Research on Route Sequence Mining Algorithms for Logistics Data Analysis

He Liang

Yunnan Vocational College of Mechanical and Electrical Technology, Kunming, 650203, China

Keywords: Logistics Transportation; Data Analysis; Path Planning; Data Mining

Abstract: The application of information technology in logistics system makes the scale of logistics data increase rapidly and the flow of data increase, which makes it more difficult for enterprises to analyze and process data and make decisions based on it. Data mining technology can effectively improve the business ability of enterprises, enhance their service ability and market competitiveness. In order to improve the efficiency of path planning in logistics system, this paper proposes a new data mining path planning algorithm, which can effectively reduce the number of database scans, generate excessive candidate path sequences, and reduce the time cost of generating frequent path sequences. This method can also be applied to the generation of frequent path sequences in network access, customer purchase and other situations.

1. Introduction

At present, data mining technology is developing at an unprecedented speed, and has been widely used in government, power, enterprises, telecommunications, finance and other industries, but its application in the logistics industry is not very common. With the improvement of logistics informationization level, logistics strategy has changed from internal integration to external integration [1-3]. Supply chain management has become a very important part of competitive strategy. The application of information-based logistics network system makes the scale of data expanding continuously and produces huge data flow, which makes it difficult for enterprises to collect and make timely decisions on these data efficiently [4-5]. Data mining method effectively promotes business process reengineering, improves and strengthens customer service, realizes enterprise scale optimization, and effectively improves the competitiveness of enterprises. Therefore, it is very important for logistics enterprises or logistics users to analyze the flow of goods in logistics through data mining technology [6-8]. Therefore, a new path planning algorithm is proposed in this paper. The algorithm scans the logistics path information database, carries out initial screening, obtains the preliminary path sequence, then generates the targeted candidate excessive path sequence, optimizes the candidate path sequence, and then obtains the frequent path sequence.

2. Logistics Information Mining Algorithms

In the logistics decision support system, the first clear goal is to discover the flow direction of goods in the future logistics market. Logistics users can discover the purpose of different shippers choosing to transport the same batch of goods separately through the decision support system [9]. Logistics enterprises can discover the possible changes in the future logistics market through the logistics decision support system. The whole process of logistics information mining is shown in Figure 1.

  DOI: 10.25236/mmmce.2019.101

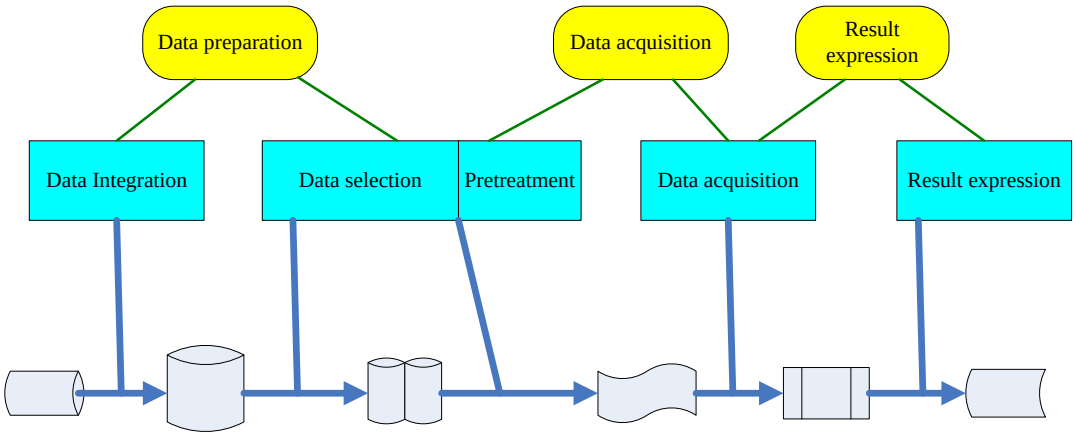

Figure 1. Flow chart of logistics information mining

2.1. Logistics data rule mining

Association rule technology in data mining is to discover meaningful links and rules between things. It can discover interesting and frequent patterns, associations and correlations between transaction databases, relational databases and item sets of large amounts of data stored in other databases [10-11]. Therefore, using association rules technology, we can analyze and mine the route data in logistics, find out the frequent path information, and find out the direction of goods in the logistics market and the possible changes in the future.

Association rules are implicative expressions like A→B, where $A \subseteq I, B \subseteq I$, where $I = \{I_1, I_2, ..., I_m\}$ is a set of items and $A \cap B = \phi$. Rule A→B is established in transaction set D and has support degree sup. Sup is the percentage of transactions in D that contain $A \cup B$ (that is, the union of sets A and B or both), and it is probability $P(A \cup B)$. Rule A→B has confidence C in transaction set D, where C is the transaction in D that contains A and also the percentage of B. This is conditional probability P (B | A).

1) Support

Sup (A→B) = P (B | A). If 80% of steel and iron plates are to be packed in 45-foot (13.72 m) containers, the support of 45-foot containers is defined as 80%.

2) Confidence

Confidence(A→B) = P (B | A).. For example, 60% of 40-foot (12.2m) containers can hold three kinds of cargo, A, B and C at the same time. It can be defined as a rule that a container with cargo A and B can simultaneously load C. It is said that a container with cargo A and B can simultaneously load C with a confidence of 60%.

Mining cargo routing rules can be divided into two steps: 1) Finding out all frequent routing sequences, which are solved by the proposed mining algorithm of routing sequences; 2) Generating associated routing rules from frequent routing sequences. These rules need to satisfy minimum support and minimum confidence.

2.2. Path mining algorithms

The specific steps of the data analysis oriented path planning algorithm designed in this paper are as follows:

Step 1: The candidate path sequence C_1 is generated by using the logistic path sequence database S with the length of 1.

Step 2: Scan database S, get the number of occurrences of each item in C_1, and generate frequent path sequence L_1. The number of elements in L_1 is less than 2, while items that do not satisfy min_sup constraints are deleted from database S.

Step 3: Generate excessive candidate sequence C'_2 and candidate path sequence C_2 with length 2 from C'_2.

The algorithm only traverses the original database once, which reduces the computational

complexity. By calculating the excessive candidate sequence C'_k, the frequent path sequence L_{k-1} is used to optimize the selection of C'_k, and the items that do not satisfy the minimum support constraint in C'_k are deleted. The number of candidate path sequences C_k generated is small, which greatly reduces the path sequence database.

3. Case Study

Database D is a logistics activity database, the specific content is shown in Table 1. Because of the large amount of data, only a part of the content is listed.

Table 1. Logistics activities

Contract	Customer number	Enterprise	Time	Place	Category of goods
1	13	222	2019-06-01	Jinan	A
2	24	333	2019-06-02	Wuhan	A
3	35	666	2019-06-03	Nanjing	A
4	46	888	2019-06-04	Guangzhou	A
…	…	…	…	…	…

Assuming that logistics enterprises distribute goods A, considering their time, customer number, path information and other data, the path sequence database D is collated, as shown in Table 2.

Table 2. Logistics route sequence

Sid	S(Logistics route sequence)
1	Wenzhou, Beijing
2	Shijiazhuang, Lianyungang, Ningbo
3	Wenzhou, Chengdu, Ningbo, Shijiazhuang
4	Wenzhou, Shijiazhuang, Ningbo
5	Lianyungang, Shijiazhuang

The data mining technology introduced in this paper is used to mine frequent path sequence in path sequence database D. The algorithm generates targeted over-candidate path sequences, which can effectively reduce the number of candidate path sequences after optimization and screening. In Table 4, the excessive candidate path sequence $C'_2 = 10$, and the optimized candidate path sequence $C_2 = 7$.

4. Experiments and Results Analysis

In order to further compare the MCP algorithm and Apriori-based basic algorithm for mining closed paths by mining location sequence and time sequence. This paper mainly compares the mining time of the same database with different minimum support, and the mining time of the path database with different sizes.

Comparing the M CP time of the two strategies under different support levels in the same path database, the experimental results are shown in Figure 2. The basic algorithm in Experiment 1 is based on a priori's idea, and some measures to check the closeness are added. The number of paths in the path database of Experiment 1 was 100,000, and the support was 0.15%-11.5%.

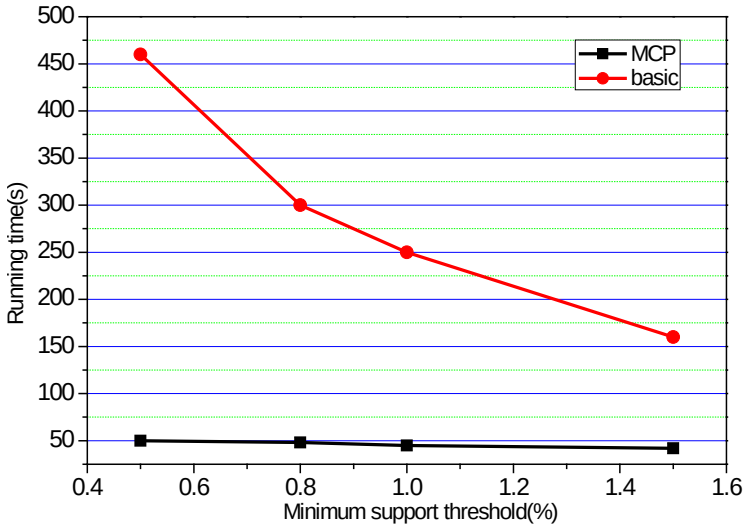

Figure 2. Runtime comparison under different minimum support thresholds

From Figure 2, we can see that the maximum length of frequent paths is relatively large when the support is small, and the advantage of MCP algorithm is obvious. When the support is large, the maximum length of frequent paths is smaller, the number of database scans by basic algorithm is reduced, and the advantage of MCP algorithm is decreased. For MCP, no matter how the length of the maximum closed path changes, the algorithm only scans the database twice, so the running time of the algorithm is small.

Test the scalability of MCP algorithm. The experimental results are shown in Figure 3. The min S up used in the experiment is 0.05 and 0.10 respectively. The number of paths contained in the path database is expanded from 500,000 to 2,000,000. It can be seen that the running time increases linearly with the increase of the number of paths in the path database.

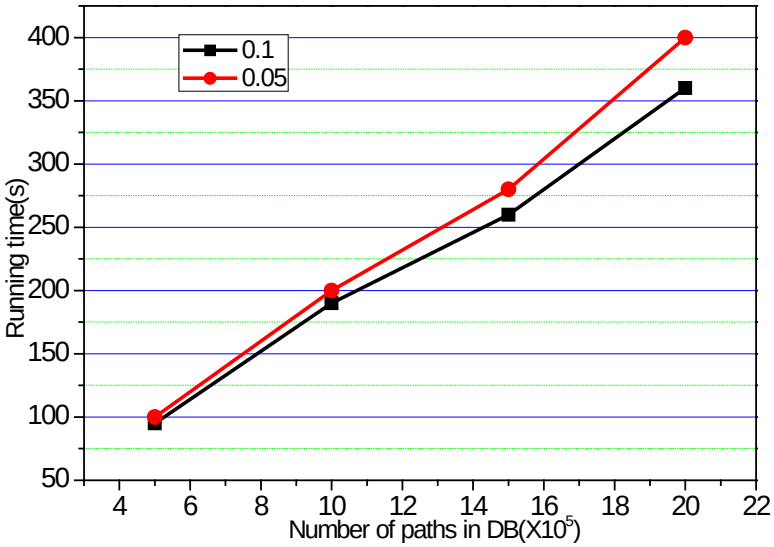

Figure 3. Running time of MCP under different path database sizes

5. Conclusion

At present, data mining technology has been developed rapidly, and has been widely used in finance, electricity, government and other fields. However, the application of this technology in logistics industry is not very mature. With the rapid development of information technology, logistics enterprises have shifted their focus from traditional internal integration to external integration when formulating strategic policies, and their requirements for the efficiency of path planning methods have gradually increased. Therefore, a new path planning algorithm is proposed in this paper. The algorithm scans the logistics path information database, carries out initial

screening, obtains the preliminary path sequence, then generates the targeted candidate excessive path sequence, optimizes the candidate path sequence, and then obtains the frequent path sequence. Association technology can mine the rule relationship between data, based on existing data, and forecast the future development trend. The application of this technology in logistics system can analyze the flow direction of logistics goods and grasp the dynamic operation state of logistics system.

References

[1] Chen D, Chen Y, Han B. Toll Policy for Load Balancing Research Based on Data Mining in Port Logistics [J]. Journal of Coastal Research, 2015, 73:82-88.

[2] Zhao W, Chen J J, Perkins R, et al. Erratum to: A novel procedure on next generation sequencing data analysis using text mining algorithm [J]. BMC Bioinformatics, 2016, 17(1):301.

[3] Xiaozhi R, Binchao B. Research and Application of Optimization Algorithm for Tobacco Logistics Delivery Route Based on Cluster Analysis [J]. Tobacco Science & Technology, 2015, 48(1):90-95.

[4] Wang J, Liu Z, Li W, et al. Research on a frequent maximal induced subtrees mining method based on the compression tree sequence[J]. Expert Systems with Applications, 2015, 42(1):94-100.

[5] Sun, Qiang. Empirical research on coordination evaluation and sustainable development mechanism of regional logistics and new-type urbanization: a panel data analysis from 2000 to 2015 for Liaoning Province in China[J]. Environmental Science and Pollution Research, 2017, 24(16):14163-14175.

[6] Ding S, Wu F, Qian J, et al. Research on data stream clustering algorithms[J]. Artificial Intelligence Review, 2015, 43(4):593-600.

[7] Stockton D J, Riham Ashley Khalil…. Cost model development using virtual manufacturing and data mining: part II-comparison of data mining algorithms[J]. International Journal of Advanced Manufacturing Technology, 2013, 66(9-12):1389-1396.

[8] Apte A, Jayasuriya A, Kennington J, et al. Class scheduling algorithms for Navy training schools[J]. Naval Research Logistics, 2015, 45(6):533-551.

[9] Cai Y, Zheng S, Ma Z. Research on Agricultural Product Logistics Efficiency and Market Factors Based on Provincial Panel Data[J]. Journal of Computational and Theoretical Nanoscience, 2016, 13(12):9804-9809.

[10] Li C, Gaukler G M, Ding Y. Using container inspection history to improve interdiction logistics for illicit nuclear materials[J]. Naval Research Logistics, 2013, 60(6):433-448.

[11] Hu Z H, Zhou J X, Zhang M J, et al. Methods for ranking college sports coaches based on data envelopment analysis and PageRank[J]. Expert Systems, 2015, 32(6):652-673.