Big Data Mining Method of New Retail Economy Based on Association Rules

Ying Liu(✉)

School of Business, Jinggangshan University, Ji'an, China
liuying1979@jgsu.edu.cn

Abstract. There are many sparse items in the data of new retail industry, and the extracted association rules are redundant, which leads to the problems of low temporal and spatial efficiency and poor mining quality when applied to the data mining of new retail economy. In order to improve the quality of data mining, a new retail economy big data mining method based on association rules is proposed. The k-means algorithm is used to subdivide the customer groups under the new retail economic model and extract the corresponding data association rules. Introduce interest threshold to filter association rules. The weight of frequent set is introduced to establish frequent item tree, and FP-Growth algorithm is improved to realize big data mining of new retail economy. Simulation results show that when the proposed data mining method is applied to the big data processing of the new retail economy, the execution efficiency and execution quality are improved by at least 27.5%.

Keywords: Association Rules · Big Data Of New Retail Economy · Data Mining · K-Means Clustering · Interest Degree

1 Introduction

New retail is a new business model that combines physical retail and e-commerce by using logistics systems to achieve "online and offline" integration, supported by digital technologies such as Internet technology, cloud computing and big data. As an emerging business model, new retail is still in the process of evolution, and driven by the new dynamics of the digital economy, it will be revitalized after precipitation. In the new retail economy, the rapid rise of Internet technology-based e-tailing has brought a huge impact to brick-and-mortar supermarkets. New retail is a business model reinvention that fundamentally transforms the business value innovation model and aims to enhance the consumer experience, so the new retail economy is very dependent on data. The new retail economy requires adequate application and integration capabilities for data so that retailers can think about digital transformation approaches at a strategic level in the context of their own business scope. However, the inability to fully analyze and use the data in the new retail economy has led to a large amount of new retail economy data being accumulated and stored in databases, forming data silos and increasing the operational costs of retail enterprises. In today's era of rapid information development,

© The Author(s) 2023
D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 1583–1590, 2023.
https://doi.org/10.2991/978-94-6463-030-5_159

because the needs of customers are always changing. The focus of the new retail industry economy is on customers, and the use of data mining technology can accurately guide companies in allocating marketing resources, which to a certain extent can maximize service quality and improve customer satisfaction within limited resources.

At present, data mining technology has been commonly used by many e-commerce companies to analyze data and obtain the required information and related business value. Foreign research on data mining technology is relatively early and its related technology is relatively mature, and data mining is used in many fields. The association rule algorithm aims to discover the strong association rules in the database by traversing and retrieving the transactional database step by step iteratively. The literature [1] addresses the problem of low mining accuracy due to inaccurate calculation of association center distance in traditional association rule mining, and uses the K-means algorithm to search and optimize association centers to improve the accuracy of data mining. The literature [2] used CNN networks to mine data with temporal relationships. However, the new retail economic development stock and thus the data generated in the navigation can have complex correlations, which puts forward higher requirements for the extraction of association rules and data mining. In order to improve the informationization level of retail enterprises and increase the data support for their management, operation and decision making, this paper will study the new retail economy big data mining method based on association rules and provide reference for the future application and promotion of data mining technology in other fields.

2 Research on Big Data Mining Method of New Retail Economy Based on Association Rules

2.1 Customer Segmentation Under the New Retail Economic Model

In the retail industry, product variety, quantity is big, the commodity type necessities in the majority, that is, to provide for people in your life related goods, so the customers to buy goods, has a certain connection between the commodity trading data lurking in the customer's buying habits and potential association. New retail economy mode, with the aid of big data technologies, such as customer after the purchase of goods, will produce goods trading data, data display a basket containing customer one-time buy all product data, and the basket analysis also is the shopping basket data correlation analysis, mining the association rules can help retailers find order, Develop corresponding marketing strategies [4]. In order to extract accurate data association rules in user consumption behavior and other data under the new retail economic model, this paper first subdivides customer groups. Compared with the traditional retail industry, the new retail industry has more comprehensive source data for customers' personal information and consumption information, consumption behavior, product preference and so on by virtue of the advantages of big data. Therefore, this paper uses k-means clustering algorithm to subdivide the customer groups of the new retail economy.

Because different customer segmentation indicators for customer segmentation importance is different; There may be outliers with abnormal consumption. Considering the disadvantages of the traditional K-means clustering algorithm, this paper will optimize the selection of K value and the selection of the initial cluster center. In the calculation of big data of new retail economy, behavioral characteristics such as daily consumption information, consumption frequency, product tendency and consumption level interval of customers are selected as clustering characteristics of customer group segmentation. The information entropy corresponding to each classification feature of customer data in the big data of the new retail economy is calculated as clustering weight to improve the clustering effect [3]. The data sample information entropy under each customer segmentation attribute is calculated as follows:

$$H(i) = -\sum_{i=1}^{n} p_j^i \lg p_j^i \tag{1}$$

In the above formula, p_j^i is the proportion of sample j under ith characteristic attribute of customer consumption behavior data in the big data of new retail economy. According to the difference coefficient of different customer consumption behavior data in feature dimension, the weight of each classification feature is defined, so as to obtain the weight distance of K-means clustering:

$$D\big(a_i, a_j\big) = \left[\sum_{i=1}^{m} w_i\big(a_i - a_j\big)^2\right]^{1/2} \tag{2}$$

In the above formula, w_i is the data feature weight of customer group segmentation, and its value is determined by the proportion of difference coefficient of feature dimension. According to the density index of the big data samples of the new retail economy and the maximum and minimum distance method, the selection of the initial cluster center is improved, that is, the initial cluster center of the data samples is selected by the density of the location of the sample data and the Euclidean distance between the data objects as a double index. The average integer of the classification characteristic information entropy was taken down as the K value of customer group classification, the density attribute value of each data point was calculated and arranged in ascending order, and K initial clustering centers were selected by combining the maximum and minimum distance method. Each sample point is classified into each cluster according to the weight distance. Calculate the mean value of the same object and update the cluster center. Repeat the above steps until the cluster center obtained reaches the maximum running times or no change occurs. The clustering algorithm to determine the clustering center and K value is used to complete the segmentation of customer groups under the new retail economic model.

2.2 Extraction and Filtering of Association Rules

In association analysis, the obtained association rules are not necessarily correct, and some rules may be generated by coincidence. Therefore, association rules need to be filtered to ensure the credibility of the applied association rules. The association rules contained in the big data of the new retail economy are presented with multi-valued attributes. Therefore, association rules can be extracted by discretizing the transaction data set by converting multi-valued attributes into binary items. Data feature dimension granulation algorithm is used to process the big data of new retail economy in order to accurately capture the data features in the frequent transaction data set.

If the number of data points in the ith dimension in the data space is N, the mean square deviation is $\bar{S}$, the mean value is $\bar{X}_i$, and the attribute accuracy is δ_i, the attribute accuracy calculation method is as follows [5]:

$$\delta_i = \bar{S}_i \big/ \bar{X}_i \tag{3}$$

If the attribute precision value in the data space is N, this paper selects the mean value of the attribute precision value as the selection parameter of dimension grain size to obtain the dimension grain size. The dimension grain is determined according to the population variance and estimation accuracy, and can be used as the number of analyzable dimensions to measure the granulation standard of dimensions. According to the above segmentation results of customer groups under the new retail economy model, the minimum support and confidence of extracting big data association rules of the new retail economy are set. Then scanning database, selected according to the given minimum support database in frequency itemsets, namely frequent itemsets. The frequent item set contains a combination of several transactions, and the permutations and combinations of these frequent transactions yield existing rules. Not all the obtained rules are useful, and the confidence degree of each association rule needs to be calculated. Only the association rule whose confidence degree is not less than the minimum confidence degree can indicate that there is strong correlation between transactions, and the association rule is also considered as strong association rule. Strong association rules of frequent item sets are mined according to screening of minimum confidence. And new retail economy mode, the big data correlation between increased, in order to reduce the redundant rules in data mining, the influence of the introduction of interest degree of filtering association rules, improve the quality of the rules.

Interest degree based on the improved association rule mining algorithm based on the coordinates of the distance between the two straight line to measure the size of the degree of interest, the smaller the distance, the greater the degree of interest. By setting the parameter threshold of interest, the uncertainty caused by the double threshold of support and confidence is improved, and redundant rules are deleted.

For association rule $B \Rightarrow V$, the coordinate system is constructed with rule antecedent B as abscissa and rule antecedent V as ordinate. It is also stipulated that the positive direction of the horizontal axis and the vertical axis is used to represent the positive item set, and the negative direction is used to represent the negative item set. Trace the points representing the association rules on the coordinate axes, and calculate the Angle formed by the straight line between the rule ante-piece and the rule post-piece.

$$\cos\langle b, v\rangle = \frac{|b_1 b_2 + v_1 v_2|}{\sqrt{b_1^2 + v_1^2} \times \sqrt{b_2^2 + v_2^2}} \tag{4}$$

The distance between the straight line of the regular antecedents and the regular antecedents is calculated by using the above included Angle formula. The calculated distance between the two straight lines is the degree of interest of the big data association rules of the new retail economy. According to different sales strategies, the threshold value of interest degree of association rules is established to filter the redundant rules of big data mining in the new retail economy.

2.3 Realize Big Data Mining of New Retail Economy

In view of the characteristics of the big data of the new retail economy, this paper optimizes the FP-growth algorithm and uses the improved FP-growth algorithm to realize the mining of the big data of the new retail economy. In the improved FP-growth algorithm, when mining and analyzing according to the constructed frequent pattern tree, it only needs to find out all frequent pre-nodes and frequent post-nodes of each non-leaf node, and then connect them to form frequent item sets.

This paper introduces the concept of weight of frequent set and optimizes the efficiency of frequent pattern tree in data mining by the feature of item set. The boundary items of frequent sets are scanned without the purpose of scanning the entire incremental data set again, and the integration support threshold weights such item sets. When the weight given to the mined boundary item set is larger, it is more likely to change frequently. In other words, the weight of the item set is calculated as follows [6]:

$$\rho(K) = [\zeta - S(K)]^{-1} \tag{5}$$

Where, $\rho(K)$ is the weight of the boundary item set K; ζ is the threshold of interest degree; $S(K)$ represents the support degree of K item set in the original data. Taking the minimum support $\min(K)$ of item set K as the frequent data node to determine the standard, the multi-frequent item support tree was established as the frequent pattern tree of the improved FP-growth algorithm. The frequent pattern tree contains not only all frequent itemsets, but also those infrequent itemsets whose support is greater than the minimum. After the frequent pattern tree is established, all non-leaf nodes in the frequent pattern tree are traversed from left to right, and their support degree is calculated and compared with the given minimum support degree to obtain frequent nodes. Through frequent nodes, the support degree of the front node and the back node is judged, and the frequent front node and the frequent back node are connected with frequent nodes, so as to obtain frequent itemsets and conditional frequent itemsets. By corresponding the frequent item set mined with the big data of the new retail economy, the deep information in the data can be analyzed to help retail enterprises make marketing plans in line with their development needs.

3 Simulation Experiments and Discussion

3.1 Simulation Content

This section validates and compares the running results and execution time of the proposed association rule-based big data mining method for new retail economy, k-means clustering-based data mining method and CNN-based data mining method above on the same platform using a synthetic dataset. The data in the dataset used for the simulation is derived from the backend transaction data of two large e-commerce retail platforms,

Table 1. Parameters of the artificially synthesized data set.

	TB1	TB2	JD	EB
Number of records	65428	23417	39842	3401
Number of projects	1274	136	252	71
Data set size/G	55.8	12.6	24.9	10.8

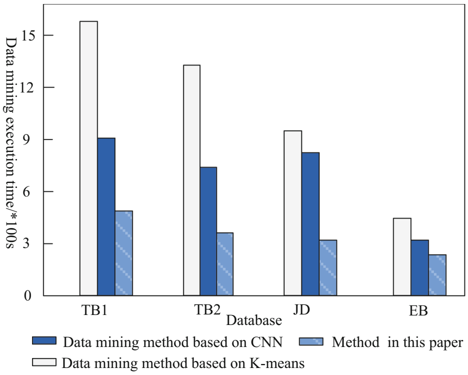

Fig. 1. Comparison of execution efficiency of data mining methods.

with a total of 163,672 pieces of data containing 2163 different products with string data type. Each of these data is a single shopping basket of a customer, containing information about at least one product. Also, relevant data were manually compiled to enrich the data types within the experimental data set. Table 1 shows the parameters of the artificially synthesized data set used in this simulation.

3.2 Simulation Results and Analysis

The three data mining methods are analyzed by taking the average value after each independent run 20 in each experimental data subset, and comparing the execution performance of the mining methods under different data set sizes by comparing the time complexity.

As can be seen from the graphical1 analysis, the performance of this paper's method remains in a dimensionally stable state for smaller data sets, and its execution time does not improve to a great extent, but on the contrary, for larger data sets, the algorithm execution time decreases significantly with the increase of item set support, and the mining efficiency achieves a great improvement. For the larger data size dataset TB1, the execution time of the method in this paper is reduced by at least 27.5% relative to the other two compared methods, with outstanding advantages of data mining execution in the big data environment.

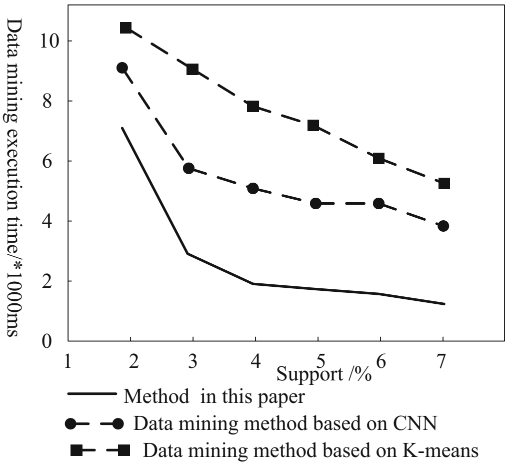

Fig. 2. Comparison of method execution time under different support degrees.

Set up different data mining methods with frequent item support and compare the execution time of the three data mining methods. Integrate the above analysis to compare the execution effect of data mining methods.

From Fig. 2, we can see that the execution time of this paper's method is overall less than the comparison methods, and as the support degree decreases, the more the time performance of this paper's method is higher than the other two comparison methods. This is because the method in this paper eliminates a large number of invalid itemsets, and as the support decreases, the algorithm produces more and more candidate itemsets, while the method in this paper uses the interest to support processing to remove the redundant invalid itemsets, thus improving the time performance of the mining method.The above results show that the new retail economy big data mining method based on association rules proposed in this paper has relatively better mining quality and efficiency.

4 Conclusions

The rise of new retail in China is closely related to the current era in which online retail is experiencing a rising bottleneck and physical retail is struggling to develop. New retail is not a simple fusion of online and offline, but a new model that transcends omnichannel based on technological innovation with big data as its core. New retail is a new retail model renewed by multiple elements, emphasizing experience and cost reduction and gain, while highly accepting emerging technologies. Therefore, in the process of new retail economy development, the use of data mining technology to uncover patterns and potentially valuable guidance information in the data can facilitate corporate businesses to implement accurate marketing and improve service quality. The association rule-based big data mining method for new retail economy studied in this paper can process complex big data for the current new retail model and obtain deep data that is beneficial to the development of the economic model, which provides technical support to promote the high-quality development of the new retail model and improve the informationization of the new retail economy in the future.

References

1. Chuen-Jiuan Jane. (2021). Analysis of Network Anomaly Data Mining Based on Association Rule Algorithm. International Journal of Uncertainty and Innovation Research, 3(2), 169–176.
2. GAN Xin-yan; TANG Xiao-nian.(2021). Mining Model of Association Rules for Temporal Data Based on CNN. Computer Simulation,38(03),282–285+326.
3. Hikmawati Erna ; Maulidevi Nur Ulfa ; Surendro Kridanto. (2021). Minimum threshold determination method based on dataset characteristics in association rule mining. Journal of Big Data, 8(1), 146
4. Meruva Reddy Subba ; Bondu Venkateswarlu. (2021). Review of Association Mining Methods for the Extraction of Rules Based on the Frequency and Utility Factors. International Journal of Information Technology Project Management (IJITPM), 12(4), 1–10.
5. Papadimitriou Georgios ; Komninos Andreas ; Garofalakis John. (2020). Predicting retail business success using urban social data mining. Journal of Ambient Intelligence and Smart Environments, 12(3), 263–277.
6. Prayitno Mokhammad Hadi ; Rasim Rasim. (2018). Analisa Penjualan Produk Retail Dengan Metode Data Mining Asosiasi. Jurnal Kajian Ilmiah, 18(3), 231–231.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

BY NC