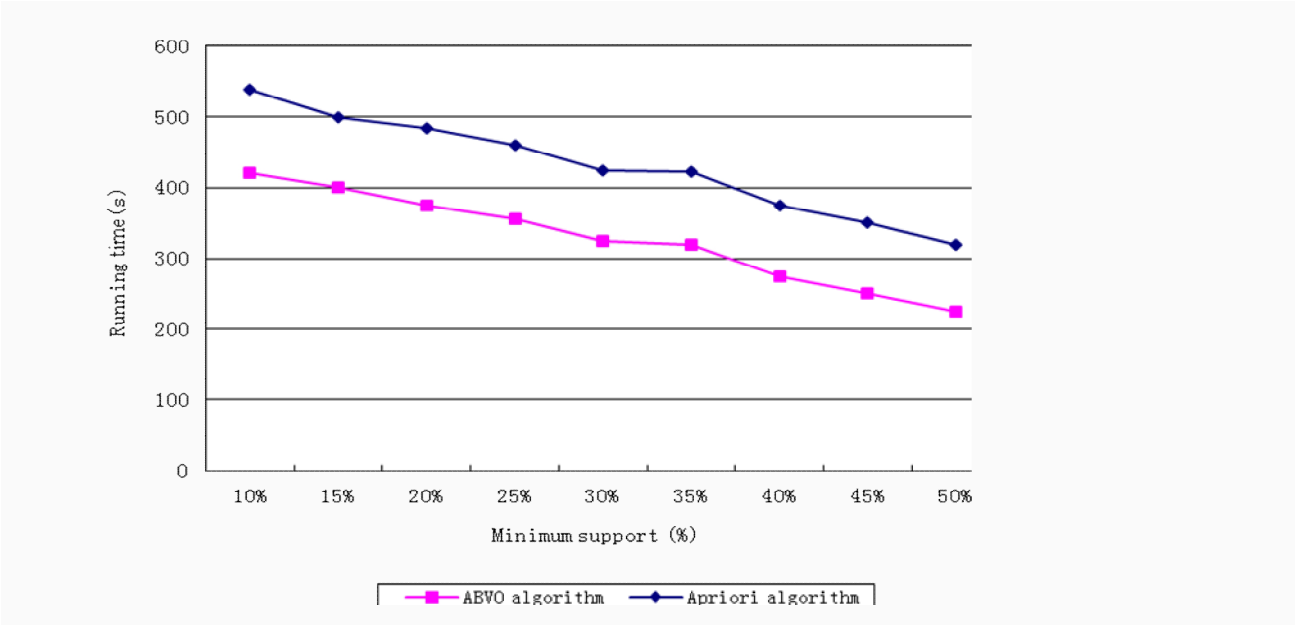

Figure1 Comparison Diagram of Algorithm Running Time

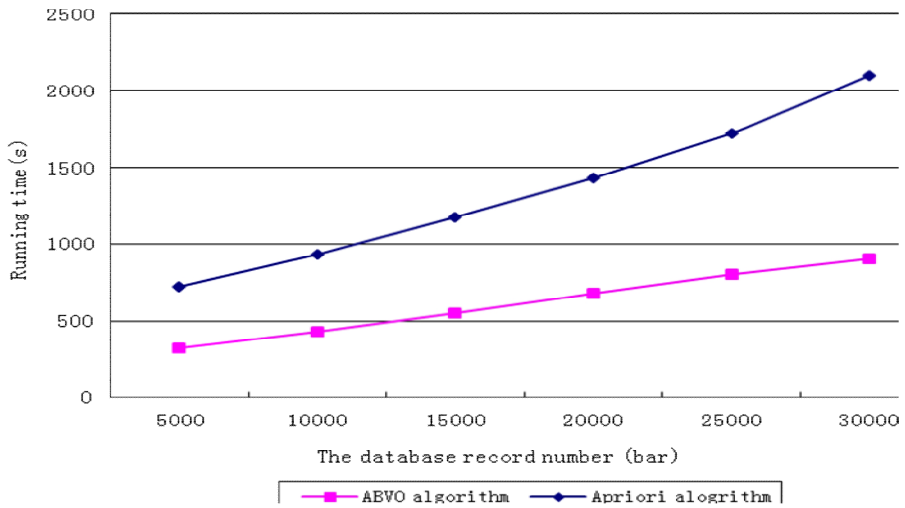

Figure2 Comparison Diagram of Algorithm Running Time

Based on the above diagrams, we can get various experimental indicators as follows: assume Minsupport is 10% and Minconfidence is 50%, the use time of improvement method proposed in the paper is 8 minutes, the frequency of scanning database is one time, and use the traditional Apriori mining algorithm, the results are the same with ABVO algorithm, but the time is 16 minutes and the frequency of scanning database is four times. By comparing and analyzing, we can get that, with the increase of database's record, the execution time of ABVO algorithm is faster than what of Apriori algorithm, with higher efficiency and achieves the intended target of the experiment.

Summary

In modern logistics management system, make full use of logistics system based on data mining technology to provide decision support for the decision maker of the logistics enterprise, will be conducive to enhance the core competitiveness of the logistics enterprise[5]. This paper studies the problems of Apriori algorithm in logistics information mining, and puts forward ABVO algorithm, which is the improvement algorithm of Apriori algorithm. The improved algorithm not only reduces

the frequency of scanning database, but also reduces the number of candidate itemsets, which cannot be called frequent itemsets, thus improves the running efficiency of the algorithm.

References

[1] Gao Xiaoting. The application of data mining in logistics management[J]. Technological Development of Enterprise (The second half of the month), 2010, 29(10):F252.

[2] Mao Guojun. Principle and algorithm of data mining,(The second edition). Tsinghua University Press 2011.

[3] Xie Zongyi. Research and improvement of Apriori algorithm for mining association rules, Journal of Hangzhou University of Electronic Science and Technology, 2006.26(3):TP311.13.

[4] Han Jiawei, Kan Bo. Data mining concepts and techniques[M].Mechanical Industry Press 2007.03.

[5] Deng Zhilong. The application of data mining in logistics management, Journal of Shanxi Youth Vocational College, 2009,(4):G714.