Keywords Research on E-commerce Platforms

-- Text Mining based on Factor Analysis and K-means Clustering

Yuan Li

Department of Economics, University of Jinan, Guangdong, China

Abstract

This paper collects keyword data of a small and medium-sized business in Alibaba International Station, and establishes a K-means clustering model based on factor analysis statistics. First, we reduce the dimensionality of its keyword indicators using Factor Analysis, then use the common factor score as the clustering indicator to perform K-means Clustering analysis, and finally establish DTM matrix and If-Idf matrix for the classified keywords to perform text Digging to find the representative characteristic words of each cluster, discovering the problems and opportunities behind them, so as to provide the merchants with relevant suggestions on product optimization and keyword selection.

Keywords

E-commerce Platform; Factor Analysis; K-means Clustering; Text Mining; If-Idf.

1. Introduction

Under normal circumstances, the search volume reflected by keywords means changes in the market. At the same time, most E-commerce platforms now allow merchants to pay to promote keywords. Therefore, the factors that affect the product ranking of this keyword in the E-commerce platform are more important. many. Meanwhile, for E-commerce platforms with B2B business models, keywords are extremely numerous and messy, and there are many product categories in the market. Most of the small and medium sized merchants are manufacturers, and they usually adopt the following more subjective operation model: choose the main product through experience, choose the keywords matched by the system by default, and carry out paid promotion provided by the platform. This subjective model has certain shortcomings, requiring a large investment in promotion costs, and the rate of return on investment is not high.

2. Related Theory

Factor Analysis (FA)[1] uses several potential random quantities that cannot be directly observed and queried to describe the covariance relationship between multiple variables. Although there are relatively few keyword index variables in this article (<10), but due to the large correlation between the variables, this article evaluates the correlation between the six keyword index variables and divides them into less Several common factors that are not linearly correlated and have practical explanatory properties, which are convenient for subsequent analysis.

The K-means clustering method[2] was proposed by McQueen (MacQueen, 1967). The basic idea is to assign each observation object to the cluster (class) closest to the center (mean), and use the Euclidean distance as the similarity measure to iteratively AN observation object is classified into the cluster (class) closest to the center until all samples can no longer be classified. The essence is to iteratively improve the variance within the cluster, aiming at high similarity within the cluster and low similarity outside the cluster.

Text Mining[3] is a method of processing unstructured data such as text in the field of Natural Language Processing (NLP). It was born from the mining demand generated by the explosion of unstructured data on the Internet, and is widely used in information retrieval and quantitative and qualitative analysis in various fields, and it can be said to play an extremely important role.

It essentially extracts and quantifies text feature words, including summaries, text classification, part-of-speech tagging (POS), sentiment analysis, and simple visualization using keyword clouds to facilitate people to conduct related research on a large amount of text., Dig out the hidden meanings, emotions, needs, etc. behind the text.

In the context of the Internet, keywords broadly refer to the subject of documents searched by users in search engines, and can reflect the core vocabulary of the content to be understood. In the category of E-commerce platforms, keywords in a narrow sense refer to the core vocabulary that users enter when searching for products that can reflect the nature and characteristics of their product needs.

Therefore, in the E-commerce platform[5], the overall popularity of keywords reflects the market supply and market demand of related products, and the traffic and clicks brought by keywords reflect whether merchant-related products faithfully meet market needs.

According to the concept and function of keywords, referring to the operating indicators on mainstream E-commerce platforms, there are three main indicators for evaluating keywords in the platform:

1. Keyword performance status: the performance status of keywords includes the exposure, clicks, visitors, feedback, transactions, etc. of keywords in related products.

2. The demand for keywords: evaluating the popularity of keyword demand includes search volume, search frequency, etc.

3. Keyword competition fever: evaluating the popularity of keyword competition is mainly based on the number of merchants for paid promotion of keywords, the paid promotion price, and the number of total keywords related products.

3. Clustering Analysis based on Factor Analysis

In this section, we introduce a new Clustering analysis based on factor analysis, we first analyses the meaning of combining these two methods. Then we present the steps of the proposed method.

3.1. The Motivation of Cobining Factor Analysis and K-means Clustering

(1) Factor analysis can provide linearly irrelevant public factors as new data indicators for cluster analysis, reduce the correlation of variables, and make up for the shortcomings of K-means clustering method that is sensitive to outliers.

(2) The few common factors retained after factor analysis are used as new data indicators for clustering, which reduces clustering variables and improves the running speed of clustering.

(3) When the number of common factors retained by factor analysis is 2, it is convenient for cluster visualization.

(4) The common factors after factor analysis can be used as variables to improve the interpretability of K-means clustering.

3.2. Clustering Analysis based on Factor Analysis Method

Our proposed method steps are described as follows:

(1) Data processing

First, suppose there are m keywords, and each keyword has n keyword evaluation index variables. Based on the differences in the data units queried in this article, in order to avoid the undesirable consequences caused by the difference in the observational dimension and its magnitude, it is necessary to standardize the sample observation data. The standardized sample data matrix is:

$$X=\begin{bmatrix} X_{11} & X_{12} & \cdots\cdots & X_{1n} \\ X_{21} & X_{22} & \cdots\cdots & X_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ X_{m1} & X_{m2} & \cdots\cdots & X_{mn} \end{bmatrix}$$

Among them, the vector coefficient of X is an observable random variable, and its mean is E(X)=0, and the covariance matrix COV(X)=1.

(2) hypothetical test of Variable Normality and Feasibility of determining factor analysis

In this step, we should perfume hypothetical test to determine the whether the data can satisfy our requirements. Firstly, we use KMO and Bartlett sphericity test to determine Feasibility of factor analysis. After confirming the feasibility of factor analysis, the normality test for univariate can be carried out by normal transformation, QQ plot, and K-S test method, and the normality test for multivariate can be carried out by Shapiro-Wilk.

(3) Calculate the correlation coefficient matrix R, as below:

$$R=\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix}$$

(4) Perform factor analysis and calculate the eigenvalues and eigenvectors of the observation data. The simple steps are: Let |R-λI|=0, find the eigenvalue and the eigenvalue vector.

(5) Calculate factor loading matrix A,as below:

$$A=\begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{m1} & \cdots & \alpha_{mn} \end{bmatrix}$$

(6) Factors for extracting principal components

Select p(p<n) main factors (marked with F), so that the sum of the variance contribution rate of the p main factors accounts for more than 85% of the total variance contribution rate, or the eigenvalue corresponding to the main factor is greater than 1, then It means that these main factors basically retain the information of the original analysis indicators, and reduce the original n analysis indicators to p factors, achieving the purpose of simplifying the optimization model of the analysis indicators.

(7) Establish a factor model of relevant significance as follows:

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p + \varepsilon_1 \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p + \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots \\ x_n = a_{n1}F_1 + a_{n2}F_2 + \cdots + a_{np}F_p + \varepsilon_n \end{cases}$$

(8) The maximum orthogonal rotation of variance (varimax)

After the first factor analysis, it is necessary to find from the factor loading table whether each main factor is related to the relevant information of the original observation data, and to sum up the information it contains to find new ideas and factors in a new sense. It is explanatory to the sample. If the factor loading table shows that the factor has little correlation with the original index, that is, the factor loading at this time does not meet the "simple structure criterion", and the typical representative variables of each factor are not very prominent, which will make it impossible for us to make sense of the factor. Explanation. For this reason, it is necessary to rotate the factor load, so that the square of the factor load is transformed from 0 and 1 in the column direction, in order to achieve the purpose of simplifying the structure and clear meaning of the factor interpretation. There are many ways to rotate factors. This article is based on multiple experiments and adopts orthogonal rotation with maximum variance, which is an orthogonal rotation of factor loading, which can make the factor loading matrix after rotation maintain the orthogonality of each column while reducing the squared variance of the data.

(9) Extract the common factor loading table and establish new variables

After factor analysis, the common factors$F_1, F_2, \dots, F_p$are taken as new variables, and the load value α_{mp} is the p-th common factor variable value of the m-th observation object. Create a new data matrix as follows:

$$B=\begin{bmatrix}\alpha_{11} & \cdots & \alpha_{1p}\\ \vdots & \ddots & \vdots\\ \alpha_{m1} & \cdots & \alpha_{mp}\end{bmatrix}$$

(10) Establish Euclidean distance matrix D as follows:

$$D=\begin{bmatrix}d_{11} & \cdots & d_{1p}\\ \vdots & \ddots & \vdots\\ d_{p1} & \cdots & d_{pp}\end{bmatrix}$$

Among them, d_{ij}represent the distance betwteen i-th and j-th value, formula is as follows:

$$d_{ij} = \sqrt[2]{\left[\sum_{k=1}^{p}(\alpha_{ik} - \alpha_{jk})\right]^2}$$

(11) Set the number of clusters and calculate the sum of variance within the initial cluster as follows:

$$E = \sum_{i=1}^{k}\sum_{p\in C_i}\left(d_{P,C_i}\right)^2$$

K is the number of clusters, C_i is the i-th cluster, c_i is the center point (mean), and P is any observation object in the data set.

Assuming that the number of k clusters is set, in the first iteration, k observation objects are randomly selected as the initial clusters, and the value p is its mean value. The remaining (m-k) objects are given to the k objects according to their Euclidean distances A cluster satisfying $\min\{d_{i\in k,j\in(m-k)}\}$.

(12) Iterate until convergence, and generate clustering results

Finally, the cluster center value (mean value) is continuously updated with the newly added observation objects, and this process is continuously iterated until E converges, and a clustering scatter plot and table are generated for analysis.

3.3. Text Mining

Since the object of observation in this article is a special non-structured data such as keywords, the keywords themselves contain a lot of information, which is of research significance. Traditional statistical analysis methods do not have relevant research on unstructured data, so keyword information is not counted. After the classification results of the K-means cluster analysis model based on the factor analysis method are generated, it is difficult to explain the messy classification results one by one under the condition of a large number of observation objects. Therefore, the purpose of text mining in this article is to extract the most important key feature words for each category by establishing a TF-IDF matrix according to the classification results of the keywords after the model is established.

We take TFIDF method[6] on text mining on keywords ,steps are as follows:

(1) Establish a corpus and transform the corpus

After the keyword text is classified, new documents are built and put into the corpus. A series of operations such as lowercase, removing symbols, removing blanks, inserting stop words, and stemming the established corpus are converted into pure documents as needed. Among them, establish stopwords according to needs, such as "and", "to", etc., and delete non-characteristic words. English stemming uses the Porter stemming algorithm, also known as Porter Stemming, to strip the word suffixes.

(2) Build document term matrix (DocumentTerm Matrix, DTM), text data structure

Assuming that the corpus S has a total of D documents, and there are n_i independent and unordered words in the i-th document, then this model of text analysis based on the corpus S composed of independent and unordered words is called a bag of words model (BOW). Based on this premise, DTM counts the number of occurrences of words in the corpus in the document. The matrix is as follows:

$$DTM=\begin{bmatrix} \mathrm{n}_{1,1} & \mathrm{n}_{1,2} & \cdots & \mathrm{n}_{1,N} \\ \mathrm{n}_{2,1} & \mathrm{n}_{2,2} & \dots & \mathrm{n}_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{n}_{D,1} & \mathrm{n}_{D,1} & \dots & \mathrm{n}_{D,N} \end{bmatrix}$$

(3) Establish IF_IDF matrix and vectorize text data

The word frequency is represented by $\mathrm{tf}_{i,j}$ In the case of non-normalization, $\mathrm{tf}_{i,j}=\mathrm{n}_{i,j}$,which is the number of times the j-th word appears in the i-th document; in the case of normalization, $\mathrm{tf}_{i,j}=\frac{\mathrm{n}_{i,j}}{\sum_{j=1}^{N}\mathrm{n}_{i,j}}$ is the frequency of occurrence of the jth word in the ith document.

The frequency of reverse entries is represented by idf_i. It is defined as

$$\mathrm{idf}_j=\log_2\frac{|D|}{|\{d|t_j\in d\}|}$$

Where $|D|$ is the total number of documents, and $\left|\{d|t_j \in d\}\right|$ is the number of documents containing the jth word t_j in S.

4. Experiments

4.1. Data

The data selected in this article are the 1000 keywords data of a certain merchant on Alibaba International Station in January 2019. According to the specific situation of the Alibaba international platform and the real situation of business operations, this article selects six indicator variables of computer-side exposure, clicks, TOP10 business exposure, TOP10 business clicks, search popularity and seller competition.

We firstly perform data standardization processing on the original observation data.Based on the differences in the data units queried in this article, the standardization method is to use the Z-Score transformation in the data as follows:

$$\mathbf{z_i = (x_i - \bar{x})/\sqrt{S}}$$

4.2. Factor Analysis

(1) Hypothesis testing before factor analysis

In order to verify the feasibility and scientificity of using the factor analysis method to analyze the keyword data in this article, the KMO and Bartlett's tests were performed using the psych package in R. The results are shown in Table 1.

Table 1. Test Results

KMO Test	KMO	0.74
Bartlett's Test	Chi-Square	5920
	Degree of freedom	15
	P-value	0.000

According to the KMO test and Bartlett's Test, this data the set can be subjected to factor analysis. The average KMO value is 0.74, and the MSA values of the six variables are all greater than 0.6. at the significance level of α=0.05, the p value of zero is significantly smaller than α, the indicators of the observation objects are fed back from the backend of Alibaba website and have a certain degree of credibility. At the same time, the merchant has not invested in paid promotion for the time being, so the source of the keyword data depends on natural drainage, eliminating certain human manipulation factors and interference, so the result analysis of this data has a certain degree of credibility and reference significance.

(2) Factor number selection

Then we uses R statistical software to perform factor analysis on the standardized observation data. First, make a gravel diagram and select the appropriate number of factors, as shown in Figure 1.

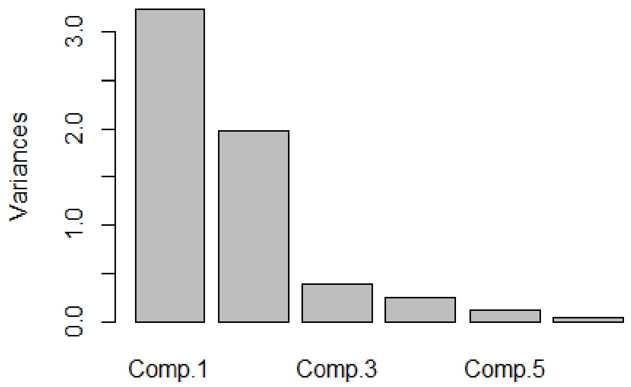

Figure 1. Strip lithograph

The bar gravel chart shows that the variance of the two factors is greater than 1, and the variance is less than 1 after the second factor. There is an obvious inflection point between the second factor and the third factor.

Then select the first two, and the total variance explanation table of the factor with the largest variance is shown in Table 2.

Table 2. Factor variance and its contribution

Factor	variance	Contribution rate	Cumulative contribution rate
Factor1	2.696	44.94%	44.94%
Factor2	2.517	41.95%	86.89%

Obviously, their cumulative variance explained the contribution rate reached 86.89%, thus two common factors can replace the original observation data variables to analyze keywords.

(3) Explanation of common factors after orthogonal rotation

Table 3. Rotated factor loading matrix

Index	factor1	factor2
Exposure	0.97638	-0.02225
clicks	0.97968	-0.01013
TOP10 business exposure	0.34486	0.82270
TOP10 business clicks	0.81209	0.50599
search popularity	0.02077	0.91434
seller competition	-0.06580	0.86460

As show in Table 3, factor1 can be represented as "word performance factor". the three variables of exposure, clicks, and TOP10 average clicks mainly reflect the performance of the exposure conversion effect of the business and the market on the keywords, When the TOP10 merchants have more clicks on this keyword, the better this keyword will perform in the market. At the same time, the remaining small and medium businesses can also use this keyword.

Factor 2 can be represented as "market popularity factor". When the three variables of Top10 average exposure, search index, and seller competition are higher, the market product demand and supply reflected behind this keyword is greater, and the overall popularity is greater. This enthusiasm includes two aspects: market demand enthusiasm and market competition enthusiasm. The market demand heat is expressed as a search index, and the variable reflecting the market competition heat is the seller competition.

(4) Cluster Analyses

Then, we extract the main common factor value Factor1 and Factor2 of Factor Analysis as new variable values for analyzing clusters. Then we select the number of clusters according to the needs. In this article, the actual test divides the clusters into 3, 4, 5, and 6 cases. According to the number and interpretation of the above main common factors, as well as the factor score map For the distribution, the cluster distribution class with k=4 that is the most suitable and has better explanatory significance is finally selected.

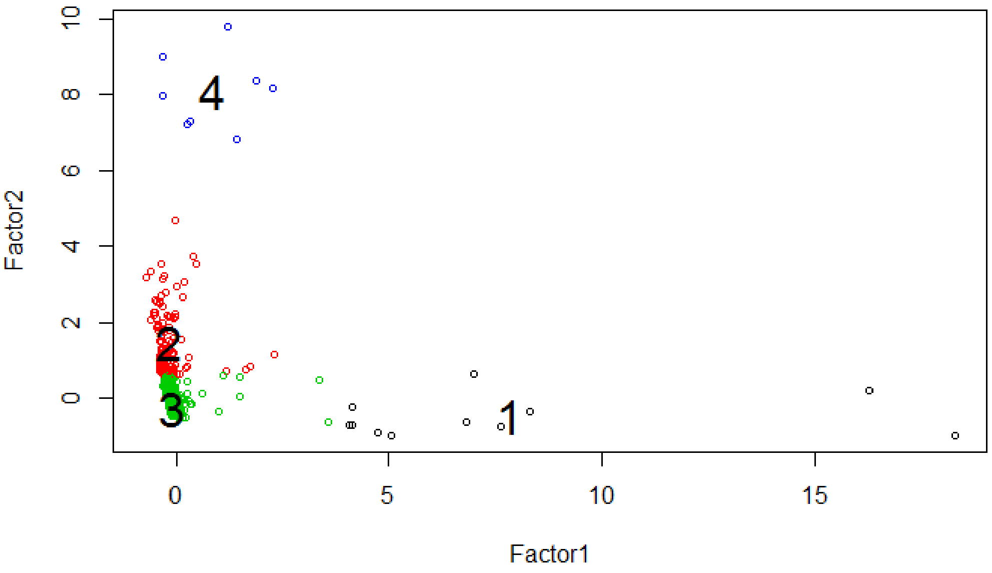

Figure 2. Scatterplot of K-means Analysis(K=4)

Table 4. Correspondence table of each cluster class and common factor

	Factor1 **(word performance factor)**	**Factor2** **(market popularity factor)**
C1	median high	low
C2	low	median
C3	median low	low
C4	low	high

As show above, among the keywords in C1,their word performance factor value is higher and the market popularity factor value is lower, which means that the product demand reflected behind such keywords has a certain potential, but the market popularity is low, and there may be keywords that are too long, which are the long-tail key of popular keywords Words, or a variety of situations such as a small number of related products in the field, the emergence of new market demands, and special vocabulary.

Among the keywords in C1, it indicates that there is a certain search interest and potential, but it is still not in the state of overheating in the market. Therefore, for conservative businesses, paid promotion You can focus on such keywords. However, there may be merchant-related products that are not in line with the keywords, which leads to poor word performance.

Therefore, merchants can optimize the related products of such keywords and make them related to products that are more in line with consumer needs to improve the performance of words.

Table 5. List of partial keywords of cluster

C1	C2	C3	C4
gold jewelry xuping jewelry jewelry set xuping jewelry 24k gold jewelry xuping pearl jewelry xuping gold necklace xuping jewelry stainless steel jewelry xuping pendant necklace xuping 18k gold jewelry xuping	online shopping india gold jewelry 2019 fashion jewelry pendant wedding rings pendant necklace earrings for women men ring tassel earrings gold ……	geometric ring aliababa com import jewelry from china white gold 925 alibaba express canada xuping jewelry sets joyas china jewelry wholesale ……	jewelry ring jewelry sets necklace online shopping free shipping accessories stainless steel jewelry online shopping

Compared with other clusters, the keyword performance and popularity in C3 are not good. It may be that the market is too small, and The consumer group searching for this keyword does not belong to the group with the exact willingness and ability to consume, or it is caused by consumers' missearch of such keywords. Therefore, the three clusters of keywords are temporarily not worth the business investment too much energy.

There are 8 keywords in the C4. The average on-word performance factor value is low and the market popularity factor value is high, indicating that this type of keyword is a hot spot in the market, and the products are homogenized, which tends to be a completely competitive market. Due to the low performance, one situation is that the product itself does not match the keywords. For example, the keywords of a certain brand that are not actually related to the product. The other situation is that the keyword competition is extremely fierce, which leads to businesses. The ranking is low. From the analysis of the merchant’s products, there may also be problems such as high prices, insufficient pictures, and failure to capture consumers’ attention and needs. For businesses with sufficient budgets, you can consider these keywords, but for ordinary small and medium-sized businesses, it is more suitable for one or two of them to be paid for promotion in the initial stage to attract traffic, and focus on those looking for such keywords. Long-tail keywords. For example, when “necklace” belongs to 4 clusters, consider the “14k gold necklace” with moderate market popularity.

4.3. Text Mining

After clustering analysis based on Factor analysis, the keywords are divided into four clusters.

We treat them as a corpus, and build a document word frequency matrix using TF-IDF. Each cluster is treated as a document,and keywords are the words in it. The value in the matrix is the TF-IDF value of each characteristic word in each cluster, which refers to the relative representativeness and relevance of this characteristic word in this cluster. The higher the value, the greater the representativeness of the feature word in this cluster. Matrix is as follows:

Table 6. Part of TF-IDF Matrix

V1	18k	24k	fashion	gold	jewelleri	jewelri	men	necklac	pearl
C1	0.42	1.00	0.42	1.66	0.42	0.00	0.42	0.00	0.42
C2	0.42	0.00	2.08	7.47	1.66	0.00	1.66	0.00	2.49
C3	9.55	10.00	6.23	38.60	26.98	0.00	21.58	0.00	14.53
C4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

From the above, we can find that the If-Idf value of relatively unrepresentative industry words such as jewelri and necklace is 0, which has significantly reduced the importance of such words. Therefore, the next step is to count the feature words with the highest tf-itf in each cluster. The following table selects the characteristic words with the if-itf value of each cluster arranged in descending order.

Table 7. List of some representative feature words of the cluster

C1		C2		C3		C4	
keywords	it-idf	keywords	it-idf	keywords	it-idf	keywords	it-idf
xupe	4.57	silver	24.00	silver	152.00	free	1.00
gold	1.66	earring	21.00	925	145.00	ship	1.00
24k	1.00	925	12.00	ring	102.10	onlin	0.83
		women	12.00	plata	58.00	shop	0.83
		sterl	11.00	sterl	53.00	ring	0.42
		ring	9.55	earring	51.00	accessori	0.42
		bracelet	6.00	anillo	42.00		
		gemstone	5.00	gold	38.60		
		plate	5.00	2019	37.00		
				stone	32.00		

It can be seen from the above table that the if-itf value can better analyze the topic feature words of each cluster category, to help businesses choose words and products.

Combining the cluster keyword table, cluster feature vocabulary and cluster product subject vocabulary content, the analysis is as follows after understanding the industry knowledge:

(1)Keywords in C1 are long-tail keywords that are mainly characterized by xuping, gold, and necklace. Among them, xuping is a top 10 jewelry supplier in the jewelry industry on the Alibaba international platform. It has been a jewelry supplier that has been established for decades and has occupied a certain market on the Alibaba international platform. Due to the establishment of the brand, it has sufficient visibility and paid promotion on the platform, so although keywords with their brand names do not have the huge traffic and search volume of normal industry keywords, they also have a certain degree of market popularity. Usually, visitors who accurately search for keywords with xuping brand characteristic words have exact consumption purposes and style preferences. In this type of words, we can also learn that xuping suppliers are mainly in gold jewelry (gold, 24k gold, 18k gold), There is a certain market share on pearls. Therefore, for merchants, setting up keywords with large and well-performing competitors' names in the store has the advantage of attracting traffic. This type of words can be used to analyze the market goals of competitors and use them for reference or avoidance.

(2)Keywords in C2 are the keyword category and product category that merchants can focus on that with potential market popularity. It can be seen that 925 silver jewelry is in great demand in the market, and it is mainly women's style, which is usually different from the domestic pursuit of Japanese and Korean styles. , European and American markets prefer exaggerated and large jewelry. At the same time, because the price of silver and gold is more expensive, and the weight of silver and gold is heavier, most foreign trade jewelry markets usually prefer electroplating, which is light, relatively low-priced and rich in styles. Therefore, for merchants, it is recommended to use this type of product as the main market direction, and select keywords with strong comprehensive popularity and performance capabilities in this cluster for promotion.

(3)The number of keywords in the C3 is huge, and the products are numerous, but it can be seen that the product elements reflected by the main representative feature words of the 3 clusters and the C2 have many similarities. Therefore, although the C3 of keywords are less popular than the C2, they can be used C3 are looking for a long tail market for C2 of products and opening up new areas. For example, “plata“ and “anillo“ mean the Spanish words "electroplating" and "jewelry", and merchants can consider researching and developing the Spanish market according to their own circumstances.

(4)C4 is a keyword category that has extremely high market enthusiasm but has a low performance level. From the perspective of representative characteristic words, there is no relevant industry market demand information, only big words such as free shipping, accessories, and online shopping. On the one hand, these words have no promotional meaning, but on the other hand, we can learn most of the customer assistance needs. First, suppliers generally need to provide sample sheets, and the jewelry industry usually adopts the FOB foreign trade method, that is, the customer’s freight. As the freight is more expensive in foreign trade, free shipping for sample orders or small sample orders is a way to attract price-sensitive consumer groups. Second, traditional B2B websites do not have the custom of consumers to place orders independently. Due to the particularity of wholesale products, orders can be placed only after consultation and customization with suppliers, and the time period is long. For fancy products, direct online ordering can facilitate customers and save time. Therefore, for merchants, they can gain insight into the psychological needs of customers from such words, portray customer group portraits, optimize services from them, or choose keywords with this characteristic word for promotion.

5. Conclusion

In this paper, a factor analysis of 1000 keyword data indicators of a certain merchant on Alibaba website, extracts two common factors of "word performance factor" and "market popularity", eliminates the correlation of variables, and uses this common factor to perform K-means aggregation. Class analysis, the keywords are divided into four clusters, and finally the feature words of these four clusters are extracted by text mining, and relevant suggestions are provided for the optimization of products and keywords.

Data-based operation of E-commerce platforms is the mainstream trend of current operations, but most of the analysis rests on the observation and intuitive analysis of simple data indicators. At the same time, because E-commerce data involves a lot of unstructured data, the psychological needs of people reflected behind it are usually very different, and it is difficult to find hidden laws. As the amount of data increases, the use of traditional and modern statistical methods for research has certain practical significance.

Among them, this article still has the following shortcomings:

(1) Theoretically, relevant multivariate statistical method theory and text mining theory are good at learning, and data analysis and processing still need to be improved.

(2) In terms of data sources, this article only uses one month's data, and there is no time-based comparative analysis. At the same time, due to other reasons such as difficulty in data acquisition and lack of data, keywords are not included in data such as the number of orders, the number of feedbacks, the relevance of related products, and information, which may have a certain degree of rigor.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China (grant numbers 2018A030313849) and the Soft Science Research Program of Guangdong Province, China (grant numbers 2016A070705028).

References

[1] Matalas N C, Reiher B J. Some comments on the use of factor analyses[J]. Water resources research, 1967, 3(1): 213-223.

[2] Rujasiri P, Chomtee B. Comparison of clustering techniques for cluster analysis[J]. Agriculture and Natural Resources, 2009, 43(2): 378-388.

[3] Hotho A, Nuremberg A, Paaß G. A brief survey of text mining[C]//Ldv Forum. 2005, 20(1): 19-62.

[4] Tan A H. Text mining: The state of the art and the challenges[C]//Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. sn, 1999, 8: 65-70.

[5] Vaidya N, Khachane A R. Recommender systems-the need of the ecommerce ERA[C]//2017 International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2017: 100-104.

[6] Aizawa A. An information-theoretic perspective of Tf–Idf measures[J]. Information Processing & Management, 2003, 39(1): 45-65.