Eng Int Syst (2019) 4: 177–184

Engineering
Intelligent Systems

Big Data Analytics in the E-commerce Retail Industry

Weiwei Cheng* and Jing Wang

Logistics and Trade College, Xi'an Eurasia University, Xi'an, Shaanxi 710065, China

The rapid development of e-commerce retailing has led to a sharp increase in the amount of information being generated, which results in an abundance of data. From this huge volume of data, much important knowledge can be extracted, particularly in regard to consumers' preferences and buying habits. Therefore, a new generation of technologies and tools is needed to conduct comprehensive and higher-level analysis of massive amounts of data, to facilitate inductive reasoning, to extract potential patterns, and to extract useful knowledge. Starting from the perspective of data mining, this paper combines e-commerce data rules and network analysis to analyze e-commerce clustering points and obtain corresponding results. Studies have shown that the method proposed in this study is effective.

Keywords: big data; data mining; e-commerce; retail; cluster analysis

1. INTRODUCTION

With the continuous development of information technology, more sophisticated applications, and the widespread popularity of the Internet, e-commerce has become increasingly popular as a new business model for modern enterprises. From reducing transaction costs to overcoming time and space constraints, from improving work efficiency to integrating upstream and downstream resources, from enhancing customer service to providing convenient and efficient online banking payment, e-commerce has had a profound impact on enterprises in various aspects such as the business model, management methods and payment methods [1]. Data mining is an information technology that has emerged in recent years with the development of database technology. It integrates the knowledge of various disciplines such as database, artificial intelligence and statistics, and applies the ideas and methods of data mining to e-commerce. Moreover, through the combing, summarizing and extracting of information from the huge volume of e-commerce transaction data, companies can obtain knowledge about current and potential customers to an unprecedented extent. This analysis can significantly assist companies with decision-making, guiding business operations [2], and providing targeted and more valuable services to customers.

2. RELATED WORK

Global research on data mining in e-commerce has also made great progress. In the web data mining technology field, Xuerong W [3] is a well-known expert in web mining and data mining. He has worked on hypertext databases and data mining, has extensive experience in real-world project development, and has secured several US patents. Although there are already many e-commerce recommendation systems on the market, there are still many problems. When designing an e-commerce system, several problems and considerations must be taken into account. These are explained below.

First there is the timeliness of web data mining. At present, most of the web data mining systems used in e-commerce systems do not have real-time functions, and the real-time requirements are not met when extracting data, mining data and integrating information from mined data [4]. It is a lengthy process from the collection and pre-processing of data to the mining of valuable data.

A holistic development platform must be provided. At present, the web data mining architectures of most

*Corresponding author: Weiwei Cheng, Email:sichengjingcst@163.com

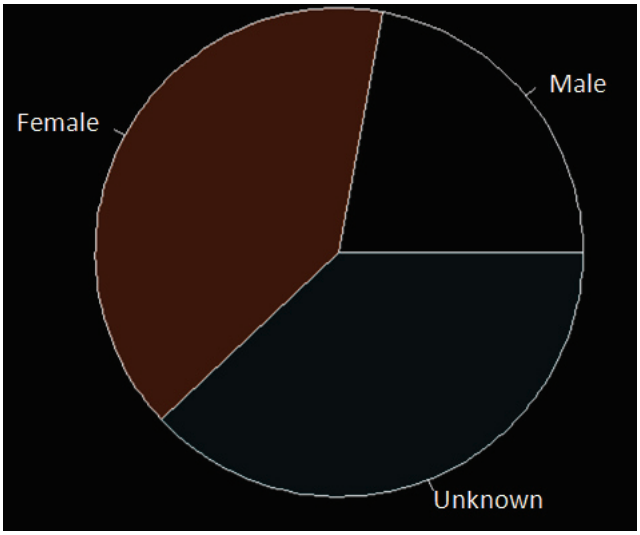

Figure 1 Online shopping groups according to gender.

e-commerce systems are not perfect. The data mining module is generally used as a module of the e-commerce system when designing the system [5]. The system is still designed according to the business logic of e-commerce, and does not include a complete data mining platform [6].

There is the problem of which data mining algorithms to choose. Data mining technology has many excellent data mining algorithms, each of which has its own features and applications [7]. Most of the current e-commerce systems are not ideal for data mining technology; hence, some data mining algorithms cannot be applied [8].

There is the problem of performance and cost. In order to meet the data mining computing requirements of e-commerce systems, companies need to invest heavily in hardware, software and system maintenance [9]. These increased costs increase the cost of system applications, and shut out small e-commerce companies, thereby preventing the uptake of system applications [10].

In China, relatively few enterprises have adopted adaptive e-commerce systems, and not all enterprises can afford to invest in the design and development of such systems [11]. With the development of information technology, data mining and encryption technologies are constantly developing and perfecting, which makes it possible for enterprises to design and implement their own e-commerce systems. Moreover, the company can design several features such as product browsing, personality recommendation, content search and management of customer payments via the system according to the company's specific requirements, so that it is completely suitable for its operations.

3. EXPLORING THE ASSOCIATION RULES OF RETAIL WEBSITES

Exploratory analysis of the data was performed using the code demonstrated in find_item. Rprior to mining the association rules. The purpose is to provide a service tool for a single online store to develop ideas. However, some interesting results have been obtained, as shown in Figures 1 and 2 [12].

The above distribution of gender and age indicates that the consumer groups of maternal and children's food are mainly women around 30 years of age in the developed areas, which is actually the main feature of the current age-appropriate mother group.

Figure 3 shows that the peak of online shopping for maternal products and children's food is Monday and then decreases over time. This suggests that because parents may be more relaxed on weekends, this may be the peak period for online shopping activity. However, it seems that everyone is free to shop online during working hours, and during the rest time on weekends the baby still needs to be cared for [13].

Figure 4 shows the online purchasing trend over an entire month. According to the weekly rule, the difference in the transaction amount is greater than the frequency. It indicates that Monday is a good time for online shoppers to spend money. Moreover, it seems to perfectly reflect the truth of "not going to work and not shopping."

4. MODEL DEVELOPMENTS

The meanings of support and confidence are easily understood, and intuitively the actual values are the two estimates mentioned in the above explanation [14]. They can be calculated with the following formula (the formula is for commodity X and commodity Y)

$$\sup\ port(X \Rightarrow Y) = P(X \cup Y)$$
$$confidence(X \Rightarrow Y) = P(Y|X)$$

The concepts corresponding to support and confidence are probability and conditional probability. However, what is the lift? In order to illustrate this problem, we need to study it further. We noticed that if a transaction has a large share of the trading database (for example, if all people bought this item), we can easily obtain a series of related association rules that meet the threshold. However, all these rules might not necessarily be interesting, and some could be obvious errors.

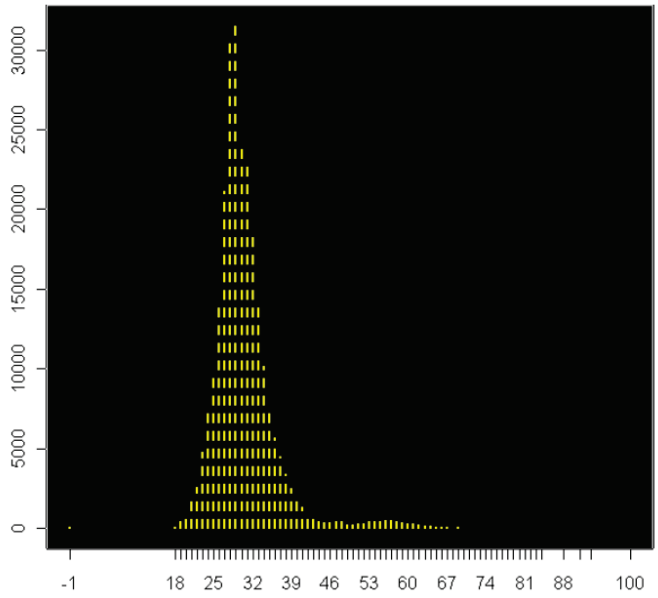

Figure 2 Age distribution of online shopping groups.

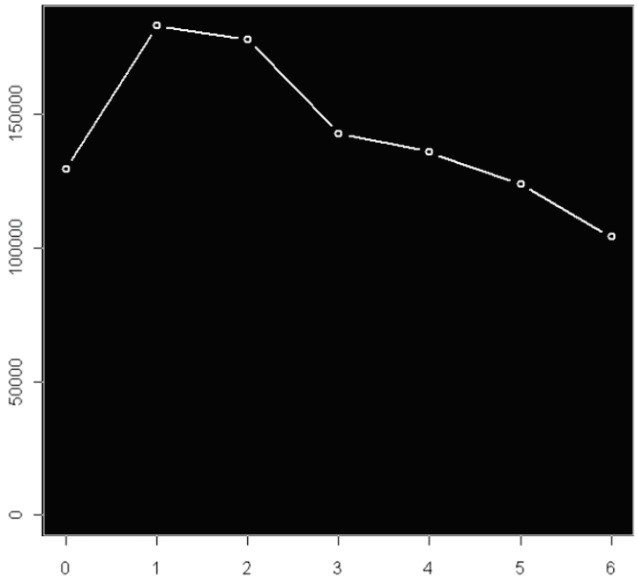

Figure 3 Trends in the number of purchases by all users from Sunday to Saturday.

It is not difficult to cite such examples. Moreover, there is an irrelevant or even negative correlation between the association rules [15]. Therefore, an indicator needs to be defined. It refers to the correlation between the commodity X and the commodity Y, which is called the degree of lift. Only itemsets that satisfy the degree of lift $lift>1$ are positively correlated. The lift is calculated with the following formula.

$$lift(X \Rightarrow Y) =$$
$$\text{support}(X \Rightarrow Y) \Big/ (\text{support}(X) * \text{support}(Y))$$
$$= P(X \cap Y) \Big/ P(X \cup Y)$$

We can use this indicator to measure the interestingness of the rules, which is particularly evident in the above analysis, especially in the case of small data sets. The reason is that if all the transactions of an item (such as X) have it, the rules are irrelevant [16]. Any rule related to it can be obtained from the above formula, and its lift value is 1. Specifically, the presence of this item in the shopping basket has no effect on any other product and is not related to any other product. Association rule mining sometimes obtains negative rules. At this time, there is $lift<1$. The lift indicator is used to distinguish which items X and Y are in a shopping basket, which is an accidental phenomenon that we do not care about.

Based on the above analysis of the results obtained by the association rules, we found that for each individual seller in all stores, our association rules mining will have some support as follows.

Suggestions for merchandise display: We need to optimize the store's merchandise structure by, for example, placing the Granville brand rice paste and the Granville brand molar stick

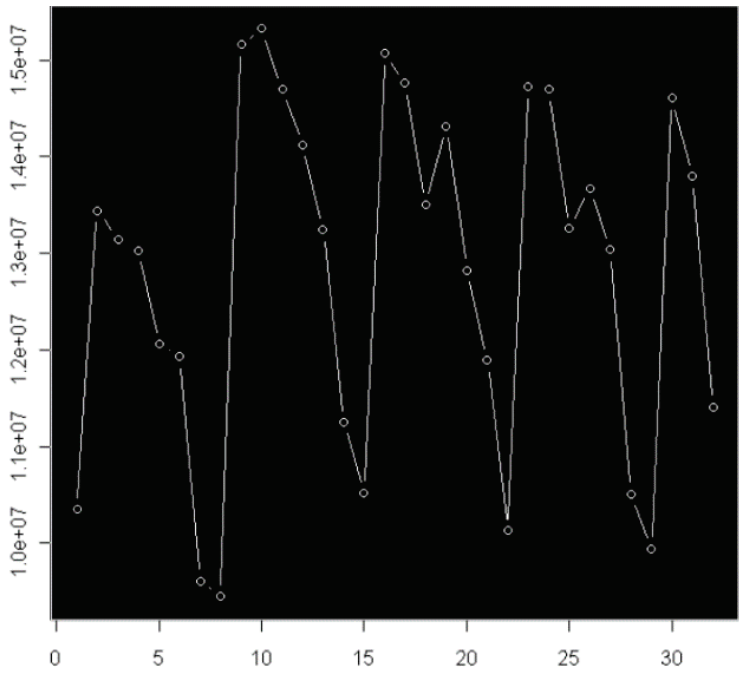

Figure 4 Trends in the purchase amount of all users from the beginning of the month to the end of the month.

under the same display, which can increase the sales of related products.

Suggestions for combined sales: We need to package the related products for sale. For example, the Granger brand rice paste and the Granville brand toothpicks are packaged in a 3-item mini package, which provides a more attractive combination of discounts.

Suggestions for product recommendation: We need to provide a page for the associated item. For example, after the user purchased the Granville brand rice paste, the Granville brand toothpick was recommended to the user, which provided an appropriate and effective product recommendation. Moreover, the conversion rate from order to purchase will increase accordingly.

5. MODEL DEVELOPMENT AND RESULTS

According to the research results of the literature, we chose to use hierarchical clustering to analyze the geographical location of 510 merchants of a group buying website. Hierarchical clustering means that each object is first divided into one class, and then the most recent class is merged. The function of hierarchical clustering in R uses the longest distance method by default; that is, the distance between groups is defined as the distance between the two objects that are the farthest between the two groups. We set the final number of categories to 6. It should be noted that we are actually clustering the samples as shown in Figure 5.

Although the details of the business cannot be displayed on the schematic diagram due to the excessive number, the figure nevertheless shows a very obvious difference between the six categories. It can be seen from Figure 6 that the position of the store is strongly related to the number of administrative divisions. Basically, if some businesses are grouped together, the distance between groups is much larger than the distances within the group. The distance function used by default in the clustering function is the Euclidean distance. In this small range (the difference between longitude and latitude does not exceed 0.4 degrees), our geographic coordinates can be expressed as Cartesian coordinates, and the geographical distance is calculated in the same way as the Euclidean distance. Figure 6 shows more clearly the correlation between classes, and classes of hierarchical clustering . It can be seen that the differences between categories are relatively large.

Since the longitude and latitude data we use is naturally suitable for representation on the map (only requiring separate horizontal and vertical coordinates), we can create a "map" showing the clustering effect. The gray box in the figure represents the center of the hierarchical clustering.

Figure 7 shows several large business districts in the main urban area of Beijing. The number of results is related to the amount of commerce being conducted in each district, but it is also related to the promotion of the group buying website in this area.

The division of a business circle for each project can provide decision support to the group marketing websites to allocate marketing resources to find more businesses in the region. Secondly, the website can provide targeted and accurate marketing for users who have purchased products in these business districts (including group messaging and subscription email delivery for various projects). Accurate marketing based on the analysis of geographic location clustering will have an obvious effect on the promotion of marketing coupons.

Finally, we present the results of hierarchical clustering using Google Earth software (the .csv file exported from R is converted to a .kml file recognized by Google Earth, and this step takes advantage of Python code). The six pointers in blue in Figure 8 are the center of the hierarchical

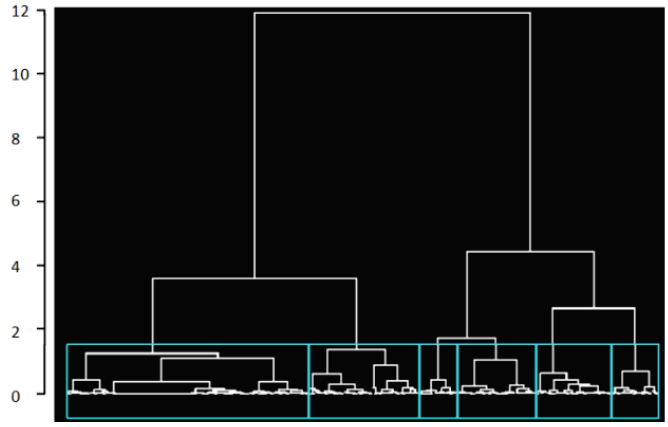

Figure 5 Schematic diagram of the cluster tree of the sample level.

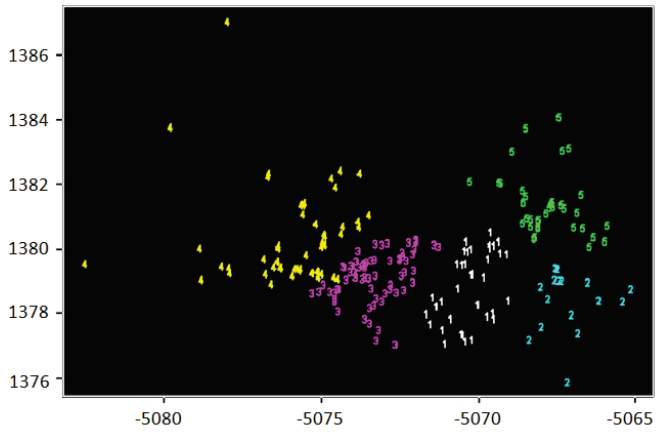

Figure 6 Diagram of the correlation between classes, and classes of hierarchical clustering.

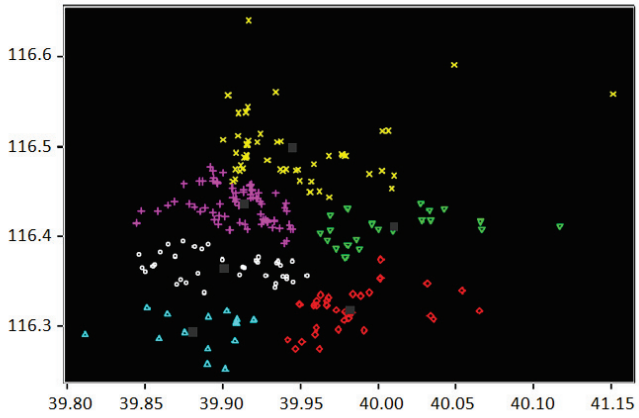

Figure 7 Diagram of hierarchical clustering effects.

clustering, and the remaining light red droplets represent all merchants.

The so-called partitioning clustering is used to classify samples according to distances and determine the cluster centers. In this paper, we mainly use the K-means algorithm for partitioning clustering. The core step of the algorithm is to recalculate the cluster center according to the newly-generated class after categorization, and then re-categorize the iterative process. The iteration stops when the cluster no longer changes or the change is below a certain threshold. Moreover, the K-means algorithm has initial value sensitivity.

For the reasons stated above, we set the final category parameter K to 6. To ensure that the results of each K-means clustering are the same, we also specify random seeds to make the initial values of the clusters consistent. The graph of the sample clustering tree is not much different from the graph above, so it will not be repeated [17]. The figure below depicts the clustering results:

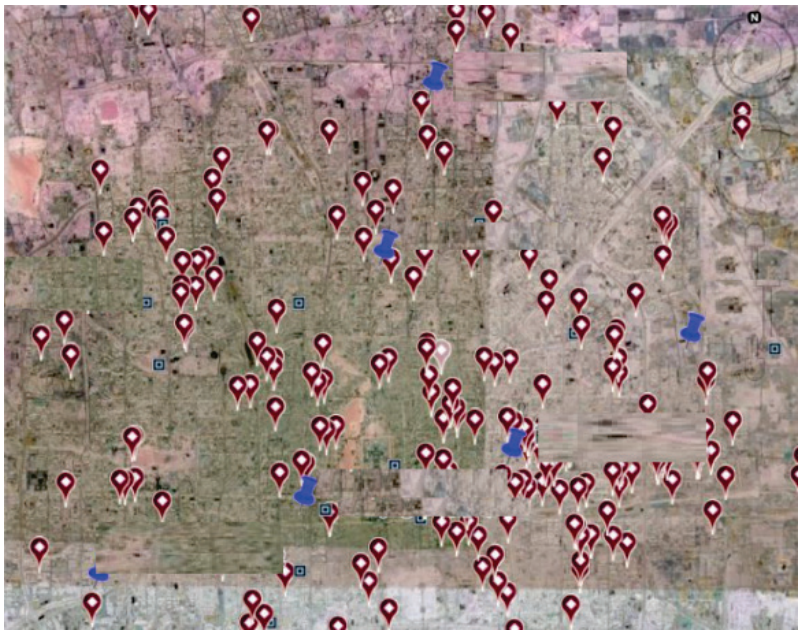

Figure 8 Hierarchical clustering results on Google Maps.

```
K-means clustering with 6 clusters of sizes 26, 46, 38, 43, 142, 43

Cluster means:
      x_point     y_point
1 40.03092236 116.4518436
2 39.93798229 116.3801131
3 39.87441828 116.3205310
4 39.98186806 116.3171077
5 39.91317870 116.4391274
6 39.92912223 116.5044292

Clustering vector:
 119  131  169  170  200  201  202  205  281  282  283  288  314  315
   3    3    5    6    5    5    6    3    6    5    4    5    3    3
 445  446  466  468  469  473  474  476  496  522  532  538  548  549
   5    1    5    2    2    3    3    3    6    2    5    2    5    2
 668  679  696  701  720  737  742  774  783  790  800  805  807  830
```

Figure 9 K-means clustering results in the program.

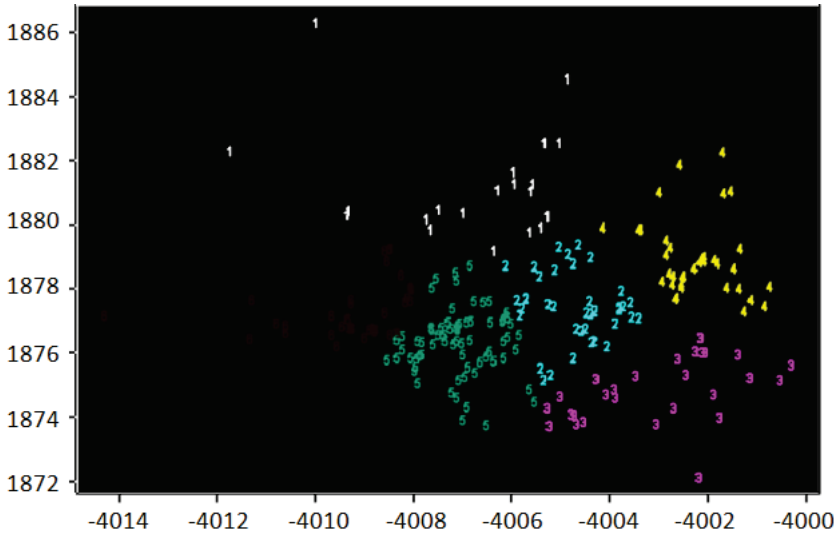

Figure 10 Correlation diagram between classes, and classes of K-means clustering.

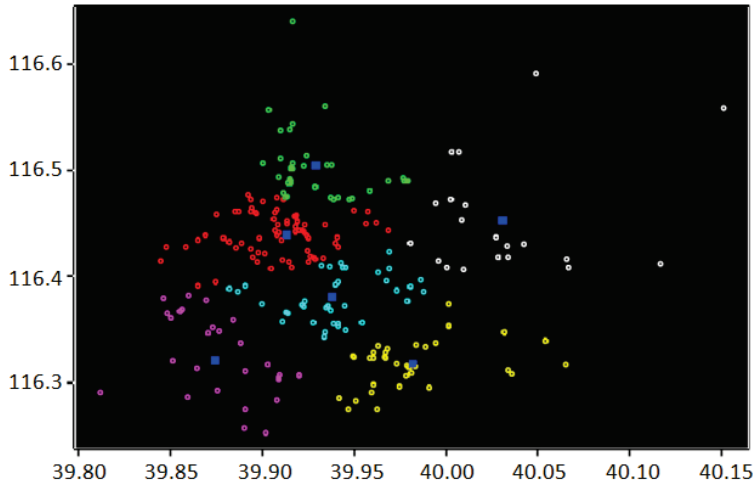

Figure 11 Diagram of K-means clustering effect.

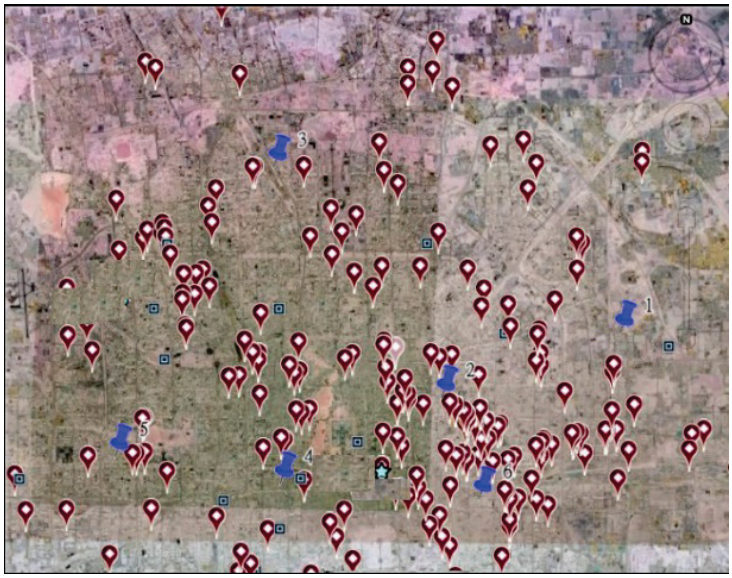

Figure 12 Display of K-means clustering results on Google Maps.

Second, we show the correlation between classes, and the classes obtained by the K-means algorithm, which is different from hierarchical clustering:

For clarity, we still create the clustering effect diagram in the form of a "map". The blue box in Figure 10 represents the center of the cluster. It can be seen from the figure that the difference from hierarchical clustering is quite obvious. Moreover, compared to hierarchical clustering, it seems that the position of the cluster center is more reasonable.

The division of such a business circle for each project can provide decision support to the group marketing websites to allocate marketing resources to find more businesses in the jurisdiction. Secondly, the website can provide targeted and accurate marketing for users who have purchased products in these business districts (including group messaging and subscription email delivery for various projects). Accurate marketing based on the analysis of geographic location clustering will have an obvious effect on the promotion of marketing coupons, and it will play a greater guiding role for the website to continue to find businesses that can be promoted.

Finally, we present the results of the K-means algorithm using Google Earth software (the .csv file exported from R is converted to a .kml file recognized by Google Earth, and this step takes advantage of Python code). The six pointers in blue in Figure 8 are the center of the hierarchical clustering, and the remaining pink droplets represent all merchants.

6. CONCLUSION

With the development of technology related to e-commerce, its design has also begun to change from the traditional C/S mode to the B/S mode. After applying data mining technology to the e-commerce system, there is a strong need to ensure the integrity of the e-commerce system. Therefore, when designing an e-commerce system based on web data mining, it is necessary to consider the specific situation of data mining in the overall design of the system. Based on the construction

of e-commerce retail data mining, this paper introduces the idea and method of data mining. Moreover, this paper collects, analyzes and mines commercial data information, and uses association rules algorithms and cluster analysis techniques to achieve the related sales of e-commerce retail and the segmentation of retail households. Through the above methods, this paper maximizes the market potential, provides enterprises with rich, multi-level information aggregation and industry data analysis, and provides decision-makers with rich decision support and strategic management to promote the further development of e-commerce retail.

ACKNOWLEDGEMENT

Construction project of the key course “Online retail” in 2018. Project editor:2018KC031.

REFERENCES

1. Hu, Junmin. E-commerce big data computing platform system based on distributed computing logistics information[J]. Cluster Computing, 2018.
2. Malhotra D, Rishi O P. An intelligent approach to design of E-Commerce metasearch and ranking system using next-generation big data analytics[J]. Journal of King Saud University - Computer and Information Sciences, 2018:S1319157817303440.
3. Xuerong W, Nianhong W. Dynamic prediction model on export sales based on controllable relevance big data of cross-border e-commerce[J]. Journal of Computer Applications, 2017.
4. Zhang B, Du Z, Wang B, et al. Motivation and challenges for e-commerce in e-waste recycling under “Big data” context: A perspective from household willingness in China [J]. Technological Forecasting and Social Change, 2018: S0040162517313501.
5. Akter S, Wamba S F. Big data and disaster management: a systematic review and agenda for future research[J]. Annals of Operations Research, 2017.
6. Liu S, Xiao F, Ou W, et al. Cascade Ranking for Operational E-commerce Search [J]. 2017.
7. Christiane L, Alexander W, Jan V B , et al. How Big Data Analytics Enables Service Innovation: Materiality, Affordance, and the Individualization of Service[J]. Journal of Management Information Systems, 2018, 35(2):424–460.
8. Nazeer H, Iqbal W, Bokhari F , et al. Real-time Text Analytics Pipeline Using Open-source Big Data Tools[J]. 2017.
9. Ahmad J, Muhammad K, Lloret J, et al. Efficient Conversion of Deep Features to Compact Binary Codes using Fourier Decomposition for Multimedia Big Data[J]. IEEE Transactions on Industrial Informatics, 2018:1–1.
10. Verma N, Singh J, Liu S. An intelligent approach to big data analytics for sustainable retail environment using apriori–map reduce framework [J]. Industrial Management & Data Systems, 2017:00–00.
11. Sathiaraj D, Cassidy W M, Rohli E. Improving Predictive Accuracy in Election[J]. Big Data, 2017, 5(4):325.
12. Chen R, Xu W. The determinants of online customer ratings: a combined domain ontology and topic text analytics approach [J]. Electronic Commerce Research, 2017, 17.
13. Wang H, Liu D. Is servitization of construction the inevitable choice of Internet Plus construction[J]. Frontiers of Engineering Management, 2017.
14. Hussaina A, Cambriab E. Semi-Supervised Learning for Big Social Data Analysis[J]. Neurocomputing, 2017, 275.
15. Vandic D. Intelligent Information Systems for Web Product Search[J]. 2017.
16. Tian X, Liu L. Does big data mean big knowledge? Integration of big data analysis and conceptual model for social commerce research[J]. Electronic Commerce Research, 2017, 17(1):169–183.
17. Song Z, Sun Y, Wan J, et al. Smart e-commerce systems: current status and research challenges[J]. Electronic Markets, 2017.