Journal of Computers Vol. 30 No. 3, 2019, pp. 252-259
doi:10.3966/199115992019063003020

Data Analysis Ability of Data Mapping in Business Circulation Statistics

Yu Guan

College of Information Engineering, Fuyang Normal University, Fuyang, Anhui 236041, China
guanyyug@163.com

Received 17 April 2019; Revised 29 April 2019; Accepted 8 May 2019

Abstract. Business circulation statistics are in urgent need of efficient data analysis methods. In this study, the application of data mapping method in business circulation statistics was studied. Taking GD-PUMA as the data processing platform and R language as the tool, the data mapping processing rules and analysis rules were designed, and then the business circulation statistics of Anhui province were used as an example to analyze the data analysis capability of data mapping. It was found that the data could be processed well using the method proposed in this study, cross heuristic had the best performance; in the data analysis, the ARIMA model could predict the data more accurately, with an average error of prediction of about 5.17%, and it showed favourable reliability in the test of ability of processing online request. The results prove that data mapping has a good performance in analyzing business circulation statistics, which provides some theoretical support for the further application of data mapping.

Keywords: business circulation, data analysis, data mapping, data processing

1 Introduction

With the development of network technology, the accumulated data of all walks of life has shown an explosive growth trend. The emergence of the Internet and social media has brought huge data [1], and the difficulty of data analysis has also increased. It takes a long time and effort to complete data processing [2]. How to effectively process massive amounts of data, obtain valuable information from it, and provide reasonable decision support is a huge test. In the statistics of business circulation, data analysis is also a big problem. The distribution of business circulation data is scattered, and the quality of data varies, making it difficult to be dealt with. Data mapping is a good method of data analysis which plays an important tole in the data integration and exchange between institutions and organizations with different data standards [3] and has been extensively studied. Reza et al. [4] applied data mapping into the processing and monitoring of health information data and made an analysis on the opportunities and advantages of the method. Shang et al. [5] designed a diagonal data mapping scheme that could be flexibly extended to support DCT/IDCT with different data throughput rates and achieve 1080 pHD resolution in real-time processing of video. Diener et al. [6] mapped threads of sharing data to a kernel with shared cache and then mapped memory pages to the memory controller, which not only reduced the cost of access, but also improved the performance of the algorithm. Matta Gómez et al. [7] studied the data mapping of multi-robot systems through the Microsoft Robotics Developer Studio framework and analyzed the local maps and global maps of the robots. Cruz et al. [8] optimized system by data mapping to solve problems of system performance and energy efficiency and found that the system performance improved by 19.2% and the energy efficiency increased by 15.7%. Serpa et al. [9] studied the impact of data mapping on machine learning and found that the implementation time of Intel Xeon and Xeon Phi KNL decreased by 25.2% and 18.5% respectively after using the mapping strategy. The current studies show that data mapping has a good performance in extracting effective information and processing and analyzing mined data. There are some problems in business circulation statistics, such as scattered data, statistics difficulty and low data utilization rate. It is difficult to achieve effective analysis of business circulation statistics by manual operation. In order to improve the analysis ability of statistical data, this

study applied computer technology to business circulation statistics. Anhui business circulation statistics was taken as the research subject, and data mapping method was used for study. It was found that data mining method could effectively improve the analysis efficiency of business circulation statistics and it had practical value in the field of economic analysis.

2 Business Circulation Statistics and Data Mapping

Data mapping [10] refers to obtaining the corresponding relationship of data through analyzing data with statistical methods. Applying data mapping in the solution of practical problems not only needs computer technology, but also needs multidisciplinary knowledge. With the development of technology, the subjects of data mapping have also evolved from structural data to unstructured data in the form of pictures and videos. In the environment of big data, data mapping has evolved from database mapping to big data mapping and has a better performance in the analysis of big data. Commonly used data mappings include entity mapping, table mapping, and attribute mapping, where attribute mapping is the most basic and common one. The flow of data mapping is shown in Fig. 1.

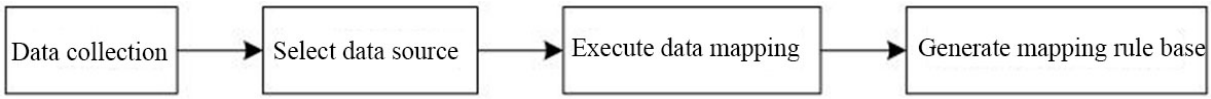

Fig. 1. Process of data mapping

Business circulation statistics is one of the work of the statistical department. The traditional statistical work mainly comes from the data of the management department. However, with the development of the network, the emergence of e-commerce and cross-border e-commerce has led to a significant increase in the data sources for statistical work, which brings a lot of difficulty to the statistical work. Anhui province in China has developed commerce. Applying data mapping technology in business circulation statistics is conducive to the good development of commerce and trade economy. In this study, GD-PUMA was used as the data processing platform, R language was used as the tool, and the data mapping method was used in the analysis of the business circulation statistics of Anhui province.

3 Data Mapping Rules

3.1 Data Preparation

Business circulation data of Anhui province were collected, and the data was mainly from the statistical information platform of the trade and circulation industry and the customs import and export data inquiry system.

3.2 Data Attribute Analysis

According to the statistics of business circulation of Anhui province, the data attributes were counted, as shown in Table 1.

Table 1. Data attributes

	Wholesale industry	Retail industry	Catering industry	Warehousing industry
Valid attribute	13	11	10	12

3.3 Data Preprocessing

The collected data were filtered and cleaned. Data whose operation could not be recorded or which had null value, error and special characters were evaluated as abnormal data, erroneous and abnormal data were deleted, data with null value were supplemented with 0, the special characters in enterprise names

were filtered, and the English bracket was replaced with Chinese bracket to reduce computation error and improve data utilization rate.

3.4 Data Processing Rules

Appearance frequency of the valid attributes of the data was counted, and the attributes with higher appearance frequency were selected, as shown in Table 2.

Table 2. Attributes with high frequency

Attributes	Frequency
Operating income	10
Total assets	5
Main business income	4
Operating profit	4
Business hours	3
E-commerce transaction amount	3
Total liabilities	3

Four attributes with the highest frequency of appearance, operating income, total assets, main business income, and operating profit, were take as key attributes.

The data in GD-PUMA were analyzed using R statistical tool. Attributes of about 20,000 items of data were analyzed. The data attributes were clustered using clustering algorithm in R language [11]. It was found that the key attributes had obvious aggregation phenomenon, indicating that the key attributes selected were reliable and could effectively filter the data.

The validity of the mapping rule was verified. Taking the customs import data as an example, the collected data was first classified according to the data types shown in Table 1. The distribution of the different types of data is shown in Table 3.

Table 3. Data distribution

Types	Percentage /%
Wholesale industry	72.91
Retail industry	84.63
Catering industry	0.21
Warehousing industry	0.06

Then they were matched according to key attributes, and the matching correct rate is shown in Table 4.

Table 4. Customs export data matching correct rate

Types	Aims	Correct rate/%
Wholesale industry	Customs import	82.44
Retail industry	Customs import	3.82
Catering industry	Customs import	42.64
Warehousing industry	Customs import	94.78

It was found from Table 4 that the warehousing industry had the highest matching rate for customs import data, followed by the wholesale industry. Good matching results were obtained according to the key attributes selected in this study, which showed that the method was effective.

3.5 Data Analysis Rules

Data analysis is based on data prediction. The data prediction methods include ARIMA model and Holt-Winters model, and the two models are introduced in the following.

ARIMA model [12] is established based on the sequence after difference. For ARIMA (p, d, q) model, it is assumed that the import and export historical data was Y_t. d-order differential processing was performed to obtain sequence X_t. Then ARIMA (p, q) model was fitted and processed by original d-

order differential reduction to obtain the prediction result.

The Holt-Winters model [13] is a statistical-based predictive model, and it can be expressed as:

$$\begin{cases} W_t = \alpha\left(X_t - S_{t-L}\right) + \left(1-\alpha\right)\left(W_{t-1} - q_{t-1}\right) \\ S_t = \beta\left(X_t - W_t\right) + \left(1-\beta\right)S_{t-L} \\ q_t = \gamma\left(W_t - W_{t-1}\right) + \left(1-\gamma\right)q_{t-1} \\ F_{t+m} = W_t + mq_t + I_{t-L+m} \end{cases} \tag{1}$$

where $W_t = \sum_{t=1}^{L} \frac{X_t}{L}$, $S_t = X_t - W_t$, and $q_t = 0, t = 1,2,\cdots,L$,

where t stands for time, X_t stands for observed value at t, W_t stands for stable component, S_t stands for seasonal component, q_t stands for trend component, m stands for the number of prediction periods, F_{t+m} stands for the predicted value of m, L stands for the length of season, and α, β and γ stand for smoothing parameters, $\alpha, \beta, \gamma \in [0,1]$..

The operating income of the wholesale industry was predicted, and ARIMA model was used as the prediction tool and the operating income data from January 2015 to September 2018 were used as the sample data. The process of data mapping includes generating time series diagram, differentiate operation, white noise detection, model order determination, validity test and prediction application.

The final comparison between the predicted and actual values is shown in Fig. 2.

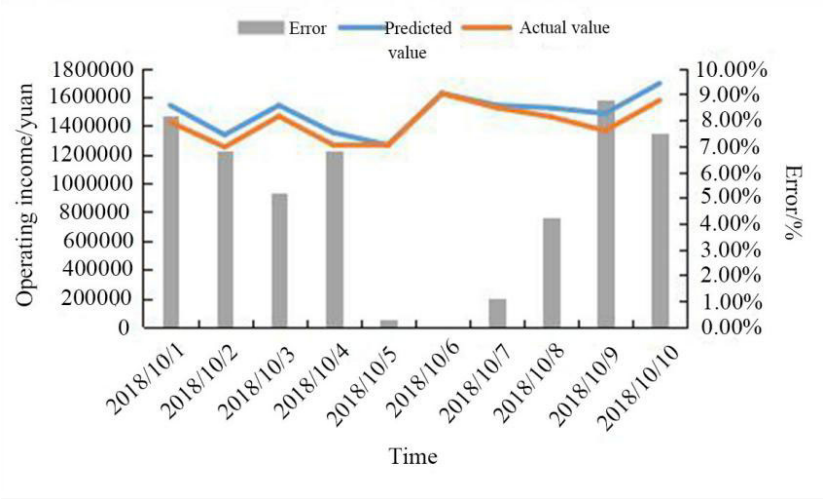

Fig. 2. Comparison of predicted and actual values

It was found from Fig. 2 that the result of prediction based on the ARIMA model was not much different from the actual value, and the average error was about 4.91%, which suggested that the ARIMA model was effective.

4 Performance Test of Data Mapping Rule

To further understand the performance of the data mapping method, both data processing rules and data analysis rules were tested.

Firstly, the data processing rules were tested. The time and space costs of the sequential, parallel and crossover algorithms were compared. All the data in the database and 5000 data which were randomly selected were used as test subjects. The results are shown in Fig. 3 and Fig. 4.

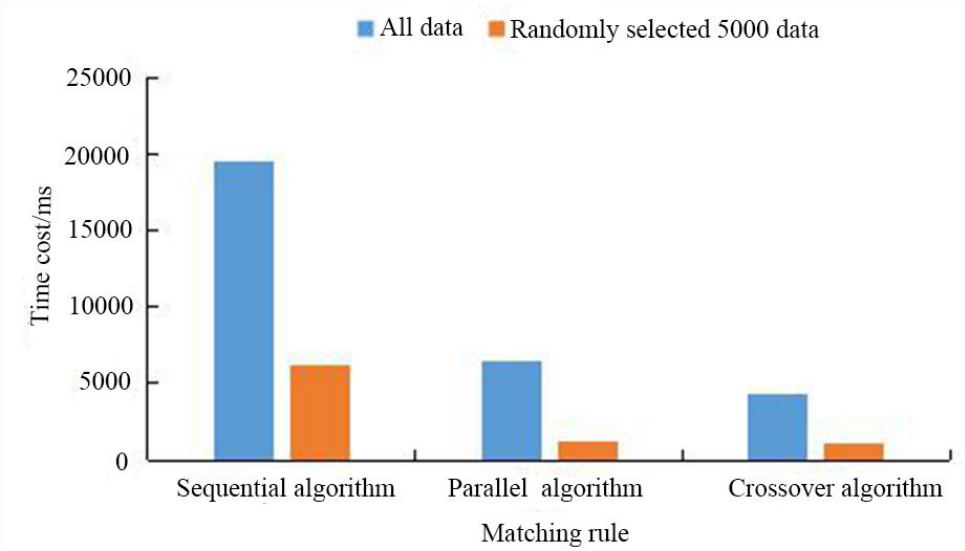

Fig. 3. Comparison of time cost

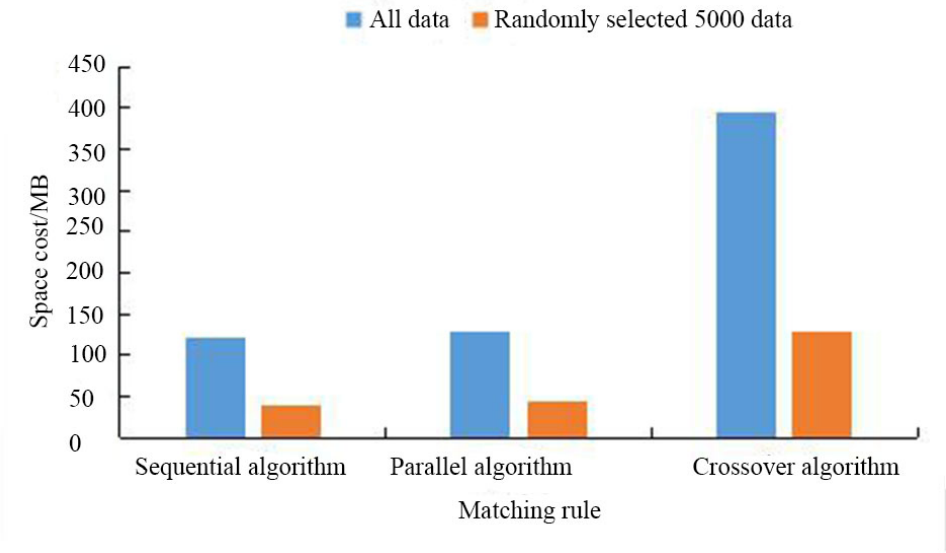

Fig. 4. Comparison of space cost

It was found from Fig. 3 and 4 that the sequential algorithm was the slowest, but the storage space required by the sequential algorithm was the smallest, and the crossover algorithm ran faster, but the storage space required by the crossover algorithm was large. The larger the amount of data, the more time and space cost. In the actual situation, data processing requires fast processing speed. Therefore, the crossover algorithm should be selected in the processing of big data.

In the test of data analysis rules, ARIMA model and Holt-Winters algorithm in R language were compared. The operating income of the wholesale industry was taken as an example, and enterprise data were selected for testing. The prediction results are shown in Fig. 5.

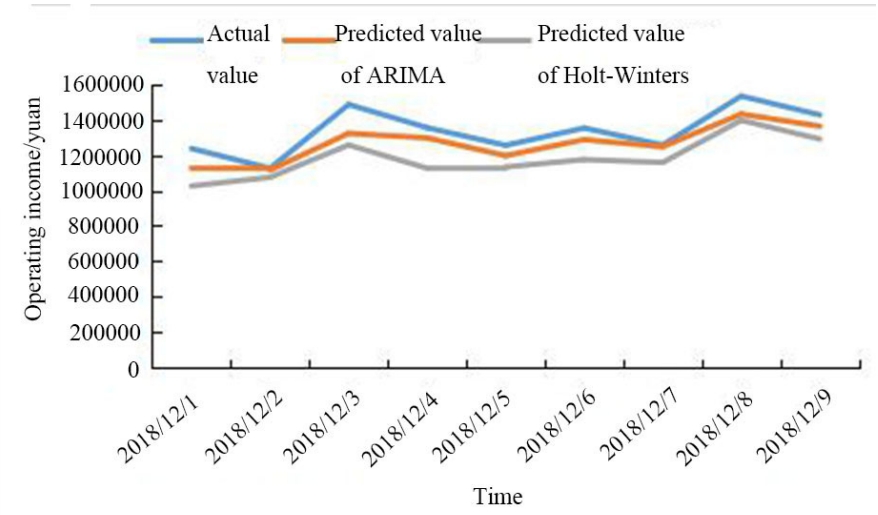

Fig. 5. Comparison of predicted value

It was found from Fig. 5 and Table 5 that the predicted value of the ARIMA model was closer to the predicted value of the Holt-Winters model, and the prediction error was also small. The average prediction error of the ARIMA model was about 5.17%, while that of the Holt-Winters model was about 11.46 %, much higher than the ARIMA model. It indicated that the data mapping method was reliable.

Table 5. Error comparison

	Error of ARIMA	Error of Holt-Winters
2018/12/1	8.96%	17.07%
2018/12/2	0.90%	4.51%
2018/12/3	10.96%	15.39%
2018/12/4	4.27%	16.91%
2018/12/5	4.67%	9.74%
2018/12/6	4.86%	13.17%
2018/12/7	0.79%	7.85%
2018/12/8	6.62%	8.92%
2018/12/9	4.49%	9.55%

The carrying capacity of the statistics of trade circulation data based on data mapping was analyzed. LoadRuner was used to simulate the number of concurrent online requests. The number of online requests was 100-500. The ability of the method to process online requests was analyzed. The results are shown in Table 6.

Table 6. Performance test results

Number of online requests	Average response time(S)	Average throughout (KB/S)	Passing rate
100	1.13	42361658	100%
200	2.64	42758658	100%
500	4.86	43255864	100%

It was found from Table 6 that the method could successfully process all the requests when the number of online requests increased from 100 to 500, and the passing rate achieved 100%, suggesting that the method was reliable.

5 Discussion

The development of the economy has brought more and more difficulties to the statistics of business circulation. The rapid development of the market economy requires the improvement of the efficiency of business circulation statistics. In the face of complex data, the traditional statistical methods are

inefficient and cannot achieve the good development of statistical work. Facing the problem of long-time and difficult data analysis [14], data mapping, as an emerging data analysis method, can efficiently process different types of data and plays an important role in data analysis [15]. It not only has a good application in the processing of pictures, videos, etc. [16], but also has a wide application in data analysis in fields such as medical analysis [17], industry and electronics [18]. Its application in business circulation statistics was analyzed in this study.

The business circulation statistics of Anhui province has large difficulties due to the large amount of data. In this study, the data were divided into six different categories and then preprocessed to prevent the data processing effect from degrading due to the existence of erroneous data. As to the data processing rules, this study extracted the high-frequency attributes of different categories, and selected the four attributes with the highest frequency of appearance as the key attributes: operating income, total assets, main business income and operating profit. The customs import data was used as an example to verify the key attributes, and the validity of these attributes for data matching was proved. As to the data analysis rules, the prediction of operating income was taken as an example to analyze the data mapping process through the ARIMA model in R language, and then the comparison between predicted results and actual values suggested that ARIMA model had good prediction result, indicating that the method was reliable.

In the performance test of data mapping method, the sequential, parallel and crossover algorithms were compared. The comparison of time and space cost suggested that the sequential algorithm had the largest time cost and the smallest space cost, while the crossover algorithm had the smallest time cost and the largest space cost. In practical applications, the time cost of algorithm is more important than space cost. Therefore, the crossover algorithm is more suitable for big data analysis. Then, the comparison between the ARIMA model and the Holt-Winters algorithm suggested that the predicted value of the former was closer to the actual value, and the average prediction error was 5.17%, which was significantly smaller than that of the latter, indicating that the former had better analysis ability than the latter and the choice of this study was correct. The method proposed in this study could successfully process all the requests, which suggested its high reliability.

But this study has some deficiencies. For example, the data collected was not comprehensive enough as data attributes were used for data mapping, which is limited and needs further study.

6 Conclusion

Data mapping method was applied in the business circulation statistics, the data processing rules and analysis rules were emphatically studied, and performance tests were carried out one by one, which proved the high reliability of the method and the excellent data analysis ability of data mapping. This work provides some references for the further application of data mapping in data analysis. Based on this study, we will select more statistical models to make analysis on the business circulation data to find out the optimal statistical prediction model.

References

[1] M. Leidig, R.M. Teeuw, Correction: quantifying and mapping global data poverty, Plos One 10(12)(2015) e0145591.

[2] Q. Hu, P. Liu, M.C. Huang, Threads and data mapping: affinity analysis for traffic reduction, IEEE Computer Architecture Letters 15(2)(2016) 133-136.

[3] D.C. Li, M. Huang, X.D. Li, Y.P. Ruan, L.X. Yao, MfeCNN: Mixture feature embedding convolutional neural network for data mapping, IEEE Transactions on NanoBioscience 17(3)(2018) 165-171.

[4] H.A. Reza, N. Devaki, S. Suparmi, N. Kusumawardani, Data source mapping: an essential step for health inequality monitoring, Global Health Action 11(sup1)(2018) 1456743.

[5] Q. Shang, Y. Fan, W. Shen, S. Shen, X.Y. Zeng, Single-port SRAM-based transpose memory with diagonal data mapping for large size 2-D DCT/IDCT, IEEE Transactions on Very Large Scale Integration (VLSI) Systems 22(11)(2014) 2422-

2426.

[6] M. Diener, E. Cruz, M. Alves, P. Navaux, A. Busse, H. Heiss, Kernel-based thread and data mapping for improved memory affinity, IEEE Transactions on Parallel & Distributed Systems 27(9)(2016) 2653-2666.

[7] A. Matta Gómez, J.D. Cerro Giner, A. Barrientos Cruz, Multi-robot data mapping simulation by using Microsoft robotics developer studio, Simulation Modelling Practice & Theory 49(49)(2014) 305-319.

[8] E.H.M. Cruz, M. Diener, M.A.Z. Alves, L.L. Pilla, P.O.A. Navaux, LAPT: a locality-aware page table for thread and data mapping, Parallel Computing (2015) S0167819115001556.

[9] M.S. Serpa, A.M. Krause, E.H.M. Cruz, P.O.A. Navaux, M. Pasin, P. Felber, Optimizing machine learning algorithms on multi-core and many-core architectures using thread and data mapping, in: Proc. Euromicro International Conference on Parallel, 2018.

[10] E.H.M. Cruz, M. Diener, L.L. Pilla, P.O.A. Navaux, Hardware-assisted thread and data mapping in hierarchical multicore architectures, ACM Transactions on Architecture & Code Optimization 13(3)(2016) 28.

[11] M.B. Ferraro, P. Giordani, A toolbox for fuzzy clustering using the R programming language, Fuzzy Sets & Systems 279(2015) 1-16.

[12] D.F. Findley, D.P. Lytras, A. Maravall, Illuminating ARIMA model-based seasonal adjustment with three fundamental seasonal models Series 7(1)(2016) 11-52.

[13] L. Ferbar Tratar, E. Strmčnik, The comparison of holt-winters method and multiple regression method: a case study, Energy 109(2016) 266-276.

[14] H. Qi, L. Peng, M. Huang, Threads and data mapping: reducing traffic via correlation and affinity, IEEE Computer Architecture Letters 15(2)(2015) 1-1.

[15] D.C. Li, M. Huang, X.D. Li, Y.P. Ruan, L.X. Yao, MfeCNN: mixture feature embedding convolutional neural network for data mapping, IEEE Transactions on NanoBioscience (2018) 1-1.

[16] F. Christoff, On attribute thresholding and data mapping functions in a supervised connected component segmentation framework, Remote Sensing 7(6)(2015) 7350-7377.

[17] A. Naveen, D. Peehoo, A.W. Toga, The GAAIN entity mapper: an active-learning system for medical data mapping, Frontiers in Neuroinformatics 9(118)(2015).

[18] W.J. Huang, Q.W. Hu, N. Zhu, Z.B. Tang, Methods for data mapping between MIL-STD-1388-2B and S1000D, Electronics Optics & Control 2014(6) 103-107.