TEST Engineering & Management

*July - August 2020*
*ISSN: 0193-4120 Page No. 5433 - 5439*

# Data Mining Simulation of Frequent Logistics Routes Relying on Wireless RFID Recognition

**Fengju Hou**[1,*]

[1]School of Business Administration, Zibo Vocational Institute, Zibo, Shandong, China, 255314

***Article Info***
***Volume 83***
***Page Number: 5433 - 5439***
***Publication Issue:***
***July - August 2020***

***Article History***
***Article Received:*** *25 April 2020*
***Revised:*** *29 May 2020*
***Accepted:*** *20 June 2020*
***Publication:*** *28 August 2020*

***Abstract***

The technology in the data mining simulation of frequent logistics routes based on wireless RFID identification has effectively solved the problem of obtaining the optimal path in mass frequent through application diagnosis. Other solutions for separable eigenbeamformers, such as instantaneous indirect trade-offs, cannot solve the problem in an effective way. The successful development of data mining simulation of frequent logistics routes that rely on wireless RFID identification will solve the problem of logistics route selection and improve the efficiency of logistics work.

***Keywords:*** *RFID Recognition, Frequent Paths, Sequential Patterns, Data Mining;*

## 1. Introduction

A large amount of data is generated in logistics, and RFID data is one of them. This paper designs the frequent sequence algorithm FSPMA to mine frequent logistics route data [1-3]. Extracting useful information such as the route status of the vehicle movement and the future movement trend based on the algorithm analysis can assist the manager to make accurate decisions, and at the same time facilitate traffic planning and design. The important significance of frequent pattern mining on RFID path data From the overall perspective, at a macro level, we can understand where frequent vehicle movements are mainly distributed, and at a deeper level, we can dig out important information such as how the road is distributed [4-6]; On the micro level, we can know whether there are vehicles passing through this node frequently, and a deeper point can be found through which frequent route the vehicle enters this node, and which frequent route leaves it. Only the analysis of a single node can't tell these Information needs to be analyzed from an overall perspective.

## 2. Construction of logistics route data model for efficient mining

### *2.1. Related concepts of frequent pattern mining*

For the convenience of description, this article first introduces several related concepts of frequent pattern mining.

Itemset: A collection of items. A collection of k itemsets is called K itemsets. For example, a collection of A, B, and C3 items is called a 3-item set, which is denoted as ABC.

Sequence: The items of the itemset are arranged in a certain order to form a sequence, and the K sequence is the arrangement of K items. For example, a sequence containing 3 items of A, B, and C, and the sequence of which is B, C, and A is called a 3 sequence and is denoted as BCA.

Support: The frequency of occurrence of an item set (sequence) in the data set, expressed as a percentage or a decimal.

Minimum support: the minimum requirement for the support of itemsets (sequences).

Frequent pattern: itemsets (sequences) whose support is not lower than the minimum support, also called frequent itemsets (sequences).

Path sequence: The items in the sequence represent the sequence of locations. For example, A, B, and C represent 3 locations. The sequence from B



*Published by: The Mattingley Publishing Co., Inc.*

July - August 2020
ISSN: 0193-4120 Page No. 5433 - 5439

to C to A is a sequence of paths, which is denoted as BCA.

Frequent paths: a sequence of paths that meet the minimum support.

Candidate set (sequence): Itemsets (sequences) that may become frequent patterns.

*2.2. Matrix-based data model in traditional frequent pattern mining*

The time consumption of the frequent pattern mining algorithm is mainly in the scanning of the data set. There have been many related studies on how to improve the scanning efficiency of the data set, and the 0-1 matrix representation is generally used. Each column of the matrix is mapped to an item, and each row corresponds to an item set in the data set. In this row, 0 indicates that the corresponding item is not in the current item set, and 1 indicates that it is in the current item set. As listed in Table 1, each column maps items A, B, C, D, and E respectively, and the item sets represented by the second and third rows in Table 1 are ACE and BCD respectively. The adoption of this data structure that can be randomly searched turns the string search in the candidate set scan into a simple logical judgment. For example, when judging whether the record ACE shown in the second row of table 1 contains items A and E, it is not necessary to search the string ACE item by item, but only to judge whether the result of the logical AND operation of the column values corresponding to A and E is 1. , 1 means that items A and E exist; otherwise, they do not exist.

**Table 1.** An example of a 0-1 matrix of the data set.

| A | B | C | D | E |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 |

In the mining of frequent sequence patterns, this kind of data model is no longer applicable, and sequence information cannot be represented simply by 0 and 1, so inefficient string search is usually used in the mining of frequent sequence patterns. However, due to the unique temporal and spatial attributes of the logistics path sequence, the same items will not appear in the sequence. Based on this feature, this paper establishes a logistics data model, in which the traditional logistics data is processed, and corresponding changes are made on the basis of the 0-1 matrix, and the traditional RFID route data is transformed into a kind of route sequence information. And it is conducive to efficiently mining data sets of frequent paths. Next, first establish a logistics route data model, and then introduce the frequent route mining algorithm FSPMA.

*2.3. A data model suitable for frequent path mining in the logistics field*

The traditional logistics data is RFID data, which can be expressed as {EPC, Location, Time}, where EPC (Electronic Product Code) is a unique mark of the item, and Location is the location at time Time (Time does not use conventional time here) Representation, but symbolically represented by numbers). Among these data, only those data whose locations represented by Location are representative starting points or transit hubs are selected. These representative points are the nodes A, B, C, D, and EF in the logistics network shown in Figure 1. The edges in Figure 1 are the paths between the nodes. Examples of RFID data related to this figure are listed in Table 2. . In addition, a table containing certain attributes of items can be established according to the mining needs, and multi-dimensional frequent path mining can be performed while realizing the separation of path sequence and attributes. For example, Table 3 is an attribute table of item type (KindID) and logistics weight (Weight), where logistics weight is a measure of route frequency, that is, the larger the weight, the more frequent the corresponding route. If these two attributes are combined for mining, the frequent paths of a certain type of item can be obtained. In order to simplify the problem, the following experiments only considered the logistics weight when comparing the efficiency, but did not consider other attributes such as item types.



Published by: The Mattingley Publishing Co., Inc.

July - August 2020
ISSN: 0193-4120 Page No. 5433 - 5439

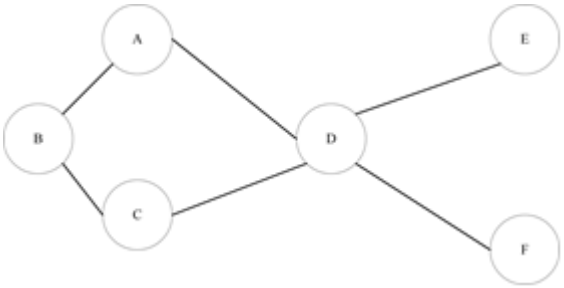


**Figure 1.** RFID data example of logistics network topology.

The RFID data is classified by EPC, and the values corresponding to the Location column in the records with the same EPC are sorted into {epc_x, location_l, location_2,..., location_i} according to the size of Time (ie, the order of time), where location_l to location_i are tables The value of EPC in 2 is the value of the Location of epc_x, and it is sorted in chronological order. This sequence is the RFID path sequence composed of the items with epe_x as EPC and the locations passed by chronological order.

### 3. Frequent path mining algorithm FSPMA based on logistics network topology information

Through the analysis of the logistics network, we know that A→B and B→A are different in logistics, and there is no inclusion relationship between A→B→C and A→C. A→C is not a child of A→B→C. Sequence is another irrelevant path, so the mining of frequent logistics paths is different from the classic association rule mining. It is a special sequence pattern mining. Since AB and BA are not the same sequence, if no pruning is performed, the number of candidate k sequences generated is $A_n^k$ instead of $C_n^k$ (n is the number of nodes in the network). Fortunately, in this special network, in addition to the traditional Apriori feature, the pruning method can also design a corresponding pruning algorithm based on the characteristics of the logistics network itself.

The process of the traditional Apriori-like algorithm for mining frequent sequences is mainly divided into two alternate steps: 1) Obtain candidate K sequences, which will use related pruning algorithms to reduce the generation of candidate K sequences; 2) Scan The data set is used to determine whether the candidate sequence is frequent, so as to obtain the frequent K sequence, and then go to step 1) to obtain the candidate K+1 sequence, and so on, until no longer sequence can be found. Among them, from step 2) to step 1), the result of step 2) is used. According to the feature that any sub-pattern of the frequent pattern is frequent, the candidate K+1 sequence is pruned, and it is impossible to remove K+1 sequence of frequent sequence. Obviously the most time-consuming of these two steps is step 2), and the one that directly affects the time-consuming degree of this step is step 1). The less frequent sequences generated in step 1), the less time-consuming step 2). The logistics network has its own characteristics, which are mainly reflected in two aspects in the mining of frequent routes.

1) Any two adjacent nodes in the sequence must be adjacent in the topological graph. For example, the AC sequence in Figure 1 is impossible to appear, only ABC or ADC may appear. Therefore, generating frequent K+1 sequences from frequent K sequences is different from traditional Apriori or other sequence pattern mining algorithms. As long as two K sequences that meet the requirements exist, the corresponding K+1 sequences can be connected to generate the corresponding K+1 sequences. These two sequences are satisfied: the K-1 sequence composed of the first K-1 nodes of one sequence is exactly the subsequence of the other sequence excluding the first node. For example, if ABC and BCD are frequent, you can connect them Get the candidate sequence ABCD;

2) The topology information of the logistics network can be used for pruning candidate frequent sequences. Since logistics tends to take the path with the least cost, theoretically, if a path between two points is frequent, then all paths with a cost less than or equal to the frequent path cost between the two points are also frequent. In the logistics network, there are one or more minimum cost paths between



Published by: The Mattingley Publishing Co., Inc.

July - August 2020
ISSN: 0193-4120 Page No. 5433 - 5439

two points. No matter whether the path is frequent or not, the cost exceeds the minimum cost so much that the unacceptable path will not be a frequent path. The specific excess is considered too much, which requires specific analysis of specific issues.

This paper defines a "cost tolerance" parameter TD for the path sequence Rij from node i to node j in the logistics network. TD represents the ratio of the cost of Rij to the minimum path cost from node i to node j. Obviously the ratio is greater than Equal to 1. When generating the candidate path sequence, the upper limit of the cost tolerance of the path that meets the requirements can be set, that is, all path sequences whose cost tolerance exceeds the upper limit are removed. For example, if the upper limit is set to 1.2, that is, all path sequences whose cost exceeds 20% of the minimum cost will be pruned.

The basic topology information of the logistics network, including the number of nodes in the network and the adjacency matrix of the network. Assuming that the number of nodes in the network is N, the corresponding adjacency matrix neib is an N×N matrix. In the implementation of the algorithm, neib is an N×N two-dimensional array, where neib[i][j] represents node i and node The adjacency distance (cost) of j. If neib[i][j] is equal to 0, it means that node i and node j are not adjacent. The shortest path algorithm can be used to find the distance between any two points from the adjacency matrix. The calculation result is similar to the adjacency matrix and is also represented by a matrix, which is called the minimum cost matrix here. In the algorithm implementation, the minimum cost matrix is represented by an N×N two-dimensional array least, where least[i][j] represents the cost of the shortest path between node i and node j, if zeast[i][j] is equal to 0, then Indicates that there is no path between node i and node j. The candidate sequence pruning method introduced in FSPMA in this paper is here referred to as the cost tolerance pruning method. The pruning method is: first obtain the cost cost of the path represented by the candidate sequence; then calculate the cost tolerance parameter (note Is the product of TD) and the minimum cost of the two points connected to the path ultimate. Suppose the path is the path from node i to node j, as shown in equation (1):

$$ultimate=least[i][j]\times TD \quad (1)$$

Ultimate is the maximum acceptable path cost from node i to node j; finally, cost is compared with ultimate. If cost>ultimate, the corresponding candidate sequence does not meet the requirements and must be removed from the candidate sequence.

The PMWTI algorithm is described as follows:

Input: logistics path data set, minimum support SUP, cost tolerance parameter TD, basic topology information in the logistics network.

Output: a collection of all frequent paths that meet the support degree SUP.

Step1 reads the data set and converts it into a matrix as listed in Table 5 and stores it in the main memory.

Step2 read the basic topology information of the logistics network, obtain the adjacency matrix neib and the number of nodes N, and then calculate the least cost matrix least according to the neib.

Step3: Obtain candidate K sequences. If the set of candidate K sequences is empty, the mining ends; otherwise, step4 is executed. The candidate sequence for the first execution of this step is the set of all nodes in the logistics network; if it is not executed for the first time, first obtain the preliminary candidate k sequence through the frequent K-1 sequence connection obtained in step 4 (the connection method has been described above), and then Use the cost tolerance pruning method for this sequence.

Step4 scans the matrix in step1 to find frequent sequences from candidate sequences. If the required frequent sequence set is empty, the mining ends, otherwise, go to step 3.

PMWTI introduces a minimum cost matrix, denote the number of nodes as n, then the time and space costs required for the calculation and storage of this matrix are O(n3) and O(n2), respectively.



Published by: The Mattingley Publishing Co., Inc.

July - August 2020
ISSN: 0193-4120 Page No. 5433 - 5439

Normally, the cost in frequent path mining The comparison is negligible, unless the minimum support is set too large, causing the algorithm to end soon, and no longer patterns are mined. Of course, this mining is meaningless.

## 4. Experimental analysis

The experiment is only performed once. The data set used for the experiment has a total of 1,206,480 records. The sum of the logistics weights of all records is 480800030, and the average weight of each record is 2328.6. Taking the minimum support as 0.001 as an example, that is If a route is frequent, the sum of its logistics weights must be greater than or equal to 480800030×0.001. First test the data set, which uses two algorithms. The first is the algorithm FSPMA proposed in this article; the second is the traditional Apriori-like mining algorithm, which is an algorithm composed of the FSPMA algorithm description and the part that uses the topological information for pruning, which is marked as TRADITIONAL here. The results of mining using FSPMA are compared with the results of TRADITIONAL mining under the same minimum support. FSPMA uses a series of cost tolerance, and the cost tolerance values are 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0. The minimum support degrees are 0.001, 0.003, o.005, 0.007, 0.01, 0.03, 0.05, 0.07. The comparison result is shown in Figure 2, where the abscissa is the minimum support, and the ordinate is the minimum cost tolerance to ensure the consistency of the mining results of the two algorithms. For example, when the minimum support is 0.001, the minimum cost tolerance is 1.6, that is, the mining results of FSPMA with a cost tolerance of 1.6 and above are consistent with the mining results of TRADITIONAL, and the mining results of 1.5 and below are inconsistent with the mining results of TRADITIONAL. Analysis of the mining results shows that the length of the longest frequent path excavated from the minimum support degree from 0.001 to 0.07 is 9, 8, 7, 6, 5, 3, 1, and 1 in order. The reason why the minimum cost tolerance is as high as 1.6 when the minimum support is 0.001 is because the minimum support is too low relative to the data set used in the experiment, so that almost all paths become frequent paths, and the mining results are meaningless. In this case It cannot give full play to the advantages of FSPMA. When the minimum support is increased from 0.001 to 0.003, the minimum cost tolerance is reduced from 1.6 to 1.3, that is, when the minimum support is slightly increased and the excavated frequent paths have a certain meaning, the effect of FSPMA is well reflected. With the increase of the minimum support, the minimum cost tolerance is gradually reduced to 1.0, which means that when the minimum support is high to a certain level, either the frequently excavated paths are the minimum cost paths, or no path meets the minimum support For example, when the minimum support is 0.05 and 0.07, the mined nodes are all single nodes instead of path sequences.

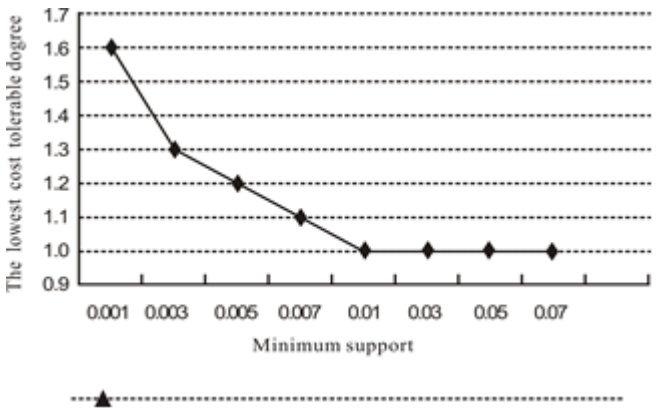


**Figure 2.** The minimum cost tolerance values under different minimum support degrees.

The efficiency comparison between FSPMA and TRADITIONAL is shown in Figure 3, where the abscissa is the minimum support degree, the ordinate is the time used for excavation (in seconds), and the cost tolerance of FSPMA is 1.6. Obviously, the efficiency of FSPMA is far superior to TRADITIONAL under any degree of support. TRADITIONAL, which does not consider any logistics network topology information at all, performs very poorly in efficiency when digging



Published by: The Mattingley Publishing Co., Inc.

July - August 2020
ISSN: 0193-4120 Page No. 5433 - 5439

frequent routes from logistics data. The two are not comparable. After analysis, it is found that the main problem of TRADITIONAL lies in the generation of candidate 2 sequences. Since the candidate 2 sequences obtained by frequent 1 sequences in TRADITIONAL are obtained by permutation, that is, if there are n frequent 1 sequences, n (n-1) candidate 2 sequences will be generated. However, some of the candidate 2 sequences are impossible to appear in the data set, because the path sequence in the logistics network has the attribute of path, and any two adjacent points in the path must be adjacent in the logistics network topology (here This sequence is called a legal sequence, and it is impossible for a sequence composed of all non-adjacent nodes to appear in the logistics data set (here, the sequence that is adjacent in the sequence but not adjacent in the logistics network is an illegal sequence). This problem can be solved by introducing basic topological information. The third algorithm is proposed here, which is to add the parameter of the adjacency matrix neib to TRADITIONAL, and judge according to neib when generating the candidate 2 sequence ij, if node i and node If j is not adjacent, remove candidate 2 sequence ij; otherwise, add it to candidate 2 sequence. The reason why this judgment is only added when generating the candidate 2 sequence is that the connection generation of the candidate K sequence (K>2) is based on the legal sequence, and all the sequences obtained through the connection of the legal sequence are legal sequences. The candidate sequence naturally obtains this pruning effect during the generation process, and the algorithm is referred to as NATURAL here. Compared with TRADITIONAL, NATURAL has a significant improvement in efficiency and is closer to FSPMA. Therefore, the comparison data given below only includes the experimental data of NATURAL and FSPMA, and TRADITIONAL is no longer involved in the comparison.

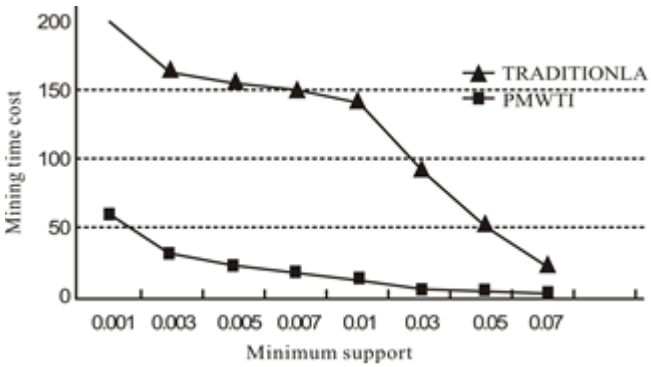


**Figure 3.** Comparison of the efficiency of FSPMA and TRADITIONAL with different minimum support.

First, compare FSPMA and NATUAL when the minimum support is 0.003, as shown in Figure 4. The ordinate represents the time used by the algorithm in seconds; the abscissa represents the cost tolerance. Since the cost tolerance is only valid for FSPMA, NATURAL does not have this parameter, so NATURAL in Figure 4 takes a constant 37.328 seconds. It can be seen from Figure 4 that with the continuous improvement of cost tolerance, the time-consuming FSPMA gradually approaches NATURAL; while the contemporary price tolerance is too low, although the mining efficiency of FSPMA is very high in comparison, the mining results may be compared with conventional mining algorithms. The results are inconsistent, so it is necessary to select an appropriate cost tolerance.

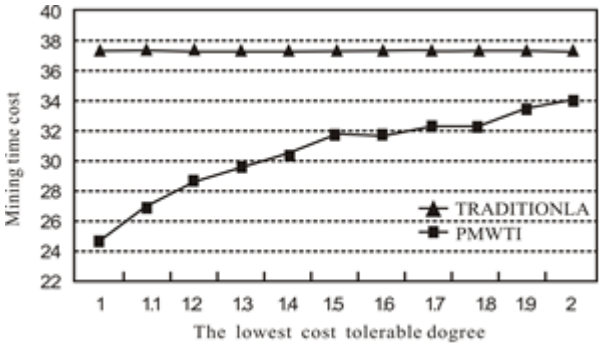


**Figure 4.** Comparison of NATURAL and FSPMA mining efficiency under different cost tolerances.

## 5. Conclusion

Aiming at the characteristics of logistics data, this paper designs a corresponding logistics data model



Published by: The Mattingley Publishing Co., Inc.

*July - August 2020*
*ISSN: 0193-4120 Page No. 5433 - 5439*

and mining algorithm FSPMA in order to efficiently mine frequent routes from these data. The data model and mining algorithm are independent of each other. The data model is based on the exchange of space for time. A general candidate path sequence pruning strategy is designed in the algorithm FSPMA—the cost tolerance pruning method. Compared with the equivalent algorithm without the pruning method, this strategy reduces the scanning of the data set and effectively improves the mining efficiency.

**References**

[1] Lapkin A A, Heer P K, Jacob P M, et al. Automation of route identification and optimisation based on data-mining and chemical intuition [J]. Faraday Discussions, 2017, 202(26): 1-10.

[2] Nezihe Yıldıran, Tacer E. Identification of photovoltaic cell single diode discrete model parameters based on datasheet values [J]. Solar Energy, 2016, 4(6): 101-111.

[3] Cho J S, Jeong Y S, Park S O. Consideration on the brute-force attack cost and retrieval cost: A hash-based radio-frequency identification (RFID) tag mutual authentication protocol [J]. Computers & Mathematics with Applications, 2015, 69(1): 58-65.

[4] Chen Y, Zhou F, Liu X, et al. Online adaptive parameter identification of PMSM based on the dead-time compensation [J]. International Journal of Electronics, 2015, 102(7-9): 1132-1150.

[5] Janosovsky J, Danko M, Labovsky J, et al. Software approach to simulation-based hazard identification of complex industrial processes [J]. Computers & Chemical Engineering, 2018, 122(4): 66-79.

[6] Nguyen C T, Bui A T H, Nguyen V D, et al. Modified tree-based identification protocols for solving hidden-tag problem in RFID systems over fading channels [J]. Iet Communications, 2017, 11(7): 1132-1142.