The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2632-0487.htm

# Railway wagon flow routing locus pattern intelligent recognition algorithm based on SST

Xiaodong Zhang
*Research and Application Innovation Center for Big Data Technology in Railway, China Academy of Railway Sciences, Beijing, China*

Ping Li
*Institute of Computing Technology, China Academy of Railway Sciences, Beijing, China, and*

Xiaoning Ma and Yanjun Liu
*Research and Application Innovation Center for Big Data Technology in Railway, China Academy of Railway Sciences, Beijing, China*

Received 17 October 2019
Revised 14 January 2020
Accepted 19 February 2020

## Abstract

**Purpose** – The operating wagon records were produced from distinct railway information systems, which resulted in the wagon routing record with the same oriental destination (OD) was different. This phenomenon has brought considerable difficulties to the railway wagon flow forecast. Some were because of poor data quality, which misled the actual prediction, while others were because of the existence of another actual wagon routings. This paper aims at finding all the wagon routing locus patterns from the history records, and thus puts forward an intelligent recognition method for the actual routing locus pattern of railway wagon flow based on SST algorithm.

**Design/methodology/approach** – Based on the big data of railway wagon flow records, the routing metadata model is constructed, and the historical data and real-time data are fused to improve the reliability of the path forecast results in the work of railway wagon flow forecast. Based on the division of spatial characteristics and the reduction of dimension in the distributary station, the improved Simhash algorithm is used to calculate the routing fingerprint. Combined with Squared Error Adjacency Matrix Clustering algorithm and Tarjan algorithm, the fingerprint similarity is calculated, the spatial characteristics are clustering and identified, the routing locus mode is formed and then the intelligent recognition of the actual wagon flow routing locus is realized.

**Findings** – This paper puts forward a more realistic method of railway wagon routing pattern recognition algorithm. The problem of traditional railway wagon routing planning is converted into the routing locus pattern recognition problem, and the wagon routing pattern of all OD streams is excavated from the historical

© Xiaodong Zhang, Ping Li, Xiaoning Ma and Yanjun Liu. Published in *Smart and Resilient Transportation*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Research supported by the National Key Research and Development Plan "Advanced Rail Transit" special project of Ministry of Science and Technology of the People's Republic of China (2018YFB1201403) and the important scientific research topics of China Academy of Railway Sciences (ADZ180142).

*Conflict of Interests:* The authors declare that there is no conflict of interests regarding the publication of this paper.

Smart and Resilient Transportation
Vol. 2 No. 1, 2020
pp. 3-21
Emerald Publishing Limited
e-ISSN: 2632-0495
p-ISSN: 2632-0487
DOI 10.1108/SRT-10-2019-0008

data results. The analysis is carried out from three aspects: routing metadata, routing locus fingerprint and routing locus pattern. Then, the intelligent recognition SST-based algorithm of railway wagon routing locus pattern is proposed, which combines the history data and instant data to improve the reliability of the wagon routing selection result. Finally, railway wagon routing locus could be found out accurately, and the case study tests the validity of the algorithm.

**Practical implications** – Before the forecasting work of railway wagon flow, it needs to know how many kinds of wagon routing locus exist in a certain OD. Mining all the OD routing locus patterns from the railway wagon operating records is helpful to forecast the future routing combined with the wagon characteristics. The work of this paper is the basis of the railway wagon routing forecast.

**Originality/value** – As the basis of the railway wagon routing forecast, this research not only improves the accuracy and efficiency for the railway wagon routing forecast but also provides the further support of decision-making for the railway freight transportation organization.

**Keywords** Intelligent transportation, Pattern recognition, Simhash algorithm,
Wagon flow routing, Similarity matrix, Clustering analysis

**Paper type** Research paper

## 1. Introduction

Railway wagon routing selection and allocation are the vital basic work in the railway freight transport organization, which is the principal reference for the railway wagon flow forecast, adjustment and daily plan making. The reasonable planning of the railway wagon routing and the accurate tracking of the railway wagon flow *locus* is the key to guide the railway freight production. However, there is a situation that actual wagon routing in the railway network does not match the planning wagon routing, which makes it hard for relevant departments to forecast the wagon flow, to trace the wagons, to get the delivered time and to optimize the freight organization in the long-distance transportation. The wagon operating records are produced from distinct railway information systems, which resulted in that the wagon routing record with the same oriental destination (OD) may be different. Some of the phenomena were because of poor data quality, while others were because of the existence of another actual wagon routings. It has already been a common problem that the actual routing trajectory of railway wagon flow was not in line with the planning routing. The planning wagon routing could not forecast wagon flow routing anymore because of the above phenomenon. Therefore, how to take full advantage of the big data of railway wagon flow accumulated by *Railway Transport Information Integration Platform* and recognize wagon routing *locus* patterns intelligently by mining the history information of wagon routing has become the chief issue of railway wagon flow routing forecast.

Experts and scholars in the field of railway transportation have made extensive research on railway wagon routing and achieved fruitful results. In the traditional research, the railway network consisting of railway lines and railway stations was defined as the directed network or the directed graph (Xu, 2001). The railway wagon routing vertex-to-arc models were classified based on two hypotheses: unlimited railway network capacity (Xu, 2001; Shen and Deng, 2003) and limited railway network capacity (Xu, 2001; Shen and Deng, 2003; Ford *et al.*, 2001). To solve the model, Dijkstra algorithm and Floyd algorithm were two major computing methods of wagon routing planning. Nevertheless, with the expansion of the national railway network scale, the execution efficiency of the above algorithm was concerned.

Besides the railway network restriction capacity, the design of the objective function was another decisive factor. Railway wagon routing problem was optimized by different goals, such as the minimum operating mileage (Wen *et al.*, 2016; Jiang *et al.*, 2004), the minimum operating cost (Wang *et al.*, 2014), the maximum merchandise revenue (Ji *et al.*, 2011), the lowest transportation cost (Jin *et al.*, 2005) and the shortest transportation time (Chang, 2010). The diverse goal of the model decided the operation effect of wagon routing selection. However, the defects in this study above are also obvious that one single goal narrowed the boundaries of the problem and reduced the complexity of the study. Therefore, multi-objective optimization models for wagon routing selection were designed thoroughly by Shi and Shi (1998), Shi and Yong (1999) and Wu *et al.* (2016).

With the development of international trade, scholars pay more attention to the railway-based multimodal transportation. For multimodal transport, the optimized railway wagon routings can lead to higher transport efficiency. The common method of multi-modal transportation freight routing planning problem contains column generation method (Caprara *et al.*, 2011) and genetic algorithm (Xiong and Wang, 2014), which are established on the mixed-integer linear/nonlinear model. In the scene of multimodal transportation, the freight commodity is integrity (Liu *et al.*, 2011; Li *et al.*, 2012), railway network situation limits the network capacity (Cho *et al.*, 2012), transportation targets are different (K. Lei *et al.*, 2014) and transportation cases are spatio-temporal stochastic (Bai *et al.*, 2014; Wang *et al.*, 2011). The characteristics of multimodal transport routing planning are similar to the main two characteristics in the railway freight routing planning.

It is not difficult to find out that there are two significant problems in the traditional research. Both problems cause that we have to use the planning routing to forecast the actual wagon routing in the work of railway wagon flow forecast. The accuracy of wagon flow spatial prediction is low, and wagon flow time prediction is even more impossible.

*Problem 1:* The optimal goal based on the comparison of the shortest and the second shortest routing is often unrealistic. The shortest wagon routing planned does not mean the best, which can only be simply called the better choice in one aspect or one target.

*Problem 2:* The cost of wagon routing calculation by the multi-objective programming is too high. The planning result does not match the actual production, which implies that the planning wagon routing is meaningless to improve the railway transport efficiency.

With the emergence of new calculation and analysis method, the big data thinking is more widely used, which converts traditional causal relationship analysis to the strong correlation analysis. To bring the research results closer to reality, domestic and foreign experts use machine learning and data mining methods to analyze the behavior of learning objects and further guide the production.The big data analytic method used in the traffic and transportation field contained the multisource data fusion (Drličiak and Čelko, 2016; Necula, 2015), the visualization of big data (Cebon and Samson, 2012), online learning (Zhou *et al.*, 2015; Lint, 2008) and SVM (Jeong *et al.*, 2013; Chen *et al.*, 2016).

Based on the excavation and analysis of railway wagon flow big data, this paper puts forward a more realistic method of railway wagon routing pattern recognition algorithm, which supports the wagon flow forecast. The problem of traditional railway wagon routing planning is converted into the routing *locus* pattern recognition problem, and the wagon routing pattern of all OD streams is excavated from the historical data results. The analysis is carried out from three aspects: routing metadata, routing locus fingerprint and routing locus pattern. Then, the intelligent recognition SST-based algorithm of railway wagon routing *locus* pattern is proposed, which combines the history data and instant data to improve the reliability of the wagon routing selection result. Finally, railway wagon routing *locus* could be found out accurately, and the case study tests the validity of the algorithm.

The rest of this paper is organized as follows. Section 2 describes the scientific problem of the wagon routing selection based on the big data. The complex reasons for the huge mismatch between the actual railway wagon path and the planning track are analyzed in detail. Based on the big data of railway wagon flow, an SST-based algorithm is developed in

Section 3. The algorithm evaluation standard is proposed in Section 4 to verify the SST-based algorithm. In Section 5, the suggested algorithm is applied to the actual cases of railway wagon flow in the China Railway Corporation, and all the data are generated from the Railway Transport Information Integration Platform. Section 6 summarizes the full text and discusses the further issue in this research.

## 2. Railway wagon routing selection issue

There is a great gap between the behavior of the railway wagon flow and other traffic flow. The unique of the railway wagon flow is mainly manifested in integrity and convergence. Integrity means that the wagons of any OD flow during the transport procedure between loading wagon station and unloading wagon station are in one train, which could not move individually. The convergence refers that the original wagon will be marshalled into a train with the same OD in the railway technical station, which would not be ceaselessly spliced in the movement.

Two characteristics improve the efficiency of the railway transport organization. The main factors of wagon routing selection are as follows:

- transport distance;
- transport time;
- restrict capability of railway stations and networks; and
- transport costs.

The traditional research on wagon routing concentrated on the influential factors above to allocate the shortest wagon routing, aiming at optimizing the railway wagon flow organization. This method, as the object entity without choosing behavior, limits the scope of the influence factors, narrows the boundary of the problem and ignores the complex and changeable situation. The direct consequence is that the simulated results of the railway wagon routing do not match the actual wagon routing.

The reasons for this phenomenon are mainly as follows. In recent years, the scale of China railway network has expanded rapidly. The transportation capacity has also been greatly improved. However, compared with the actual domestic freight demand, there are still significant shortcomings, which are directly reflected in the lack of railway network capacity. The railway department formulates the freight plan based on the historical experience and railway network capabilities and plans the wagon routing. In the actual transportation process, the capacity of the marshalling stations is limited, the distributary station is blocking and the freight trains avoid passenger trains. These reasons have made it difficult for freight trains to accurately follow the planned wagon routing. In this case, to carry out the operation plan and to ensure the railway network unobstructed, the temporary adjustment of the wagon routing on some congested stations of the railway network is required. Then, it causes the actual wagon routing in the railway network to mismatch with the planning wagon routing, as shown in Figure 1, which may not provide decisive support for the transport organization and production.

In Figure 1, there are three loci from Pingdingshan (central city of Henan Province, China) East Railway Station to Sanming (northwestern city of Fujian Province, China) Railway Station: Locus 1 is the planning shortest wagon routing, Locus 2 and Locus 3 are the other two actual wagon routing, which means the difference is obvious and brings difficulties to the wagon flow forecast.

As shown in Figure 2, in the China railway network, the number of active wagons per day is about 600,000. The number of daily wagon report is about 3,000,000. The capacity of one report is about 1 KB. It can be observed that the reported capacity is approximately 3 GB by day and nearly 100 GB by month. All the records are stored in the Railway Transportation Information Integration Platform. Many railway vehicle data contain many influential factors, which reflect the discipline of the wagon routing, and have great significance for analyzing the wagon routing pattern. All the wagon routing patterns in the records are actual trajectory of the trains, while they may be different from the planning ones, which have brought practical difficulties to the wagon flow forecast in railway freight transportation.

According to the big data of the actual wagon routing, a hypothesis is proposed that the routing selection behavior mode of any wagon flow in the railway network has certain potential rules. Wagons with individual characteristics have obvious differences in the routing selection mode. In addition to the four factors above, wagon routing behavior mode is also affected by the natural environment, railway network status, management level, human factors, uncontrollable noise and other factors. In other words,behavior modes are determined by multiple influential factors. It is not hard to find that the traditional analysis method constructs multi-objective model for various influential factors to obtain the optimal wagon routing, which is not always suitable for the actual situation. However, ignoring some of the factors, the optimization result (the planning wagon routings) does not match the actual situation, which cannot support the transport organization. In recent years, Chinese railway has accumulated huge wagon operation data, which contains the wagon routing loci. Using railway wagon big data to excavate vehicle path mode is a new solution.

## 3. Intelligent recognition SST-based algorithm

The intelligent recognition SST-based algorithm consists of three important parts:

(1) routing metadata model description;

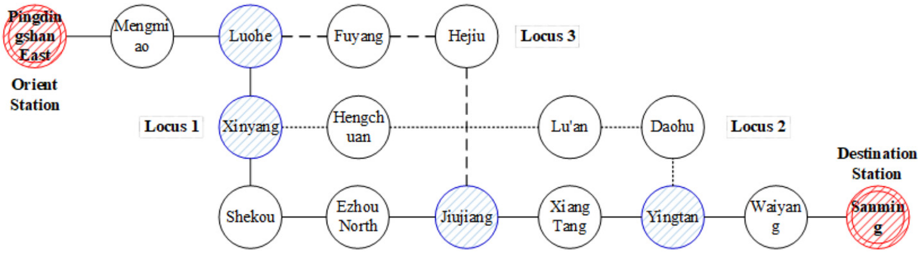


**Figure 1.** Pingdingshan East-Sanming actual wagon routing loci

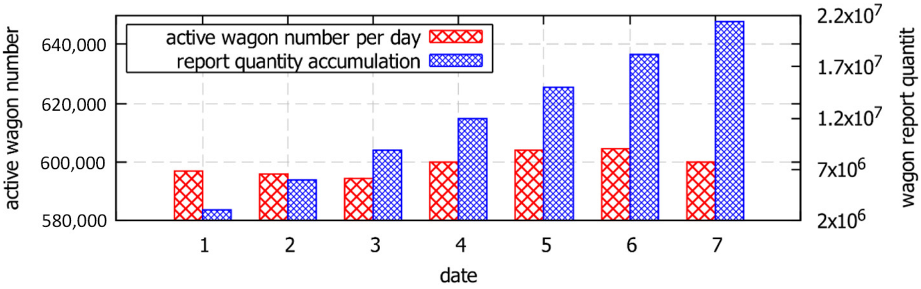


**Figure 2.** The daily active wagon number and seven-day quantity accumulation of wagon report in the China Railway Network

(2) routing fingerprint construction based on Simhash; and
(3) routing locus pattern mining and recognition.

Part 1 (Chapter 3.1) processes and fuses the wagon operating records in the different railway information systems, which form the complex wagon routing locus strings. Part 2 (Chapter 3.2) reduces the dimension of the complex wagon routing locus strings and keep the authenticity, which would improve the efficiency of the wagon flow routing locus recognition. Part 3 (Chapter 3.3) designs the recognition algorithm to excavate the actual wagon routing *locus* from the railway big data based on the above scheme. The total flow chart is shown as Figure 3.

### *3.1 Routing metadata model description*

*3.1.1 Routing metadata characteristic.* The railway wagon routing data is not stored centrally. When the wagon arrives in or departs from one railway station, the operation event may be recorded and stored separately in the database. It is hard to get the related information about the wagon routing directly. Therefore, it is necessary to use metadata model to process and analyze the railway wagon flow data.

Metadata is the description of the data that is used to process the data rapidly. The major class described by the routing metadata model is routing metadata. Routing metadata consists of four basic characteristics, $RM = (T, S, C, D)$, which provides services for the semantic recognition, the semantic matching and the wagon behavior decision of railway wagon routing.

Figure 4 points out the relationship of the characteristics in the routing metadata.

3.1.1.1 Time characteristic $T$. The characteristic $T$ is the value that represents the start or end time of a wagon behavior event.

3.1.1.2 Spatial characteristic $S$. The characteristic $S$ is the related attribute of the railway station. The attribute mainly refers to the geographical location where the wagon behavior

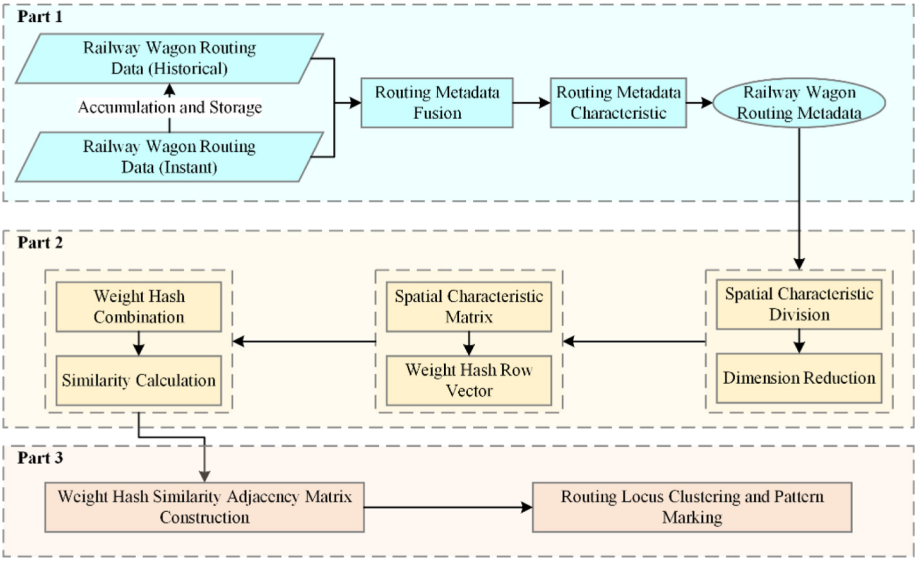

Part 1
Railway Wagon Routing Data (Historical)
Accumulation and Storage
Railway Wagon Routing Data (Instant)
Routing Metadata Fusion
Routing Metadata Characteristic
Railway Wagon Routing Metadata
Part 2
Weight Hash Combination
Similarity Calculation
Spatial Characteristic Matrix
Weight Hash Row Vector
Spatial Characteristic Division
Dimension Reduction
Part 3
Weight Hash Similarity Adjacency Matrix Construction
Routing Locus Clustering and Pattern Marking

**Figure 3.**
The total flow chart of the intelligent recognition SST-based algorithm

event occurred. The spatial feature uses the spatial basis data to establish the mapping table, which is spliced into the wagon routing *locus* as the chronological order, to make the foundation for mining the Content Feature C.

3.1.1.3 Content characteristic *C*. The characteristic *C* refers routing *locus* feature of the wagon flow. The wagon routing pattern indicates the OD routing pattern in routing metadata, which provides decision support for future wagon routing planning.

3.1.1.4 Data source characteristic *D*. The characteristic *D* is divided into instant wagon flow data and history wagon flow data. The instant data is reported immediately by the railway station and railway bureau. After analysis, the instant wagon flow data is transferred to the history wagon flow data. The history wagon flow data is used to build the wagon routing knowledge base, which enhances the variety of the study.

*3.1.2 Routing metadata fusion.* The real-time wagon flow data and historical wagon flow data are fused in the routing metadata model, and the fusion process is describedin the dashed box part in Figure 4, which is described in detail in Figure 5. Historical wagon flow data will no longer be unchangeable, which enriches the information value of the training data and enhances the credibility of the wagon routing analysis and selection.

Figure 5 reflects the fusion process of history data and real-time data. Real-time data submitted by reporting clients is the source of the historical data. Through data cleaning, classifying and weighting, real-time wagon flow data is fused with historical wagon flow data of different storage types, and the training data is constructed. The training data

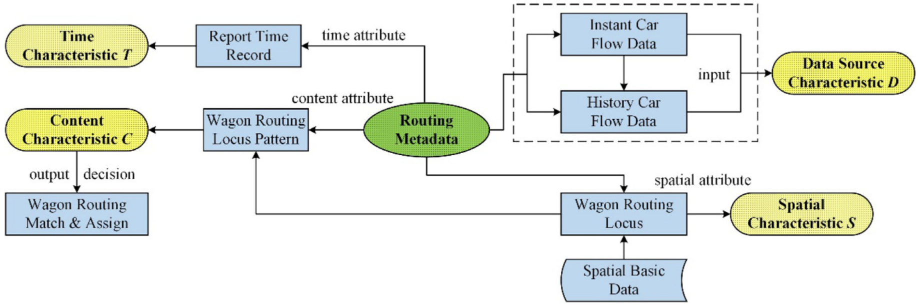


**Figure 4.**
The descriptive model of wagon routing metadata

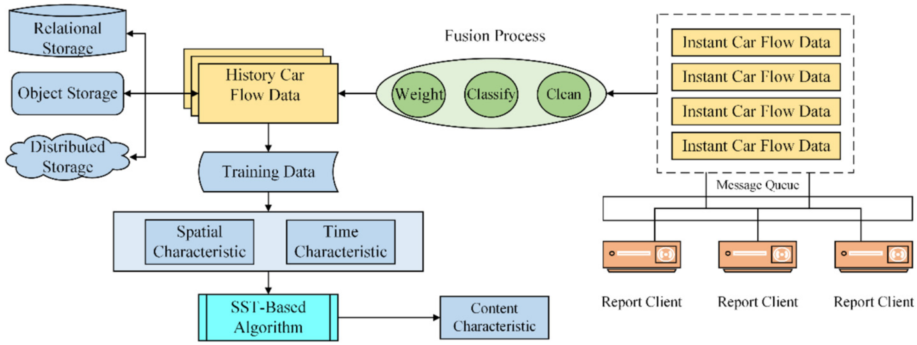


**Figure 5.**
Fusion process of the history data and instant data

provides spatial characteristics and time characteristics for the routing metadata. At this point, the data preparation has been prepared for extracting the content characteristics of the wagon routing *locus*.

The spatial characteristic of routing metadata contains all the basic data of passing stations, which are spliced into a completed character string, called *Wagon Routing Locus Character String* (WRLCS). The larger data length increases the computational cost and declines the efficiency of pattern recognition. In Figure 6, there are two stitching strings of the actual wagon routing *locus*, which means contrast calculation is huge.

In the field of semantic recognition, the word segmentation method is used to segment the text and give different weights according to the degree of importance to judge the similarity of the text. Based on this idea, the dimension of spatial characteristics, especially the WRLCS, are reduced by the Simhash algorithm. According to the different influencing factors, the higher dimensional eigenvector is mapped into the lower dimensional fingerprint, and metadata wagon routing mode can be excavated.

### *3.2 Routing fingerprint construction based on Simhash*

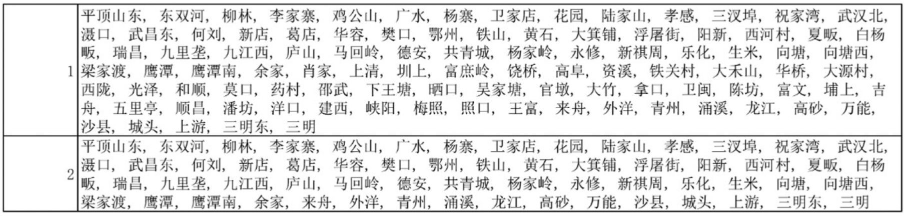
*3.2.1 Spatial characteristic division and dimension reduction.* The railway stations can generally be divided into two categories: *Distributary Station* and *Non-Distributary Station*. Distributary station refers to the point of the railway network, like *Luohe Railway Station* given in Figure 1. The L1 norm of the edge between the distributary point and its adjacent point in the railway network is greater than 2. On the contrary, *Susong Railway Station* is a non-distributary station. Assuming *Factor 3* would not exist, in the certain direction, the wagon routing WRLCS of any two neighbor distributary stations is the single one. That means all the elements between the neighbor distributary stations are not-distributary stations.

Assuming that the WRLCS is represented by a spatial characteristic $S$ of routing metadata $RM$, which consists of separator symbols and passing station symbols. These are $a-1$ distributary station and $b$ non-distributary station. The passing station quantity of the spatial characteristic $S$ is calculated by formula (1). The spatial characteristic $S$is divided into a vector space with $a$ eigenvector $V_a$. It is necessary to initialize $f$-dimensional fingerprint row vector $F$, and any element of $F$ is 0, descripted by formula (2).

$$\|S^{station}\|_1 = |a-1| + |b| + 2 \tag{1}$$

$$F = [0, 0, \cdots, 0]_{1\times f} \tag{2}$$

To improve the performance of data matching and mining, the eigenvector, which is composed of separator symbols and passing station symbols, is processed by the md5 hash algorithm, converting $V_a$ to *f-bit* 0-1 hash value $H_a$ (Chen *et al.*, 2016).



| | |
|---|---|
| 1 | 平顶山东，东双河，柳林，李家寨，鸡公山，广水，杨寨，卫家店，花园，陆家山，孝感，三汊埠，祝家湾，武汉北，滠口，武昌东，何刘，新店，葛店，华容，樊口，鄂州，铁山，黄石，大箕铺，浮屠街，阳新，西河村，夏畈，白杨畈，瑞昌，九里垄，九江西，庐山，马回岭，德安，共青城，杨家岭，永修，新祺周，乐化，生米，向塘，向塘西，梁家渡，鹰潭，鹰潭南，余家，肖家，上清，圳上，富庶岭，饶桥，高阜，资溪，铁关村，大禾山，华桥，大源村，西陇，光泽，和顺，莫口，药村，邵武，下王塘，晒口，吴家塘，官墩，大竹，拿口，卫闽，陈坊，富文，埔上，吉舟，五里亭，顺昌，潘坊，洋口，建西，峡阳，梅照，照口，王富，来舟，外洋，青州，涌溪，龙江，高砂，万能，沙县，城头，上游，三明东，三明 |
| 2 | 平顶山东，东双河，柳林，李家寨，鸡公山，广水，杨寨，卫家店，花园，陆家山，孝感，三汊埠，祝家湾，武汉北，滠口，武昌东，何刘，新店，葛店，华容，樊口，鄂州，铁山，黄石，大箕铺，浮屠街，阳新，西河村，夏畈，白杨畈，瑞昌，九里垄，九江西，庐山，马回岭，德安，共青城，杨家岭，永修，新祺周，乐化，生米，向塘，向塘西，梁家渡，鹰潭，鹰潭南，余家，来舟，外洋，青州，涌溪，龙江，高砂，万能，沙县，城头，上游，三明东，三明 |

**Figure 6.**
Stitching string of the actual wagon routing locus (From Pingdingshan East to Sanming)

*3.2.2 Spatial characteristic matrix and weight hash row vector.* In the process of traditional semantic lexical analysis, based on the frequency and property of the word, Simhash algorithm divides the fingerprint row vector into several parts and gives different weights. However, for the issue of the railway wagon routing *locus* recognition, freight cars do not repeatedly pass a certain railway station (circuitous phenomenon) in general. Thus, the traditional semantic lexical analysis method is meaningless to the issue of this paper. To solve this problem, the property of every railway station is used to estimate how to give the weight for the spatial eigenvector $V_a$, and then the 0-1 hash value $H_a$ is transformed into eigenvector weight hash value $W_a = \{W_i^a\}$, as shown in the constraints below.

In the formula (3), $v_i^a$ is the *i*th element of the *a*th eigenvector $V_a$. $I_a$ is the element quantity of the *a*th eigenvector that is divided by the separator. The form of the separator among the stations is free, which can be expressed as a semicolon (";"), comma (","), vertical line ("|") and other forms.

$$V_a = \{v_i^a | i \in I_a\} \tag{3}$$

Formula (4) states all the constants of the involved station *s* in the $OD_{\text{station}}$. $\lambda_s$ is the unique constant generated at random.

$$P = \{\lambda_s | s \in OD_{station}\} \tag{4}$$

$w_i^a$ is the weight value of the *i*th element in the *a*th eigenvector. $\alpha$, $\beta$, $\varphi$ and $\mu$ are four different weight constants when the passing station $v_i^a$ is the distributary station, non-distributary station, boundary station or marshalling station, respectively.

$$w_i^a = \begin{cases} \alpha\lambda_s \;\; v_i^a \in distributary\, station, s = v_i^a \\ \beta\lambda_s \;\; v_i^a \in non - distributary\, station, s = v_i^a \\ \varphi\lambda_s \;\; v_i^a \in boundary\, station, s = v_i^a \\ \mu\lambda_s \;\; v_i^a \in marshalling\, station, s = v_i^a \end{cases} \tag{5}$$

$w_a$ is the summation of the element weight value in the *a*th eigenvector, as shown in formula (6).

$$w_a = \sum_{i \in I_a} w_i^a \tag{6}$$

$W_j^a$ is the *j*th weight value of the *a*th eigenvector, which is calculated by constraint (7). When the *j*th bit of the *a*th *f-bit* 0-1 hash value $H_a$ is 1, $W_j^a$ will be $w_a$. Conversely, it will be $-w_a$.

$$W_j^a = \begin{cases} w_a & H_j^a = 1, j \in f \\ -w_a & H_j^a = 0, j \in f \end{cases} \tag{7}$$

All the constraints and formulas above transform the spatial characteristic matrix to the *f*-bit weight hash row vector.

*3.2.3 Weight hash combination and similarity calculation.* Formula (8) sums the value of every hash bit from $W_a$ to get the *Weight Fingerprint* (WF).

$$WF'_j = \sum_a W_j^a \quad j \in f \tag{8}$$

Aimed at reducing the dimension of the spatial eigenvector, the merged weighted result is mapped to the binary value, transformed to 0-1 string. The fingerprint set $F' = \left\{F'_j\right\}$ is calculated by formula (9).

$$F'_j = \begin{cases} 1 & WF'_j > 0, j \in f \\ 0 & WF'_j \leq 0, j \in f \end{cases} \tag{9}$$

At this point, the comparison of the spatial characteristics of the routing metadata is converted to the similarity comparison of the routing *locus* fingerprint, which optimizes the wagon routing *locus* mining method. The Hamming distance between two vectors measures similarity by comparing the number of different characters (Pi *et al.*, 2009; Jarrous and Pinkas, 2009). To estimate the similarity, hamming distance compares the fingerprint *d* with the fingerprint *e* to calculate the number of distinct characters (Jayram *et al.*, 2008), which is described by formula (10).

$$\delta(d, e) = \sum_{j \in f} \left| \overset{d}{F'_j} - \overset{e}{F'_j} \right| \tag{10}$$

*3.3 Routing locus pattern mining and recognition*

*3.3.1 Weight hash similarity adjacency matrix construction.* Comparing the similarity of any two fingerprint element $F_i$ and $F_j$, which are extracted from the routing fingerprint set *F*in the wagon routing metadata, similarity indicator $\pi(i, j)$is calculated via the bit length $f$ of the fingerprint *F* and hamming distance $\delta(i, j)$. Similarity indicator $\pi(i, j)$ constitutes the similarity matrix $\Pi = \{\pi(i, j)\}$ as shown in formula (11).

$$\pi(i, j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} = \frac{|f - \delta(i, j)|}{|f + \delta(i, j)|} \quad 1 \leq i, j \leq n \tag{11}$$

According to the similarity matrix, for wagon routing fingerprint set, clustering analysis is an effective method to collect the wagon routing *locus* pattern.

In recent years, clustering algorithm based on the similarity matrix was the focus of the study, which was divided into spectral analysis method and graph analysis method. However, the two main algorithms need to design the scale of the clustering. To avoid the problem above, *Squared Error Adjacency Matrix Clustering* (SEAM Algorithm) (Yu and Lin, 2008) was an excellent choice. The core idea of SEAM algorithm was that hypothesizing the similarity matrix reflected the structure of original data, the similarity matrix should be close to an adjacency matrix with the same similar weight measurement. In the ideal case, the two matrices were the same. SEAM algorithm did not need to specify other parameters, and the clustering result was depending on the given similarity matrix.

In principle, the consequence of clustering analysis will not be affected, when the similarity matrixmultiplies a similarity weight. Theoretically, there should be a certain adjacency matrix $E$ that is similar to the similarity matrix $\Pi$ in theory, and the ideal situation is that the two matrices are equal. In the paper, $uE = [ue(i, k)]_{n \times n}$ is defined as the similarity weighted adjacency matrix and offset weight. The optimum $E$ and $u$ can be calculated by least square method, and the loss function is shown in formula (12).

$$f(u,E) = \sum_{i \neq j} (\pi(i,j) - u \times e(i,k))^2 \tag{12}$$

The distributary phenomenon of railway wagon flow in the distributary station causes a higher coincidence degree of the wagon routing, generally more than 65 per cent, which would produce the situation of similarity over-iteration. Therefore, SEAM algorithm was adjusted by formula (13)-formula (16), and the adjusted algorithm was called as *Squared Error Offset Adjacency Matrix Clustering* (SEOAM Algorithm) in this paper.

The formula (13) transforms $f(u, E)$ into the relationship among the offset adjacency matrix $\Pi' = 1 - \Pi(i, j)$ that is shown in formula (16). The offset matrix $E'$ and the offset weight $u'$ are explained as below:

$$f(u',E') = \sum_{i \neq j} \left(\pi'(i,j) - u' \times e'(i,k)\right)^2 \tag{13}$$

The necessary condition of the loss function described by formula (13) is converted to formula (14) and (15).

$$e'(i,j) = \begin{cases} 1 & \pi'(i,j) \geq u'/2 \\ 0 & \pi'(i,j) < u'/2 \end{cases} \tag{14}$$

$$u' = \frac{\sum_{i \neq j} \pi'(i,j) \times e'(i,j)}{\sum_{i \neq j} e'(i,j)} \tag{15}$$

$$e^{(l)}(i,j) = 1 - e^{'(l)}(i,j) \tag{16}$$

For the adjacency matrix $E'^{(l)}$ of wagon routing fingerprint offset, SEOAM Algorithm is an efficient method to iterate the matrix, and the attribute $l$ presents the iteration count. Finally, the Tarjan algorithm is used to find the connected branches in the adjacency matrix, and then the clustering recognition results of the routing trajectory pattern are obtained. The calculation process was designed:

*INPUT*: $l$-iterative number, $e'^{(0)}(i, j)$-offset adjacent matrix, $u'(0)$ -offset weight, $\Psi$-iterative limit.

*OUTPUT*: similarity adjacency matrix $E^{(l)}$.

Step 1:Initialize$l = 0$, $e'^{(0)}(i, j) = 0$, $u'^{(0)} = \max_{i \neq j} \pi'(i, j)$, $\Psi$.

Step 2: Update offset adjacent matrix $E'^{(l+1)} = \{e'^{(l+1)}\}$ by formula (14) and formula (15).

Step 3: Calculate and update the offset weight $u'^{(l+1)}$.

Step 4: Terminate when $l + 1 = \Psi$ or $u'^{(l)} = u'^{(l+1)}$; otherwise, $l = l + 1$, and return to Step 2.

Step 5: Output the offset adjacency matrix $E'^{(l)}$ and the offset weight $u'^{(l)}$.

Step 6: Calculate the similarity adjacency matrix $E^{(l)}$ by formula (16).

*3.3.2 Routing locus clustering and pattern marking*. To get the clustering recognition result of the wagon routing *locus*, Tarjan algorithm (Liao, 2006) is asked to seek the connected branch in the adjacency matrix $E^{(l)}$. Tarjan algorithm is used to find the closed loop for the directed graph. However, the adjacency matrix was an undirected graph, and the features of the wagon routing fingerprint should be considered. In this paper, it is necessary to improve the Tarjan algorithm to fit the situation.

The process of the amended Tarjan algorithm was shown in Figure 7. The amended parts were in the shadows.

(1) There are two parts to be amended in the process of $P = \{i, \ldots, h, j\}$:

- Considering that the undirected graph has no direct follow-up vertex, the adjacent vertex $k$ of $j$ is used to extend $P$. When the vertices are selected from the adjacent vertex sets of $j$, they are accessed from small to large according to the label of the adjacent vertex. It can be ensured that all the vertices which are adjacent to $j$ can be accessed to only once.

(2) According tothe property of the undirected graph, if there is only $i$ in $P$, $k$ never appears in the loop which is started from $i$:

- When $P$ goes back to the vertex $v$ and the flag of $v$ is true, only all the adjacent vertices of $v$ are accessed, the tag of $v$ is released. However, in the traditional Tarjan algorithm, the tag of $v$ is immediately released when $P$ is accessing $v$.

Now, the clustering set C was found out by the amended Tarjan algorithm, which got the connected component of the adjacency matrix $E^{(l)}$.

## 4. Algorithm evaluation standard

The evaluationwas performed in each WRLCS set of the spatial characteristic. The manual annotation was used as the gold standard for the evaluation (Artiles *et al.*, 2009). In general, the recognition algorithm or system was evaluated using the standard clustering metrics *purity* and *inverse purity*, and the algorithm evaluation standard was called *P-IP* (Artiles *et al.*, 2007). This measure focuses on the frequency of the most common category in each cluster and rewards the clustering solutions that introduce less noise in each cluster (Khabsa *et al.*, 2015). *Purity P*, well known in *Information Retrieval*, is related to the precision measurement of the wagon routing *locus* pattern. The emphasis of *Inverse Purity IP* is the cluster with a maximum recall for each wagon routing pattern, which focuses on the comprehensiveness of recognition.

Being $C$ the set of wagon routing loci cluster $C_i$ recognized automatically to be evaluated and $L$ the set of wagon routing pattern category $L_i$ annotated manually, purity is computed

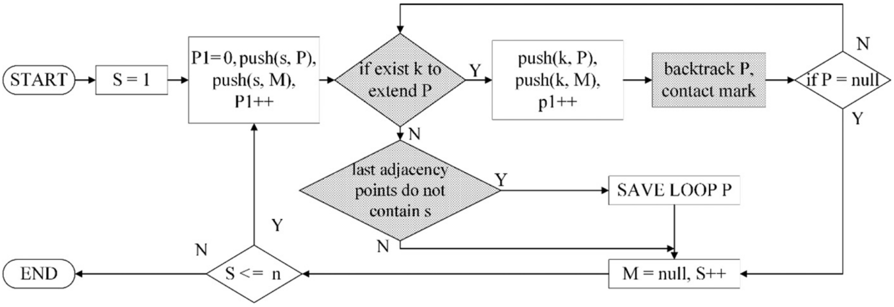


**Figure 7.**
The flow chart of the amended Tarjan algorithm

by taking the weighted average of maximal intersection between $C_i$ and $L_j$ for a given category $L_j$, as shown in formula (17).

$$P = \frac{\sum_{C_i \in C} \max_{L_j \in L} \|C_i \cap L_j\|_1}{\sum_{C_i \in C} \|C_i\|_1} \tag{17}$$

Inverse purity is calculated by getting the weighted average of maximal intersection between $L_i$ and $C_j$ for a given category $C_j$, defined as formula (18).

$$IP = \frac{\sum_{L_i \in L} \max_{C_j \in C} \|L_i \cap C_j\|_1}{\sum_{L_i \in L} \|L_i\|_1} \tag{18}$$

The relationship between the two indexes is not inevitable, but mutual restraint. Thus, we used the harmonic constant $F_\sigma$ of purity and inverse purity to rank the recognition algorithm finally. The $F$ measurement is defined by the following formula (19).

$$F_\sigma = \frac{(\sigma^2 + 1)P \cdot IP}{\sigma^2 P + IP} \quad \sigma \in [0, 1] \tag{19}$$

$\sigma$ is included as an additional measure giving more importance to the aspect of purity or inverse purity, which is a ratio constant in nature. Therefore, achieving a high inverse purity should be rewarded more than having high purity.

## 5. Case study

### *5.1 Experiment data and test environment*

The paper adopted the national railway historical report of wagon flow in the latter half of 2016 as the experimental data. The scale of the railway report of wagon flow is nearly 600 GB, generating about 130,000 OD. After processed by the routing metadata model, every OD routing metadata contains the spatial characteristics, which are consisted of 4,000-20,000 passing station records. Based on the actual production data, the experimental environment was established. The recognition algorithm was running on the computer with 2.5 GHz CPU, 3 GB random memory, developed by Java, and executed in the JVM environment.

### *5.2 Experiment result and analysis*

*5.2.1 Experiment procedure.* Using the routing metadata model, about 130,000 routing metadata were constructed. Taking the wagon routing from *Pingdingshan East Railway Station* to *Sanming Railway Station* as the example, six wagon routings with significant differences were found out from the records of 16,729 railway wagons by the traditional method. The traditional method refers to the direct comparison of the stitching strings of the actual wagon routing with the same OD. Because the stitching string of the passing station code is incomplete, the accuracy of comparison is low.

Table I shows that the situation and proportion of the seven loci are different from each other, but compared with the result annotated manually, in fact, there are only three actual wagon routing loci, as shown in Figure 1. In Table I, the percentage of *locus* 1 and *locus* 3 accounts for 24.694 per cent and 37.037 per cent, and the quantity of the wagon routing is 10,327. The quantity of passing station of *locus* 4 is the same with *locus* 5, while the WRLCS Unicode length of them are different, which means the traditional method treats them as two distinct routing loci.

First of all, the spatial characteristic $S_{L–Z}$ of routing metadata $RM_{L–Z}$ was divided by the distributary station, and the dimension of $S_{L–Z}$ was reduced to the form of 0-1. Second, while calculating the weight hash value of spatial eigenvector, the weight value of distributary station $\alpha$was set as 2, the non-distributary station $\beta$ was set as 1, boundary station $\varphi$ was set as 4 and marshalling station $\mu$ was set as 3. Then, the wagon routing spatial characteristic fingerprint was calculated by the improved Simhash algorithm, to compute the hamming distance of any two routings, whereby hamming distance matrix was obtained, as shown in Figure 8.

From Figure 8(a), we could not find the same *locus*, and all the routing fingerprint hamming distances were different. Computing the similarity matrix $\Pi_{L–Z}$ of the wagon routing *locus* pattern continuously, seen in Figure 8(b), $\Pi_{L–Z}$ could not support to cluster the actual routing *locus*. The routing fingerprint offset matrix $\prod'_{L-Z}$ in Figure 8(c) was calculated by formula (11). Assumed $\Psi$ was 10, the routing fingerprint offset adjacency matrix was set by SEOAM algorithm, as shown in Figure 8(d). $E_{L–Z}$ in Figure 8(e) was calculated by formula (16). Finally, combined with Tarjan algorithm, the clustering result of routing fingerprint undirected graph was solved, which was illustrated in Figure 8(*f*). Figure 1 indicated that the experimental result was consistent with the manual recognition result, and the SST algorithm found the actual wagon routing *locus*.

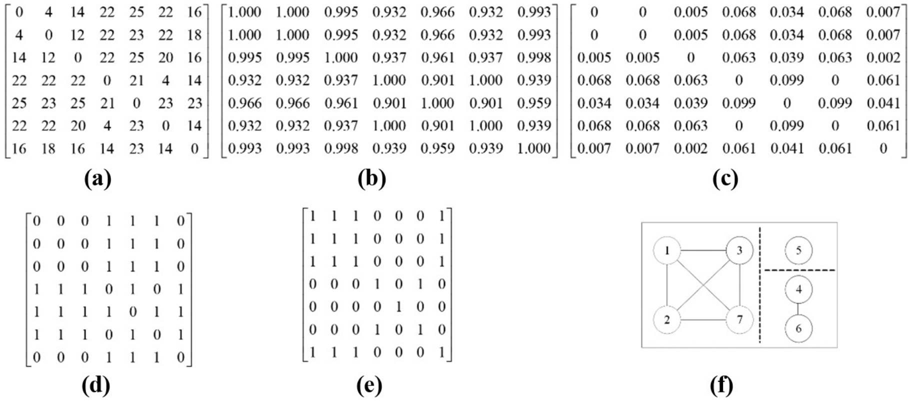
*5.2.2 Pattern recognition assessment.* Considering the scale of 130,000 routing metadata, it is hard to recognize the wagon routing pattern manually. Hence, the first 100 routing metadata is selected, which have rich and detail spatial characteristics, and P-IP evaluation standard is used to analyze the experiment result. In the harmonic coefficient $F_\sigma$, $\sigma$is 0.5, that means accuracy and comprehensiveness are equivalent important. The purity $P$, inverse purity $IP$ and $F_\sigma$ are shown in Table II.

The $F$ values of the three methods are over 80 per cent. The harmonic coefficient $F$ of the wagon routing recognition method in this paper is the best, as high as 94.739 per cent, which is 14.550 per cent higher than the traditional algorithm and 4.264 per cent higher than the semantic method. The experimental result demonstrated that the algorithm in this paper could effectively recognize the wagon routing *locus*.

*5.2.3 Calculation efficiency analysis.* The program was programmed by Java on the eclipse and run on the JDK 1.8. The OD set size was taken from 10 to 500. Respectively using different recognition algorithm to analyze the wagon routing metadata. Figure 9 showed the

**Table I.** Pingdingshan East-Sanming actual wagon routing locus analysis result by traditional method

| Wagon routing name | Passing station quantity | WRLCS Unicode length | Wagon routing statistics | Wagon routing ratio (%) |
|---|---|---|---|---|
| Locus 1 | 97 | 870 | 4131 | 24.694 |
| Locus 2 | 101 | 902 | 1652 | 9.875 |
| Locus 3 | 63 | 574 | 6196 | 37.037 |
| Locus 4 | 31 | 288 | 516 | 3.084 |
| Locus 5 | 31 | 300 | 826 | 4.938 |
| Locus 6 | 32 | 296 | 310 | 1.853 |
| Locus 7 | 39 | 373 | 3098 | 18.519 |


$$\begin{bmatrix} 0 & 4 & 14 & 22 & 25 & 22 & 16 \\ 4 & 0 & 12 & 22 & 23 & 22 & 18 \\ 14 & 12 & 0 & 22 & 25 & 20 & 16 \\ 22 & 22 & 22 & 0 & 21 & 4 & 14 \\ 25 & 23 & 25 & 21 & 0 & 23 & 23 \\ 22 & 22 & 20 & 4 & 23 & 0 & 14 \\ 16 & 18 & 16 & 14 & 23 & 14 & 0 \end{bmatrix}$$

(a)

$$\begin{bmatrix} 1.000 & 1.000 & 0.995 & 0.932 & 0.966 & 0.932 & 0.993 \\ 1.000 & 1.000 & 0.995 & 0.932 & 0.966 & 0.932 & 0.993 \\ 0.995 & 0.995 & 1.000 & 0.937 & 0.961 & 0.937 & 0.998 \\ 0.932 & 0.932 & 0.937 & 1.000 & 0.901 & 1.000 & 0.939 \\ 0.966 & 0.966 & 0.961 & 0.901 & 1.000 & 0.901 & 0.959 \\ 0.932 & 0.932 & 0.937 & 1.000 & 0.901 & 1.000 & 0.939 \\ 0.993 & 0.993 & 0.998 & 0.939 & 0.959 & 0.939 & 1.000 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 0 & 0 & 0.005 & 0.068 & 0.034 & 0.068 & 0.007 \\ 0 & 0 & 0.005 & 0.068 & 0.034 & 0.068 & 0.007 \\ 0.005 & 0.005 & 0 & 0.063 & 0.039 & 0.063 & 0.002 \\ 0.068 & 0.068 & 0.063 & 0 & 0.099 & 0 & 0.061 \\ 0.034 & 0.034 & 0.039 & 0.099 & 0 & 0.099 & 0.041 \\ 0.068 & 0.068 & 0.063 & 0 & 0.099 & 0 & 0.061 \\ 0.007 & 0.007 & 0.002 & 0.061 & 0.041 & 0.061 & 0 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

(d)

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(e)

(f)

**Notes:** (a) Routing fingerprint hamming distance matrix; (b) routing fingerprint similarity matrix; (c) routing fingerprint offset matrix; (d) routing fingerprint offset adjacency matrix; (e) routing fingerprint similarity adjacency matrix; (f) routing fingerprint udirected graph clustering

**Figure 8.** Thewagon routing locus fingerprint matrix and clustering partition

**Table II.** The analysis of comprehensive performance

| Method name | Purity $P$ (%) | Inverse purity $IP$ (%) | Harmonic coefficient $F_{\sigma=0.5}$ (%) |
|---|---|---|---|
| Traditional routing recognition method | 79.931 | 81.237 | 80.189 |
| Semantic recognition method (Artiles *et al.*, 2007) | 89.916 | 92.783 | 90.475 |
| Routing recognition method in this paper | 94.578 | 95.386 | 94.739 |

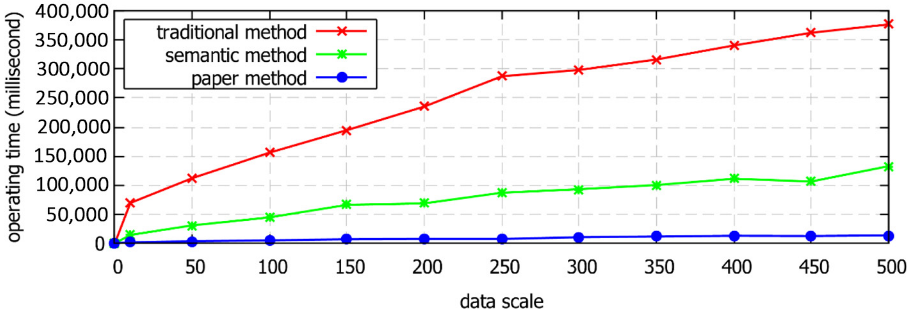


**Figure 9.** The comparison of the algorithm execution efficiency

comparison result of execution efficiency by traditional routing recognition method, semantic recognition method and wagon routing recognition method in this paper.

With the expansion of routing metadata scale, the time consumption of the traditional method is obvious, and the operating time growth rate is high. When the size of the routing metadata reaches 500, the operation time of the traditional method is close to 7 min. The semantic method improves the operating effectiveness, but the recognition effect remains to be improved, as shown in Table II. The algorithm of this paper has significant improvement not only on the recognition accuracy but also on the operating efficiency.

In the actual production process, for large-scale data sets, the time cost of traditional algorithms and literature algorithms is difficult to accept, and the algorithm operation time growth rate is low, it can maintain an acceptable stable state.

Using the algorithm mentioned above to identify parts of actual railway wagon routing loci, Figure 10(a) shows three actual wagon routing loci from Lianyun Railway Station (UHK) of Shanghai Railway Administration of China to Zhongning Railway Station (VNJ) of Lanzhou Railway Administration of China, Figure 10(b) shows eight actual wagon routing loci from Jingtanggang Railway Station (JGV) of Taiyuan Railway Administration of China to Yian Railway Station (YAV) of Taiyuan Railway Administration of China, Figure 10(c) shows two actual wagon routing loci from Qingdao Railway Station (QDK) of Jinan Railway Administration of China to Wangjiayingxi Railway Station (KNM) of Kunming Railway Administration of China, Figure 10(d) shows three actual wagon routing loci from Sanming Railway Station (SMS) of Nanchang Railway Administration of China to Jishan Railway Station (JSQ) of Guangzhou Railway Corporation of China, Figure 10(e) shows two actual wagon routing loci from Shuangyashan Railway Station (SSB) of Ha'erbin Railway Administration of China to Bohai Railway Station (BED) of Shenyang Railway Administration of China, and Figure 10(f) shows one actual wagon routing *locus* from Hami Railway Station (HMR) of Urumqi Railway Administration of China to Zhicheng Railway Station (ZCN) of Wuhan Railway Administration of China. All the results above indicate the proper recognition effect.

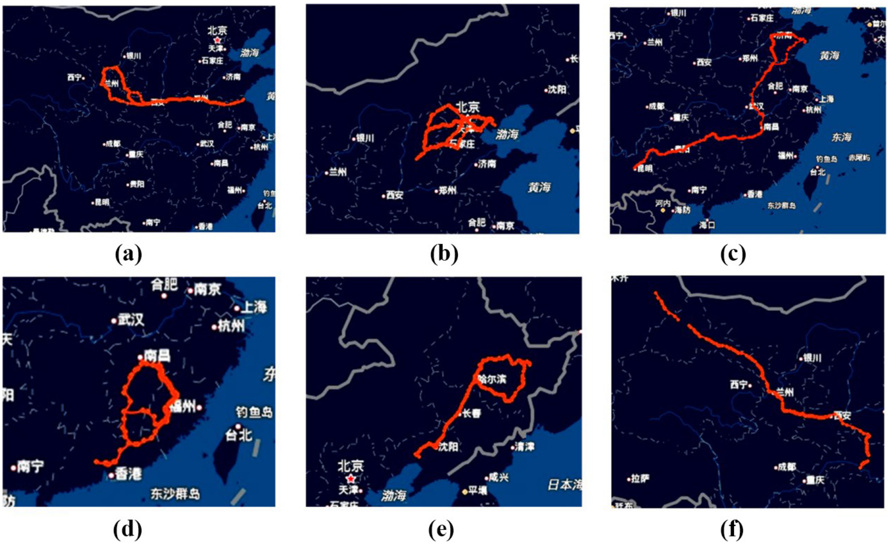


**Figure 10.**
The railway wagon flow actual routing locus recognition effect

**Notes:** (a) UHK-VNJ: 3; (b) JGV-YAV: 8; (c) QDK-KNM: 2; (d) SMS-JSQ: 3; (e) SSB-BED: 2; (f) HMR-ZCN: 1

## 6. Conclusions

In this paper, an intelligent algorithm of railway wagon routing is presented to accurately recognize the actual wagon flow loci from the railway big data. Four main characteristics of the railway wagon flow data are considered in the routing metadata model to ensure the comprehensiveness of information. Taking full advantage of the historical data and real-time data, the spatial characteristic is extracted from the fruitful context characteristic. Based on improved Simhash algorithm, SEOAM algorithm and Tarjan algorithm, SST wagon routing pattern recognition algorithm is designed properly and then successfully tested through a real-world case study. The results argued that the routing metadata model is effective to improve the efficiency of characteristic extraction. The results also point out that the SST-based algorithm is better than the traditional method not only in the efficiency but also in the accuracy.

The intelligent recognition for the wagon routing *locus* is one of the key parts in the railway wagon flow space forecast. However, different routing loci patterns provide various choices for the railway wagon. Hence, future research should attempt to explore the wagon routing decision pattern to guide the actual railway freight transport organization and improve the efficiency of the railway wagon coordination.

## References

Artiles, J., Gonzalo, J. and Sekine, S. (2007), "The SemEval-2007 WePS evaluation: establishing a benchmark for the web people search task", *International Workshop on Semantic Evaluations*. Association for Computational Linguistic, pp. 64-69.

Artiles, J., Gonzalo, J. and Sekine, S. (2009), "Weps 2 evaluation campaign: overview of the web people search clustering task", *Proceedings of the www Web People Search Evaluation Workshop*.

Bai, R., Wallace, S.W., Li, J. and Chong, A.Y.L. (2014), "Stochastic service network design with rerouting", *Transportation Research Part B: Methodological*, Vol. 60, pp. 50-65.

Caprara, A., Malaguti, E. and Toth, P. (2011), "A freight service design problem for a railway corridor", *Transportation Science*, Vol. 45 No. 2, pp. 147-162.

Cebon, P. and Samson, D. (2012), "Using real time information for transport effectiveness in cities", *City Culture and Society*, Vol. 2 No. 4, pp. 201-210.

Chang, J.S. (2010), "Assessing travel time reliability in transport appraisal", *Journal of Transport Geography*, Vol. 18 No. 3, pp. 419-425.

Chen, C., Chen, L., Xiong, J. and Yu, H. (2016), "Research and improvement of data de-duplication based on Simhash algorithm", *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, Vol. 3, pp. 85-91.

Cho, J.H., Kim, H.S. and Choi, H.R. (2012), "An intermodal transport network planning algorithm using dynamic programming-a case study: from Busan to Rotterdam in intermodal freight routing", *Applied Intelligence*, Vol. 36 No. 3, pp. 529-541.

Drličiak, M. and Čelko, J. (2016), "Implementation of transport data in to the transport forecasting in Slovakia ☆", *Transportation Research Procedia*, Vol. 14 No. 5, pp. 1733-1742.

Ford, W., William, F. and Topp, W. (2001), *Data Structures with C++*, TSINGHUA UNIVERSITY PRESS, Beijing.

Jarrous, A. and Pinkas, B. (2009), "Secure hamming distance based computation and its applications", *International Conference on Applied Cryptography and Network Security*, Vol. 40, pp. 107-124. Springer-Verlag.

Jayram, T.S., Kumar, R. and Sivakumar, D. (2008), *The One-Way Communication Complexity of Hamming Distance*, Vol. 4 No. 1, pp. 129-135.

Jeong, Y.S., Byon, Y.J., Castro-Neto, M.M. and Easa, S.M. (2013), "Supervised weighting-online learning algorithm for short-term traffic flow prediction", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14 No. 4, pp. 1700-1707.

Ji, L., Lin, B., Qiao, G. and Wang, J. (2011), "Car flow assignment and routing optimization model of railway network based on multi-commodity flow model", *China Railway Science*, Vol. 32 No. 3, pp. 107-110.

Jiang, N., Xia-Miao, L.I., Zhu, Y.H. and Wei, C.Y. (2004), "Mathematical problems in car flow routing", *China Railway Science*, Vol. 25 No. 5, pp. 121-124.

Jin, L., Ye, Y., Zhao, Y., *et al.* (2005), "Choosing of regional railway network car flow routing", *Chinese Railways*, Vol. 12, pp. 49-51.

Khabsa, M., Treeratpituk, P. and Giles, C.L. (2015), "Online person name disambiguation with constraints", *Acm/ieee-Cs Joint Conference on Digital Libraries*, ACM, pp. 37-46.

Lei, K., Zhu, X., Hou, J. and Huang, W. (2014), "Decision of multimodal transportation scheme based on swarm intelligence", *Mathematical Problems in Engineering*, Vol. 2014, pp. 1-10.

Li, Y., Zhao, J., Wu, G. and Chen, J. (2012), "Solving the mode selection problem with fixed transportation cost in intermodal transportation", *J. Southwest Jiaotong Univ*, Vol. 47, pp. 881-887.

Liao, J. (2006), "The recognition arithmetic study of special structures for palm diagnosis", Dissertation, Harbin Institute of Technology.

Lint, J.W.C.V. (2008), "Online learning solutions for freeway travel time prediction", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 9 No. 1, pp. 38-47.

Liu, J., He, S.W., Song, R. and Li, H.D. (2011), "Study on optimization of dynamic paths of intermodal transportation network based on alternative set of transport modes", *Journal of the China Railway Society*, Vol. 33, pp. 1-6.

Necula, E. (2015), "Analyzing traffic patterns on street segments based on GPS data using R", *Transportation Research Procedia*, Vol. 10, pp. 276-285.

Pi, B., Fu, S., Wang, W., Han, S. and Inc, R. (2009), "Simhash-based effective and efficient detecting of near-duplicate short messages", *Proceedings of International Symposium on Computerence and Computational Technology*, Vol. 4, pp. 20-25.

Shen, Z. and Deng, X. (2003), *Transportation Engineering*, CHINA COMMUNICATIONS PRESS, Beijing.

Shi, Q. and Yong, S. (1999), "Multi objective linear programming model and its algorithm for car flow routing with bidirectional heavy and empty cars in railway network", *Journal of the China Railway Society*.

Shi, Y. and Shi, Q. (1998), "Multi-Objective decision model for car flow routing with bidirectional heavy and empty Cars in railway network", *Shanghai Tiedao Daxue Xuebao*, pp. 78-82. Z2.

Wang, L., Ma, J., Lin, B., Chen, L. and Wen, X. (2014), "Optimal route choice model for loaded and empty car flows in railway network", *Journal of Beijing Jiaotong University*, Vol. 38 No. 6, pp. 12-18.

Wang, Q.B., Han, Z.X., Ji, M.J. and Li, Y.M. (2011), "Path optimization of container multimodal transportation based on node operation randomness", *J. Transp. Syst. Eng. Inf. Technol*, Vol. 11, pp. 137-144.

Wen, X., Lin, B. and Chen, L. (2016), "Optimization model of railway vehicle flow routing based on tree form", *Journal of the China Railway Society*, Vol. 38 No. 4, pp. 1-6.

Wu, W., Dong, B. and Chen, G. (2016), "Optimization of car flow routing based on vague sets", *Railway Transport and Economy*, Vol. 10, pp. 42-47.

Xiong, G.W. and Wang, Y. (2014), "Best routes selection in multimodal networks using multi-objective genetic algorithm", *Journal of Combinatorial Optimization*, Vol. 28 No. 3, pp. 655-673.

Xu, L. (2001), *Modern Mathematics Handbook*, Huazhong University of Science and Technology Press, Wuhan.

Yu, J. and Lin, Z. (2008), "Squared error adjacency matrix clustering", Technical Report on Dept. of Computer Science, Beijing Jiaotong University.

Zhou, Z., Lu, X., Peng, W. and Zeng, W, Transportation, S. O. and University, S (2015), "Vehicle shadow detection algorithm based on superpixel and SVM", *Journal of Southeast University*, Vol. 45 No. 3, pp. 443-447.

**Corresponding author**
Xiaodong Zhang can be contacted at: xdcheung@126.com

For instructions on how to order reprints of this article, please visit our website:
**www.emeraldgrouppublishing.com/licensing/reprints.htm**
Or contact us for further details: **permissions@emeraldinsight.com**