An approach to merging railway ticket records into origin-destination pairs based on sparse matrix characteristic

LIN Ruixi¹, LIN Boliang²

(1. Department of Electronics and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, 999077;

- School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China)
 Abstract: This work describes a data merging method under the background of railway tickets data integration. Origin-destination (OD) pairs are fundamental data in transportation engineering, and are widely used in various applications. For example, predictive OD pairs of the long-term future are the basis of optimizing transportation network design. Optimization of transportation plans also relies on existing OD pairs. The size of an OD matrix is always huge in a real transportation system. Each
- 15 element of the matrix, i.e. an OD pair, is accumulated by numerous passenger tickets or freight invoices. The size of an OD matrix is huge in real transportation systems. In a railway system, each element of the matrix, i.e. an OD pair, is accumulated by numerous passenger ticket or freight invoices. In 2013, China railway system transported 2.075 billion passengers. In other words, an equivalent number of tickets were sold. Each ticket consists of information on the name of origin and destination, train
- 20 number, and seat class etc. Similarly, a railway freight invoice consists of information on the name of origin and destination, freight category (e.g. coal, oil, grain, and ore), and volume etc. Obviously, the tickets usually contain some identical fields of information. For example, two tickets can share the same origin and destination. Researchers use only particular fields of the tickets for different analysis purposes. Consequently, an effective approach is needed to merge the ticket records based on interested
- keywords and creates the corresponding compressed OD matrix. A simple direct merging method costs a lot of computations. This work proposes an effective and efficient approach, named origin-associated merging, to merge ticket records based on the sparse matrix characteristic. For experiments, 30 samples are created according to the characteristic of China railway passenger flow, ranging from 200 thousands to 6 million records with a step of 200 thousands records. The experimental results show that the time using origin-associated merging is about 1% of that of direct merging. It is worth mentioning
- that the number of passenger tickets in China railway system is approximately 6 million per day. **Key words:** data integration; rail transportation; OD pair; sparse matrix; ticket records

0 Introduction

35

40

5

Data merging is an important method in data processing. With the progress of information technology, the scale of data has been rapidly expanded, especially in areas involved in computer applications. The work of Ying^[1] proposes a method to merge and integrate catalog data in merged university libraries. Chen^[2] gives a preliminary study on patterns of library network construction and data integration since university merging of China. In the paper of Zhang^[3], several issues are discussed including the integration of bibliographic data and merging of data to export, splitting bar code number, unifying the length of bar code number, deleting duplicate records and merging data. However, works on the merging of OD data in railway passenger tickets or freight invoices has not been found explicitly.

An optimization of passenger or freight train plan depends on the analyses of current or 45 annual prediction of OD data^{[4][5][6]}. While the optimization of railway network design usually depends on the prediction of OD data ten years later^{[7][8]}. At present, China railway operation mileage is over 100 thousands kilometers, among which the length of high-speed railway is over 11,000 kilometers, covering thousands of stations. There have been 4894 passenger trains

Brief author introduction:LIN Ruixi(1993-),Female,Undergraduate student,Information Iechnology,IC Design Correspondance author: LIN Boliang(1961-),Male,Professor,Railway Operation Management,Transportation Systems Network Design,Transportation and Logistics,Intelligent Transportation System. E-mail: bllin@bjtu.edu.cn

including 2660 high-speed trains(EMU) in China railway system since July 2014^[9]. According to

- 50 statistics, in the past year, 2.075 billion passengers travelled by China railway^[10]. It should be noted that 2.075 billion tickets does not impose the same amount of OD pairs. In general, an OD pair consists of a great number of passengers. To merge 2.075 billion tickets into the needed OD pairs, one ticket is compared to another by their origin station names and destination station names respectively. An estimated comparison times will exceed 100 trillion, which is hard to realize on
- 55 computers. For a large-scale railway network such as China railway network, only a few pairs of stations present passenger demands. In other words, for a given station, passengers travel to only a few stations. For instance, assuming 2000 stations in the railway network, at a given station *A*, the passengers can travel to all the other 1999 stations in theory. In reality, they actually travel to only a small number of stations among the 1999 stations. From the matrix point of view, the railway
- 60

65

70

OD matrix is quite sparse. Based on this sparse characteristic, this paper proposes an origin-associated OD records mergence approach.

The rest of the paper is organized as follows: Section 1.1 describes the direct merging method. In Section 1.2 we provide a complexity analysis of direct merging. In Section 2.1 we illustrate the origin-associated merging method, with a complexity analysis in section 2.2. Numerical analyses utilizing both methods are described in section 3.

1 Direct merging method

A typical train ticket contains information of the name of origin and destination, train number (which implicitly determines the type of train), class of service (hard seat, hard sleeper, or soft sleeper etc), and date etc. Assuming that there are N ticket records to be merged, the direct merging method is described as follows.

1.1 Description of the direct merging algorithm

Let S^{init} be a set of N ticket records. Without loss of generality, we assume that each record only consists of origin and destination stations. Let S^{merge} stores the OD pairs after merging the records. We assume that each OD pair is comprised of three fields: origin, destination, and number of tickets. Firstly, empty the set S^{merge} , move the first record r_1 of S^{init} into S^{merge} , name this record as OD pair h_1 and set the field of number of tickets as 1. Secondly, compare the second record r_2 of S^{init} with h_1 of S^{merge} . If the origin of r_2 is the same of that of h_1 and the destination of r_2 is the same of that of h_1 , merge r_2 into h_1 and increase the number of tickets of h_1 by 1. If either the origins or destinations are different, move r_2 into S^{merge} and set the number of tickets of h_2 as 1. Continue moving and comparing records. When it comes to the record r_k , assuming there are already m OD pairs in S^{merge} , compare r_k with the elements of S^{merge} successively. If the origin of r_k is the same of that of h_i in S^{merge} and the destination of r_k is the same of that of h_i , merge r_k into the number r_k is the same of that of h_i and increase 1 to the number

of tickets of h_i . If either the origins or destinations are different for all *m* OD pairs in S^{merge} , 85 add r_k to S^{merge} and note it as h_{m+1} , set its field of number of tickets as 1. Repeat the above operations until all records in S^{init} are compared. The detailed algorithm is illustrated in the flow chart (Fig. 1).

http://www.paper.edu.cn



Fig 1. The flow chart of direct merging algorithm

90 **1.2 Complexity analysis**

Assume there are M origin-destination pairs after merging N ticket records. Define the reduction ratio θ as $\theta = M/N$. Apparently, the record r_1 does not need to be compared, so the computation time is 0. The last record r_N will be compared M times in the worst case, while it needs to be compared only once in the least case. Therefore, as an approximate estimation, the average times of comparison a record is about (1+M)/4, and dealing with N records needs

95

中国对技论文在线

a total comparison time of N(1+M)/4, i.e, the complexity β_1 of the direct merging algorithm is

$$\beta_1 = N(1+M)/4$$
 (1)

We can infer from (1) that the computation time of this algorithm grows linearly with respect 100 to the number of records when M is a constant.

2 Origin-associated merging method

There will be $Y \times (Y-1)$ station pairs for a railway network with Y stations theoretically, but for a large-scale network, not all station pairs have traffic demands. For example, China railway network consists of 5,544 stations in 2007 with 520,000 freight OD pairs^[6]. Each station has only 93.8 OD pairs on average. In other words, for a certain freight origin, the number of corresponding destinations is less than two percent of all stations. It is obvious that the railway OD matrix is highly sparse. Utilizing this characteristic, we can first compare the origins of two records, if their origins are same, compare their associated destinations. In this way the number of comparison times can be effectively reduced. Based on this idea, we propose a novel merging approach named origin-associated merging method. The algorithm is described as follows.

Description of the origin-associated merging algorithm 2.1

Let F be a set of all origins. The size of the set is Y. Let D(i) be a set of all possible destinations of origin i. Let $n_{D(i)}^k$ be a variable of the number of tickets whose origin is the *i*th element in F and whose destination is the kth element in D(i). Firstly, empty the set F, copy the origin of the first record r_1 of S^{init} to F, copy its destination to D(1) and set $n_{D(1)}^1 = 1$. When it comes to record r_k in S^{init} , assuming there are already n_f origins in F, compare the origin of r_k with that of all n_f origins. If the origin of r_k matches the *i*th element of F, compare the destination of r_k with each elements of D(i) respectively. Stop comparing when the destination of r_k matches the *h*th element in D(i). Increase the value of $n^h_{D(i)}$ by 1. If no matching element is found, add the destination of r_k to D(i) and set 120 $n_{D(i)}^{|D(i)|+1} = 1$. However, in the case where the origin of r_k does not match any elements of F, add the origin of r_k to F and increase the value of n_f by 1. Meanwhile, add its destination to $D(n_{f})$. Repeat the above operations until all records of S^{init} are compared. Detailed algorithm is shown in the following flow chart.

105

110

115



Fig 2. The flow chart of origin-associated merging algorithm

2.2 Complexity analysis

130

125

Skip the first record r_1 . The origin of the last record r_N will compare Y times in the worst case, while it needs only one comparison in the least case. An average comparison time is the mid-value of the two extreme values, i.e. (1+Y)/2. Consequently, the average comparison time of the origin of all elements in S^{init} is (1+Y)/4. The origins of all N records needs a comparison time of N(1+Y)/4. Now assume each origin is associated with D destinations.

Given a certain origin $u \in F$, suppose there are N_u records with origin u, among which the destination of the first record need not be compared; for the last record, its destination needs to be compared D times in the worst case, but only once in the least case. In this analogy, for any record, the average comparison time should be the mid-value, that is (1+D)/2. Therefore, for all N_u records, the average comparison time is (1+D)/4. We can infer that the complexity β_2 of this method is

$$\beta_2 = N[(1+Y)/4 + (1+D)/4]$$
⁽²⁾

140 Clearly, the relative efficiency α of the two algorithms is

$$\alpha = \frac{\beta_2}{\beta_1} = \frac{N[(1+Y)/4 + (1+D)/4]}{N(1+M)/4} \approx \frac{Y+D}{M}$$
(3)

Suppose N = 6,000,000 (which is roughly equal to the number of tickets sold per day in China railway), M = 200,000, Y = 2000, and D = 100, then $\alpha \approx 1\%$. Apparently, origin-associated merging method improves the merging efficiency significantly.

145 **3 Numerical Analyses**

We observed from our work supported by China Railways Corporation that, in recent years, freight stations with lower freight demands have been closed and the number of freight stations in operation has reduced to about 3,239. In addition, the number of OD pairs is 336,384, which suggests that each station is associated with about 104 OD pairs. As for rail passenger transportation, the number of stations with an departure or arrival volume over 10,000 passengers per year is about 2,100. The number of the OD pairs of these stations is about 200,000. On average, each station is associated with about 95 OD pairs. In conclusion, the OD matrix of China railway network is highly sparse.

- For the convenience of analysis, we may neglect some stations whose transportation
 demands are low, so we can assume there are 2000 major passenger stations, labeled from A0001~A2000. The origins of the records are generated randomly from A0001to A2000. In theory, for a certain station, there should be 1999 OD pairs. Without loss of generality, We assume that each origin is associated with 100 destinations. We also assume the destinations are randomly picked from 100 stations, labeled by B0001~B0100. In fact, B0001~ B0100 should be parts of A0001~ A2000. We separate the notations of origins and destinations for the convenience of
- analysis, without loss of logical reasoning. According to the above hypothesis, we generate 30 samples. The numbers of records in these samples are 200,000, 400,000, 600,000, 800,000,, 6,000,000 respectively.

3.1 Analysis of records reduction ratio

165

150

We adopt direct merging and origin-associated merging approaches respectively to do computations on the samples described above. All experimental analyses are done on a personal computer with Intel(R) Core(TM) i5 processor, CPU frequency at 3.19GHz, and a memory of 4G bytes. The computational results are shown in table 1.

Tab. 1 Major computational results									
No.	Ν	M	θ	Time_1/(s)	Time_2/(s)	α			
1	200000	126242	0.63	98	1	1.02%			
2	400000	172681	0.43	296	4	1.35%			

3	600000	189789	0.32	524	6	1.15%
4	800000	196185	0.25	673	8	1.19%
5	1000000	198557	0.20	872	10	1.15%
6	1200000	199385	0.17	1054	11	1.04%
7	1400000	199697	0.14	1239	14	1.13%
8	1600000	199835	0.12	1449	15	1.04%
9	1800000	199871	0.11	1639	18	1.10%
10	2000000	199897	0.10	1843	20	1.09%
11	2200000	199912	0.09	2057	22	1.07%
12	2400000	199914	0.08	2303	24	1.04%
13	2600000	199915	0.08	2510	26	1.04%
14	2800000	199919	0.07	2671	29	1.09%
15	3000000	199914	0.07	2926	32	1.09%
16	3200000	199915	0.06	3082	35	1.14%
17	3400000	199910	0.06	3336	37	1.11%
18	3600000	199911	0.06	3540	39	1.10%
19	3800000	199912	0.05	3715	39	1.05%
20	4000000	199923	0.05	3864	42	1.09%
21	4200000	199913	0.05	3987	42	1.05%
22	4400000	199915	0.05	4222	44	1.04%
23	4600000	199922	0.04	4410	46	1.04%
24	4800000	199924	0.04	4577	47	1.03%
25	5000000	199929	0.04	4820	49	1.02%
26	5200000	199923	0.04	4977	51	1.02%
27	5400000	199925	0.04	5251	52	0.99%
28	5600000	199928	0.04	5622	55	0.98%
29	5800000	199921	0.03	5824	55	0.94%
30	6000000	199916	0.03	5906	59	1.00%

170

The forth column of table 1 is the reduction ratio, which is calculated by the value of the third column over that of the second column, i.e. $\theta = M/N$ Take the number of records as the horizontal axis, reduction ratio as the vertical axis, we can obtain the curve of reduction ratio shown in figure 3.

http://www.paper.edu.cn

中国科技论文在线



Fig. 3 Reduction ratio of railway passenger ticket records

Figure 3 indicates the reduction ratio θ becomes smaller and approaches 0 as the number of tickets grows. This feature will be more evident if we set the merged OD pairs as vertical axis shown in figure 4. It can be easily seen that as the number of records exceed 1 million, the corresponding value of the y-axis saturates at 200,000.



Fig. 4 Number of records before/after merging

3.2 Calculation time analysis using direct merging

185

175

180

We can draw the calculation time curve using the data in column 5 of table 1 (Fig. 5). The calculation time grows linearly as the number of records grows, which meets our theoretical formulation N(1+M)/4. Since the structure of passenger flow depends on the population distribution and passenger train plan in a region, M is approximately a constant within a given period.

山国對技论文在线



190

Fig. 5 Computation time of direct merging

Calculation time analysis using origin-associated merging 3.3

We can draw the calculation time curve using the data in column 6 of table 1 (Fig. 6). The calculation time also grows linearly as the number of records grows, which basically meets our theoretical formulation N[(1+Y)/4 + (1+D)/4]. It will take about 1 second to merge every

100,000 records. For a give railway network and a stable population distribution, Y and D are 195 roughly constants. Of course, Y and D will change slowly as the scale of railway network expands and the incomes of residents grow.



Fig. 6 Computation time of origin-associated merging

Because the computation speed of the origin-associated merging is nearly 100 times faster than direct merging, log scale representations can be more intuitive. Hence we plot a time curve with a log-scale y-axis in figure 7.

200



Fig. 7 Comparison between the computation times of two methods

In the figure above, the red curve represents the direct merging method and the blue one is the other method. It is not difficult to find that the vertical interval between the two curves is approximately constant. This further confirms the accuracy of (3).

4 Conclusion

205

Based on the sparse matrix characteristic, this paper proposes the merging methods to 210 generate railway OD pairs. The first method, namely the direct merging method, compares the records origin by origin, then destination by destination. Then we refined the direct merging method considering the sparse characteristic of the OD matrix, and proposed a second method, namely the origin-associated merging method. Origin-associated merging significantly enhances the merging performance. The computation speed rises roughly 100 times compared to direct 215 merging. The approach in this work issues from the demands of the optimization for railway transportation operation and management. Nevertheless, the underlying method can also be applied in different transportation systems, such as highway transportation systems, urban transport management and traffic control, and urban rail transit. We believe that this method can achieve a similar good performance as long as the corresponding OD matrix has a highly sparse 220 feature. Take the railway freight transportation as an example. A railway freight invoice record includes information of the name of origin and destination, freight category (e.g. coal, oil, grain, and ore), car type, volume, and cars needed etc. In railway network design problem, researchers only care about the volumes between two zones or stations. While for the design of a coal

225 freight categories in the records. Moreover, in the researches of railway empty car distribution, the interested fields of information are the origins, destinations and railcar types. In general, when dealing with the design of transportation network, traffic plan and etc., we frequently encounter the problems of merging huge quantities of passenger or freight transportation records according to different needs. The raw records often come in tens of thousands, and even in millions.

transportation rail line, besides the origin and destination, people also need the information of

230 Repeated mergence of data will consume a lot of time, so researchers are trying various techniques to find efficient merging methods. Approaches to merging records according to the interested fields have become an important research topic in large-scale transportation systems. This work

can be used as technical support for further research works.

Acknowledgements

235

260

265

The author gratefully acknowledges PhD students Xuhong WEN and Lei CHEN from Beijing Jiao Tong University for advice on the refinements and embellishments of figures in this work, and master student Jianping WU from Beijing Jiao Tong University for advice on the tests and computational analyses.

References

- [1] YING H. The mergence and integration of catalog data in merged university libraries[J]. Journal of academic libraries, 2004, 21(5): 63-65
 [2] CHEN M H. Preliminary Study on Patterns of Library Network Construction and Data Integration Since University Merging[J]. Journal of Library and Information Sciences in Agriculture, 2005, 17(3): 14-16
 [3] ZHANG W H. Technologies of the mergence and integration of catalog data in the university libraries[J].
 Journal of Library and Information Sciences in Agriculture, 2010 (4): 81-83
 [4] LIN B L, WANG Z M, JI L J, et al. Optimizing the freight train connection service network of a large-scale rail system[J]. Transportation Research Part B: Methodological, 2012, 46(5): 649-667
 [5] BARNHART C, JIN H, VANCE P H. Railroad blocking: a network design application[J]. Operations Research, 2000, 48 (4), 603-614.
- [6] MARTINELLI D R, TENG H. Optimization of railway operations using neural networks[J]. Transportation Research Part C, 1996 4 (1), 33-49.
 [7] LIN B L, XU Z Y, HUANG M, GUO P W. An optimization model to railroad network designing[J]. Journal of the China Railway Society, 2002, 24 (2), 1-6.
- [8] KUBY M, XU Z Y, XIE X D. Railway network design with multiple project stages and time sequencing[J].
 Journal of Geographical System, 2001, 3: 25-47.
 [9] News China. New train schedule will come into effect from July, 2014. [OL]. [2014].

http://news.china.com.cn/txt/2014-06/11/content_32632874.htm

[10] China Railway Corporation. Cumulative completion of main indicators of national railway from January to December in 2013[OL]. [2014]. http://www.china-railway.com.cn/gkl/tjxx/201403/t20140326_42558.htm

基于稀疏矩阵特点的铁路原始客货流归并 方法研究

林睿熙1,林柏梁2

(1. 香港科技大学,电子与计算机工程系,香港,999077;

2. 北京交通大学, 交通运输学院, 北京, 100044)

- 摘要:在铁路客货票数据大规模增长的背景下,本文提出了一种数据归并方法。OD (origin-destination pairs)数据是交通运输系统的基本数据,远期预测的 OD 数据是交通运输网络设计的基础,而当前的 OD 数据是设计各种优化运输方案的主要依据。由于在一个大规模的运输系统中,不仅仅 OD 矩阵非常大,而且每个 OD 对本身就是大量的原始客货票据
 270 累加而成。例如,2013年中国铁路就发送了 20.75 亿旅客,每张客票的数据包括了:发站、到站、车次、席位种类等信息。类似地,体现铁路货运 OD 的货票记录就包括了:发站、到站、货品名、使用货车种类、运输量等字段。海量的记录往往包含一些共同的字段信息,如两张不同的火车票可能有相同的发站、到站等。而在不同的应用研究中,往往关注的只是其中的部分字段。因此,如何针对有效字段进行铁路客货票据的合并,以形成所需要的 OD 数
- 275 据,就成为铁路运输系统后续优化模型计算的重要前期工作。由于直接合并铁路客货票记录将产生大量的计算时间,数据处理产生的时间成本很大。本文针对这一现象,提出了一种基于稀疏 OD 矩阵特征的关联归并法,有效地提高了原始 OD 记录的的合并效率。作为检验例子,本文根据中国铁路客流分布特征随机产生了一组数据,以 20 万条记录为步长,从 20 万到 600 万(相当于铁路一天的客票数量)共计 30 文件,采用直接归并法和关联归并法进
- 280 行比较,对于 600 万条客票记录的数据量,关联归并法时间消耗仅仅是直接归并法的百分之一,验证了本文所提方法的有效性和实用性。 关键词:数据归并;铁路;OD 对;稀疏矩阵;票据 中图分类号:TP274