Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 1444755, 10 pages
https://doi.org/10.1155/2021/1444755

*Research Article*

# Social Network Big Data Hierarchical High-Quality Node Mining

**Dongning Jia**[1,2] **Bo Yin**[1,2] **and Xianqing Huang**[2]

[1]*Ocean University of China, Qingdao Shandong 266100, China*
[2]*Pilot National Laboratory for Marine Science and Technology (Qingdao), Qingdao Shandong 266237, China*

Correspondence should be addressed to Bo Yin; ybfirst@ouc.edu.cn

Received 9 April 2021; Revised 26 April 2021; Accepted 7 May 2021; Published 18 May 2021

Academic Editor: Shan Zhong



Compared with the conventional network data analysis, the data analysis based on social network has a very clear object of analysis, various forms of analysis, and more methods and contents of analysis. If the conventional analysis methods are applied to social network data analysis, we will find that the analysis results do not reach our expected results. The results of the above studies are usually based on statistical methods and machine learning methods, but some systems use other methods, such as self-organizing self-learning mechanisms and concept retrieval. With regard to the current data analysis methods, data models, and social network data, this paper conducts a series of researches from data acquisition, data cleaning and processing, data model application and optimization of the model in the process of application, and how the formed data analysis results can be used for managers to make decisions. In this paper, the number of customer evaluations, the time of evaluation, the frequency of evaluation, and the score of evaluation are clustered and analyzed, and finally, the results obtained by the two clustering methods applied in the analysis process are compared to build a customer grading system. The analysis results can be used to maintain the current Amazon purchase customers in a hierarchical manner, and the most valuable customers need to be given key attention, combining social network big data with micro marketing to improve Amazon's sales performance and influence, developing from the original single shopping mall model to a comprehensive e-commerce platform, and cultivating their own customer base.

## 1. Introduction

With the rapid development of computer technology, the Internet electronic information resources play an indispensable role in everyone's life and work, and people interact with each other more and more through the Internet [1]. Social networking is a change from online social networking, the predecessor of online social networking is email, and online social networking pushes BBS one step forward [2]. Email and BBS are upgraded versions of social tools are instant messaging and blogs, which have significantly improved in terms of transmission speed and parallel processing; blogs have begun to reflect sociological and psychological theories. In recent years, the development of social networks has been remarkable [3–5]. At present, more than half of Chinese Internet users communicate and share information through social networks, which have become the Web 2.0 business with the largest communication impact, the widest coverage of users, and the highest commercial value. Social networks have become a part of people's lives and play a very important role in people's lives, having an undervalued impact on people's access to information, thinking, and life [6–8].

Social networks have become a window for people to get information, show themselves, and market and promote their functions. As of December 2012, the total number of Internet users is 564 million, and 309 million are microblog users, in which tens of millions of Internet users are active every day, posting microblogs to share what is new, etc. [9, 10]. Social networks have three major advantages: (1) higher user viscosity. According to the survey data, an average Internet user spends about 17 to 20 minutes a day on social networks. (2) Low maintenance cost. The number of editors needed for Web 1.0 portal websites exceeds the number of employees for Web 2.0 websites. (3) Information preparation. Social network users are asked to fill in real, detailed personal information. This facilitates developers to conduct data analysis as well as business applications. Web data is growing rapidly, and having access to this data on websites

can be of great help in analyzing topical social issues and trends and also can be of great help and impact on social network operations [11]. Social networks concentrate the youngest and most active web users, so they are also the most densely populated places for online speech, and mining their information resources can obtain a lot of valuable raw data for relevant user analysis and decision-making [12]. Data mining is a thriving disciplinary frontier concerning data and information systems and applications, the result of a multidisciplinary field and drawing from several disciplines, and a natural evolution from information technology [13–15]. It is able to precisely mine data hidden knowledge from massive amounts of data, and data mining is applied in any type of information repository and transient data. In data mining applications, the most basic forms of data are database data, data warehouse data, and transactional data [16, 17]. Data mining has now become one of the most cutting-edge research directions in the field of information decision-making and databases internationally, mainly because of its great potential for business prospects and has attracted wide attention from industry and academia [18, 19]. Data mining can be applied to financial data analysis, retailing, telecommunications, biological data analysis, and other scientific applications. In this paper, data mining is applied to data mining in social networks. The huge amount of data in the Internet and the data will grow rapidly, provides a good basis for data mining [20, 21].

In recent years, due to the limitations of search engine query information, many scholars improve the efficiency of retrieval is by eliminating the near mirror pages, the research of algorithms for near text detection. Researchers at the University of Arizona, USA, found similar documents existing in a large file system by employing a method that computes the degree of overlap of documents. The results of the studies presented above are generally based on statistical methods and machine learning methods, but some systems use other methods such as self-organizing self-learning mechanisms and concept retrieval. The diversity of web data forms brings new challenges to data mining, and it is the main effort and development direction in the future to combine clustering, support vector machines, neural networks, etc. with various database techniques in its large amount of graphical and complex spatiotemporal data, and to investigate data mining in new databases. In this paper, we study the application of data mining in social networks. The social network object studied in this paper is Sina Weibo. The data of popular topics and participating users in Sina Weibo are extracted by python, and then, the clustering algorithm and collaborative filtering algorithm of data mining are used to analyze and recommend the user data. The main work done is as follows.

(1) It briefly introduces the basis of web data extraction and methods of extracting data, then introduces data mining techniques and clustering algorithms and hierarchical methods in clustering methods

(2) This article describes in detail the extraction of user data from social networks in Python, using two methods: the Weibo API interface and the simulated browser login method

(3) Clustering of the extracted trending topics and topic recommendations to users based on their participation in the topics

## 2. Data Mining Techniques and Tools

*2.1. Data Mining Tools.* The Oracle database management system is used as the tool for data storage, which is a relational database and one of the current mainstream database management systems. Compared with other database management systems, it is easy to install, easy to use, and easy to understand interface. The simple and easy-to-use SQL statements are not necessary for beginners to consider how to process data with complex algorithms, and they can get the desired data with suitable SQL statements. Currently, there are two main categories of popular data mining tools: one is a data mining tool specifically for a specific industry or field, and the other is a general mining tool with a wide range of applications and more situations. Data mining tools for a specific field or industry, the tool itself has some optimization or preprocessing work on the data. In this paper, we choose the more general data mining tools, and the main general and free data mining tools are WEKA from New Zealand, R language with powerful plotting performance and statistical analysis, and the Clementine system from SPSS.

Among these three data mining tools, Clementine is the industry's leading data mining platform that can apply complex data mining algorithms and machine learning techniques to process data to help companies uncover the value behind transaction data, with the strongest ease of use and the most beautiful interface. The main algorithms applied in this data mining process are K-Means clustering algorithm, two-step clustering algorithm, RFM structural model, etc. The analysts using Clementine do not need to spend much time on the algorithms themselves, but only need to apply a reasonable data model to observe whether the data analysis results are consistent with the business needs.

*2.2. Clementine Has Three Main Features*

(1) *Beautiful Interface and Visualization of Data Analysis Process*. As a general-purpose data mining tool, the beautiful and visualized operation interface is one of its great advantages. Users only need to select the nodes they want in the modules of source, record option, field option, graph, modeling, output, and export and connect several nodes with lines to complete the creation of a basic model.

(2) *Powerful Data Processing Capability*. Even a novice who has no contact with data mining algorithms can use Clementine to build data models by reviewing some theoretical knowledge of algorithms and knowing the application scenarios of algorithms. The simple model in Clementine does not require any parameter setting; just input the preprocessed

data in the data source and then run and output the results. The expert model requires the analysts to select the established data model according to the business requirements and customize some model analysis parameters according to the actual needs in the analysis process. The setting of parameters mainly relies on the experience value of business personnel in the practice process to set and through the adjustment of parameters to continuously optimize the data model and build a model that meets the actual needs of the enterprise, in order to achieve the purpose of outputting ideal analysis results, providing data support for the development of business and providing a source of power for the advancement of the enterprise.

(3) *Follow the Standard Data Mining Process of CRISP.DM*. Unlike the traditional mining process built on the technical level, Clementine can effectively control the entire mining process, equating the data mining process with a business analysis process and making business purposes and business needs the latest goal of data mining. A complete data analysis process includes six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, and different stages correspond to different project management requirements.

Following the CRISP.DM process enhances the user's understanding of the business model and achieves a perfect alignment between business requirements and the data model. More than half of the data mining platforms in the industry follow the CRISP.DM standard, which has gradually become the industry standard. As shown in Figure 1. Data source is an efficient hierarchical clustering algorithm, which can deal with large data clusters simultaneously. It is mainly used to divide the dataset to be processed into several small datasets to complete the quasi-clustering process of data.

*2.3. Core Ideas and Algorithms.* Clustering is the process of dividing a collection of physical or abstract objects into multiple classes composed of similar objects. The cluster generated by clustering is a collection of data objects that are similar to each other in the same cluster and different from the objects in other clusters. As the saying goes, "things come together in groups, people come together in clusters," and classification problems can be found everywhere in the social and natural sciences. Cluster analysis, also known as cluster analysis, is a statistical analysis method to study the problem of classifying samples or indicators.

Although clustering and classification have some commonality in that they are both aimed at classifying data from a given dataset, the difference is that clustering has a learning process, while classification simply divides the existing dataset into different categories according to the specified needs.

The number of categories for clustering can be specified or obtained automatically based on algorithms, and it is up to the analyst to decide which approach to use based on different business needs. The current development of clustering technology is promising and covers a wide range of fields, including statistics, data mining, machine learning, and marketing. Clustering analysis, as an important branch of data mining, is a popular research topic.

K-Means clustering is a classical bottom-up clustering method. It is characterized by a very fast clustering speed compared to other clustering algorithms even with a very large amount of data and a relatively simple execution process. However, the disadvantage is that the K-value must be specified before clustering, and the analyst usually does not know how many classes should be clustered in the actual application. This requires us to select different K-values for multiple clustering, which is solved by selecting the best clusters.

K-Means clustering is used to divide the existing dataset with $n$ samples into clusters with high similarity. The clustering process can be roughly described as follows: $k$ samples are randomly selected as clustering centers, the remaining ones calculate the distance from each sample to each center, then the object is assigned to the center closest to it, the center of each new cluster is recalculated, and the above process is repeated until the convergence condition is satisfied. The usual convergence condition is that the centroids no longer change or a certain number of iterations are reached. The squared error criterion is generally used, and the relevant definition is as follows:

$$E(n) = \sum_{i=1}^{k} \sum_{P \in Q}^{n} (\lambda t)^{n+1} e^{\lambda t}. \tag{1}$$

$E$ is the sum of the squared errors of all objects in the database, $P$ is the point in space, and im is the average of cluster ic. The objective function is to make the generated clusters as compact and independent as possible, and the distance measure is the Euclidean distance, but it is possible to use other distance measures if desired.

$$P = e^{-2\pi T^2} = e^{-2G}. \tag{2}$$

The second-order clustering algorithm is a common hierarchical clustering algorithm, mainly used in the two major fields of cross-sectional and data mining of multivariate statistics. Compared with the K-Means clustering algorithm, the biggest advantage of the two-step clustering algorithm is that it can determine the K-value of clusters by itself, and there is no need to determine the best clustering class by different K-values, which eliminates many steps of manual discrimination.

But the disadvantage lies in the fact that because the categories are determined by the algorithm itself, without combining the actual needs, the clustering results often fail to achieve the expected results. The choice of the two algorithms needs to be made by analysts according to different needs and application scenarios. The following is a brief description of the algorithmic process of the two-step clustering algorithm for your reference.

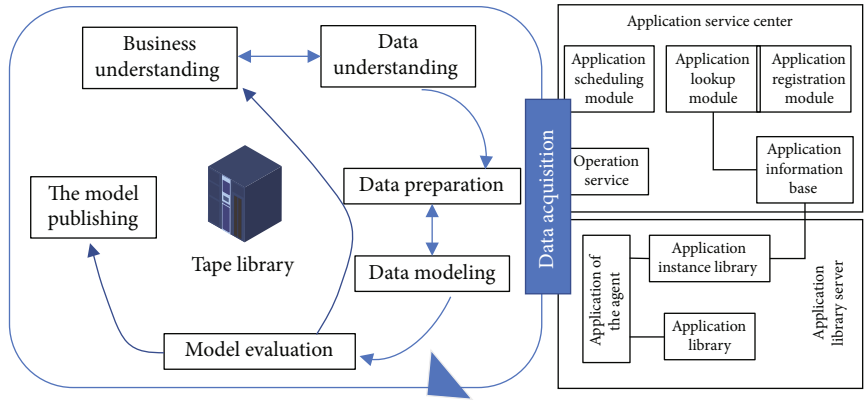


Figure 1: The 6 stages of CRISP.DM.

Input: the dataset with $n$ samples and the number of clusters $k$.
Output: $k$ clusters and minimizes the squared error criterion.
Steps.
(1) Arbitrarily select $k$ samples as initial cluster centers
(2) Repeat the above steps
(3) Calculate the mean of the samples in the cluster and reassign the values of the samples to the most similar cluster
(4) Updating the cluster means, i.e., calculating the mean of the samples in each cluster
(5) Until no more changes occur

Algorithm 1:

*Step 1.* Quasi-clustering process. The main work of this step is to apply a comprehensive hierarchical clustering algorithm BIRCH (Balance Iterative Reducing and Clustering using Hierarchies), which is an efficient hierarchical clustering algorithm capable of handling large-scale data clusters at one time. It is mainly used to divide the dataset to be processed into many small datasets and complete the quasi-clustering process of the data.

*Step 2.* Secondary clustering process. This step uses the log-likelihood function to calculate the distance between samples and implements secondary clustering based on the small dataset preprocessed in the previous step. The recursive algorithm merges the input subclusters using the "hierarchical coalescence method" until the last cluster includes all the datasets. The second-order clustering algorithm is integrated in Clementine, and the user only needs to select the appropriate model node to use. The algorithm uses probability-based distances as a measure function:

$$d(k,n) = (\lambda t)^{n+1} e^{\lambda t} + \xi_i, \tag{3}$$

$$\zeta_n = H_{N,v}\left(\sum_{i=1}^{k}\sum_{P\in Q}^{n} (\lambda t)^{n+1} e^{\lambda t} + Q_n\right), \tag{4}$$

$$\tilde{E}(n) = \sum_{i=1}^{I_k}\sum_{P\in Q}^{n} (\lambda t)^{n+1} e^{\lambda t} \log_n \frac{N_v}{N_n}, \tag{5}$$

where KA is the number of input domain ranges, KB is the number of input domain symbols, KL is the number of categories of the $k$th symbolic domain of the input, VN is the number of records in cluster $V$, VKLN is the number of records belonging to the $k$th symbolic domain of the $l$th category in cluster $V$, 2k is the estimated deviation of the $k$th continuous variable from all records, and 2vk is the estimated deviation of the $k$th continuous variable from the $V$th cluster.

RFM model is a means for enterprises to achieve database precision marketing, by setting three core indicators to carry out customer segmentation. In most cases, management decision-makers tend to focus only on the single dimension of sales amount and believe that those with high spending amount are important value customers and should be given key marketing. However, there are many customers with large single purchase amounts, but the frequency of sales is very low, and some even come into the store once in a few years; these customers can basically be judged as lost customers. For different customer groups, we should use differentiated marketing strategies, if the marketing of lost customers and important value customers using the same marketing approach will result in a waste of corporate resources. Currently, this kind of rough marketing model is gradually improving; the United States after a lot of research data shows that the following three factors constitute the key factors can be customer segmentation. They are freshness (the latest consumption time), consumption frequency (the number of times of consumption over a period of time), and consumption amount (the total amount of consumption over a period of time).

(1) Freshness (last consumption time)

Freshness refers to the most recent consumption time of the customer, which category of goods the customer has recently purchased, whether the category has changed compared to the previous purchase of goods, what is the reason for the recent change in consumption habits, if the change from the purchase of daily goods to frequent purchase of baby products can be determined that the customer has a baby at home recently, and the change to the purchase of home building materials can be determined that the customer recently purchased a new house has the demand for decoration. For customers in different stages of life, merchants need to dig out customers' recent needs from purchase data and push the latest information and promotion information of products that customers are concerned about according to their consumption needs. Send targeted information through data analysis, so that the customer's acceptance of the information will be greatly enhanced. Do not make customers feel that you are sending information only to sell products; you need to reduce the customer's aversion to promotional information.

Once the customer becomes disgusted with the merchant, he or she will not go through the merchant's catalog book, the cell phone sets up SMS blocking, and the contact channel between the merchant and the customer will be cut off. If you send relevant information after three months or even six months, you will also miss the best marketing time period, and customer acceptance will be reduced. The actual situation shows that the acceptance of customers with recent consumption time is inevitably higher than that of customers with long consumption time. Through data analysis to dig out the information behind the customer, form effective interaction with the customer, repeated contact, so that the customer always feel the merchant's concern for him, enhance the customer's goodwill towards the merchant, and enhance the customer experience. Then, the customer's arrival rate will certainly be improved, and with the merchant's customized marketing strategy, the transaction rate will be significantly increased. By increasing the frequency of contact with the customer, the customer's arrival rate will be improved, thus increasing the closing rate.

(2) Consumption frequency

This is the actual number of times a customer has visited the store over a period of time. These customers can be considered separately in the actual analysis process. The characteristics of these customers are that they love to take advantage of small bargains and will not come to the store without the merchant's activities and will only choose cheap goods, which will not form related sales and upselling. However, most of the customers with high consumption frequency are customers who often buy with high satisfaction to the merchant. The merchant's brand influence and service quality are what keep customers spending at high frequencies and loyalty. Customers have a limited time to buy, and increasing the frequency of customers' consumption will, to a certain extent, grab market share from competitors and increase their own market share.

(3) Spending amount

Spending is usually the most important metric for management and decision-makers, as it directly affects revenue and profitability. According to the traditional "two to eight" rule, 20% of customers contribute 80% of the company's revenue. It can be seen that how to distinguish the important value customers is very important to enterprises; in the case of limited resources and marketing expenses, you can give priority to 20% of the important value customers for precision marketing, because these customers are the main contributors to corporate revenue, as shown in Figure 2.

When classifying customers, we need to consider the above three core indicators and set different weights for the three indicators according to the actual situation of our own enterprise when using the RFM model in Clementine, and the weights can be adjusted by referring to the experienced values of marketing personnel.

## 3. Experimental Design

Freshness (last consumption time), consumption frequency, and consumption amount are three very important indicators in customer value assessment and customer relationship management, which have a strong guiding meaning for our daily marketing activities, customer maintenance, and enhancing customer loyalty. The freshness is the most important of them. Using RFM model in Clementine, we can transform these three variables as long as the analyzed dataset has three indicators: unique customer identification, last time, and spending amount. Each of these variables is given a different weight according to the actual situation, and then, the customer value is evaluated based on the RFM score. We classify customers into $5^*5^*5 = 125$ categories, as shown in Figure 3, and analyze their data to customize marketing strategies.

Referring to the amazon-meta dataset of Amazon, the amount of review data shown in Dataset statistics is 7781990, which is the total sum of review statistics table of products, including 42919 undownloaded data, 145827 invalid data, the total number of reviews (count), the average value of evaluation scores (AVG_rating), the average value of the number of votes (AVG_votes), the average value of the number of customer reviews (AVG_votes), and the average value of the number of customer reviews (AVG_votes). For (AVG_helpful), the total number of data is 1555171.

Companies must carry out marketing activities before the consumer habits and buying behavior data mining, to understand the customer in order to develop marketing strategies. Traditional market research methods have been tested in practice and have great limitations. The customer base is narrow and unrepresentative. Data mining makes up for this shortcoming. Big data mining based on social network evaluation data can be carried out from the following aspects.

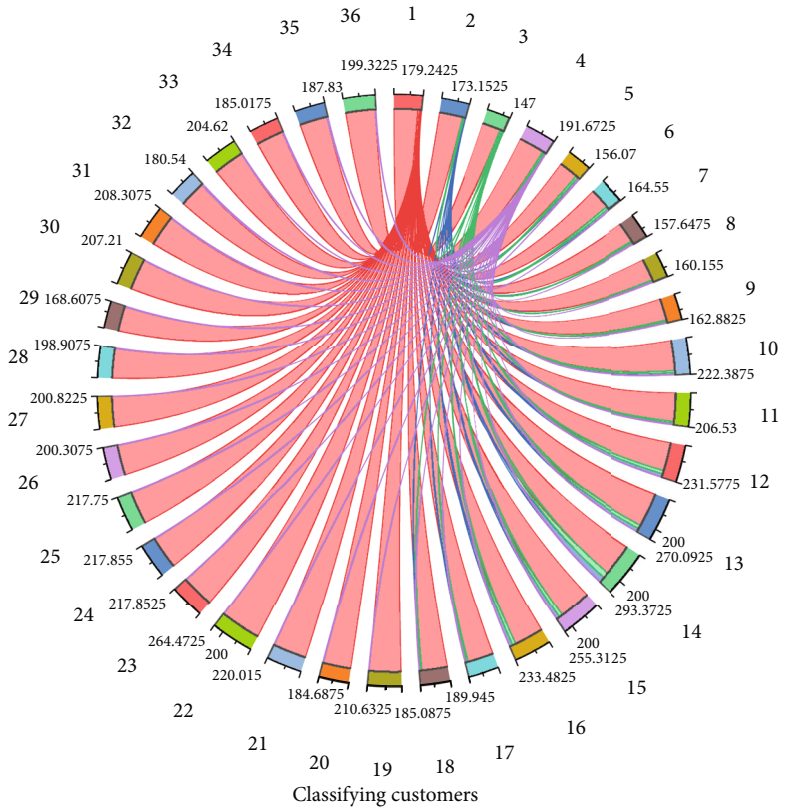


Figure 2: RFM model.

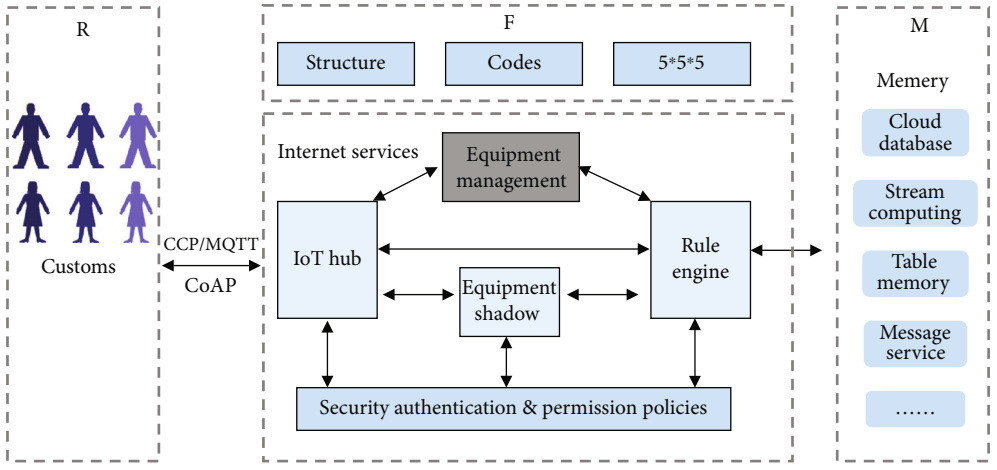


Figure 3: RFM coding structure.

(1) Preprocessing of data and clustering using different clustering algorithms

(2) Analysis of the differences between the customer groups formed by the clustering results

(3) From the differentiation of data, find out the characteristics of customer reviews and the interesting behavior of customers in purchasing goods and conduct a series of analysis and summary, so as to provide users and enterprises with more targeted and valuable information

## 4. Results and Analysis

*4.1. Quality Node Mining. RFM model*: R (recency) indicates the most recent purchase time of the customer, F (frequency) indicates the frequency of the customer's consumption in a period of time, and M (monetary) indicates the amount of the customer's consumption in a period of time. The original fields analyzed are three: customer ID (unique identification of customers), consumption time (date format), and consumption amount, which are processed by data mining software to obtain RFM scores by weighting the three indicators,

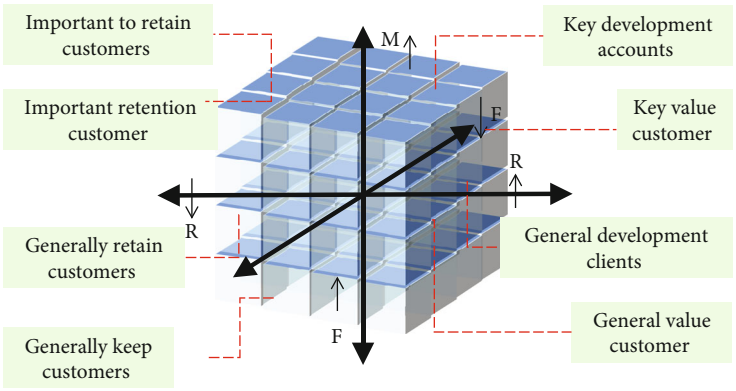


Figure 4: Traditional RFM model and Amazon's RFM model.

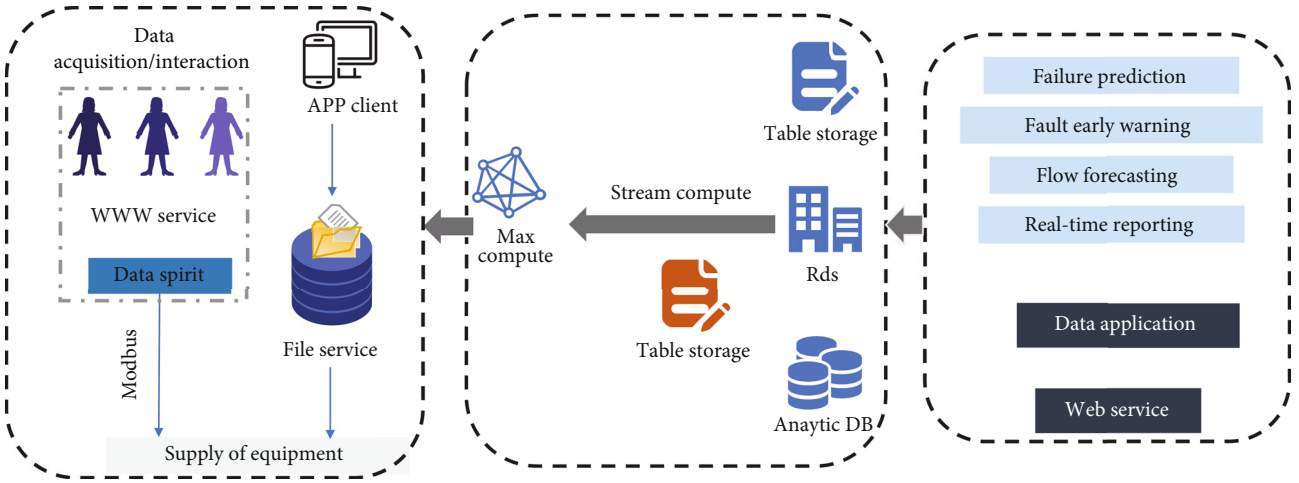


Figure 5: Customer segmentation based on RFM model.

completing customer class classification, and sorting the obtained RFM scores to implement accurate database marketing for different customer classes. Figure 4 shows the transformation of traditional RFM model into Amazon's RFM model.

RFM model is a data processing method, and in this paper, we found that the traditional RFM model classifies to form 125 customer categories, which is too segmented customers and difficult to operate in practical application. We need to adopt a new clustering method for customer segmentation based on RFM model, and the characteristics and behavioral habits of segmented customer groups need to reach our expected clustering under over.

We continue to use Clementine to cluster the three fields of R, F, and M. The clustering analysis mainly uses the following: K-Means and two-step algorithms. Before clustering, we found that directly using the three variables of R, F, and M for clustering, the measurement scales between the variables are very different and cannot achieve the expected effect, so we need to have descaling. In addition, the weights of R, F, and M should be different because of the differences in the importance of these three indicators in the displayed evaluation system. By comparing the functionality of a topic, the following statistics can be obtained for the number of times a given user has posted each topic. Here, we do not use a weighting method for the three variables (in practice, an expert or a corresponding marketer is needed to evaluate them), and through continuous testing and evaluation, we can choose the specific clustering method and the number of clusters, and also, we need to compare which of the two algorithms is more desirable. The RFM model-based customer segmentation quadrant is shown in Figure 5.

Drawing a tree graph starts by constructing a specific function whose return value is the overall height of the given cluster. It is important to know the overall height of the clusters when determining the overall height of the graph and placing the different node positions. If the cluster is a leaf node, its height is 1; otherwise, the height is the sum of all branch heights. In addition, the overall error of the root node must be known. That is because the length of the lines is adjusted accordingly to the error of each node, so a scaling factor is generated based on the total error value. The depth of error of a node is equal to the maximum possible error of each branch to which it belongs.

The drawdendrogram function creates an image with a fixed width and a height of 20 pixels for each of the final generated clusters. The scaling factor is obtained by

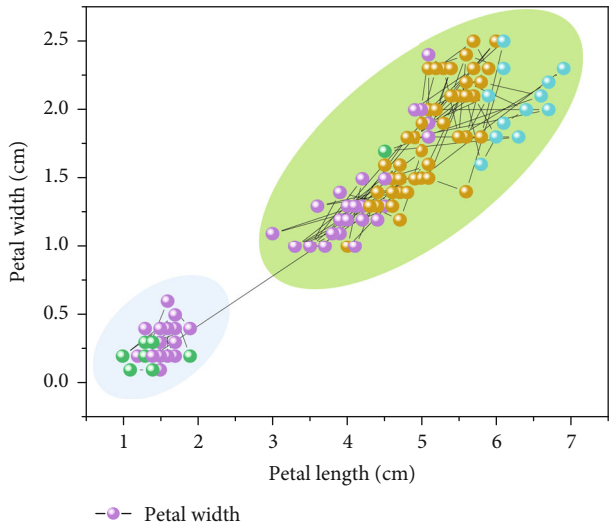


Figure 6: Clustering similarity.

dividing the fixed width by the total depth value. This function creates the corresponding draw object for the image and then calls the drawnode function at the position of the root node and places the node at the left center of the entire image. When the drawnode function accepts a cluster and its position as input parameters, the function takes the height of the child nodes and calculates their exact position in the picture and then connects them with lines. The diversity of Web data forms has brought new challenges to data mining. Combining clustering, support vector machine, neural network, and other database technologies, in the massive graphical and complex spatiotemporal data, is the main effort and development direction of future data mining. There are two horizontal lines and one longer vertical line. The length of the horizontal line is determined by the error condition in the clustering. The length of the line is related to the clustering result, the more different the two clusters are combined, the longer the line is drawn, and if the two clusters are more similar, the shorter the line is drawn, as shown in Figure 6.

According to the result of the above figure, there are eight levels. First, topic 4 and topic 5 are clustered, assumed to be cluster 1, and topic "Little Times 3" and topic "120 years of Wudao" are clustered, assumed to be cluster 2. At this point, topic 1 and topic 2 are clustered to form cluster 3 and so on, and finally, all the clusters are merged into one class. Finally, all clusters are combined into one class. And from Figure 7, we can see that after topic 3 and other topics are clustered into one class, the topic "Where did flowers go" is much farther away from the aggregation point, so we can conclude that the classes formed by other topics are closer to the aggregated class. There are 28 topics in the above figure, and 31 topics are extracted, so we can conclude that the remaining 3 topics are outliers and cannot be clustered with other classes.

The function will return a value between -1 and 1, and the number of evaluations must be greater than or equal to 0, so the final value returned is a value between 0 and 1. A value of

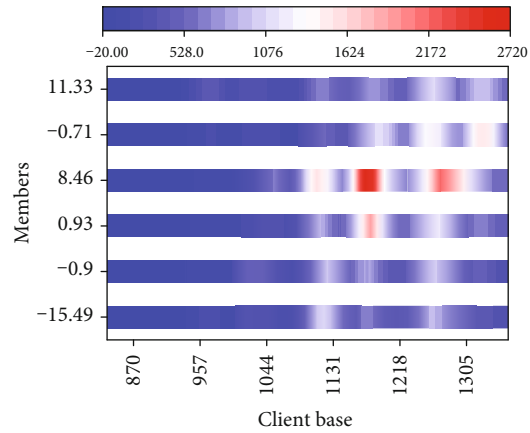


Figure 7: Clustering algorithm viewer.

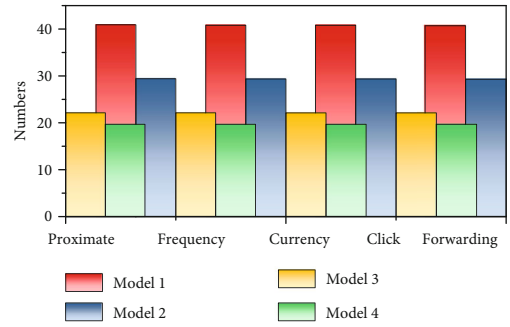


Figure 8: Differences in metrics between categories.

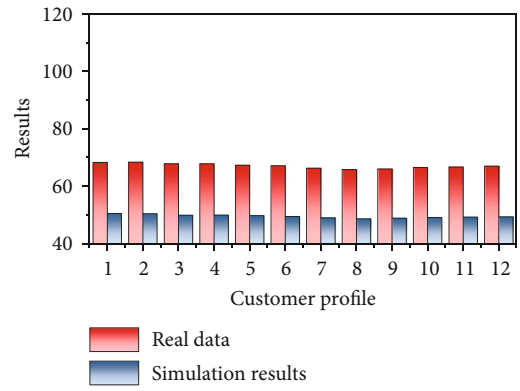


Figure 9: Distribution of the number of customers by category.

1 means that each user has the same number of posts for both topics. With the function of comparing topics, the following statistics can be made on the number of times a given user has posted on each topic, and if the number of times is higher, it means that the user has a higher interest in the topic, and the interest can be expressed by scoring the evaluation, as shown in Figure 8. However, there may be a user who is particularly enthusiastic about a topic, which will affect the final evaluation result. In order to solve this problem, the authors propose a weighted evaluation value to score the topic, and then, the ranking of the topic evaluation will be obtained. Finally, the similarity obtained is multiplied by the

evaluation value obtained for each topic to reach the final result. The program obtains an ordered list of topics that are of similar interest to a given person and that have not been tweeted by that user. The specified person is randomly generated in the program.

Since the 31 popular topics and the users involved in each topic were extracted in the previous section, it is possible to recommend topics with high similarity to users based on this data, which is also known as collaborative filtering technique. Collaborative filtering algorithms usually delineate a large group of people and then search for them to find a group of people with similar tastes. The algorithm focuses on the content that the group likes and ranks the content together to get a list of recommendations, as shown in Figure 9. After collecting the data, the number of tweets posted by each user on each topic is determined, and then, the similarity between topics is determined based on the number of tweets posted by each user on each topic, which can be used to calculate the similarity between each topic and other topics.

## 5. Conclusion

Discovering the hidden value in the data is the main purpose of data mining applications. In recent years, the rise of e-commerce networks, the development of people's online shopping habits, the development of social networks is remarkable, and people's interactive behavior on the Internet has formed a large amount of social network data, the traditional statistical analysis methods have limitations for dealing with big data, and the hidden patterns and values in the data can be discovered with the method of big data mining which is a new application trend. In this paper, we combine traditional statistical analysis methods and data mining algorithms to analyze Amazon's customer evaluation data. Around this work, this paper finds that the customer evaluation data of Amazon is analyzed by cluster analysis method, and two different clustering algorithms are applied to classify customer groups using four evaluation indexes of customers, and the results of clustering are compared and evaluated, and the optimal clustering results are selected to make a customer portrait for Amazon's evaluation customers. In addition, this paper applied the clustering algorithm based on RFM model to classify the customers of Amazon customer evaluation data, observe the proportion of important value customers and whether there are missing growth customers, and provide relevant suggestions for customer development and maintenance.

## Data Availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

## Consent

Informed consent was obtained from all individual participants included in the study references.

## Conflicts of Interest

We declare that there is no conflict of interest.

## References

[1] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: a survey," *IEEE Transactions on Big Data*, vol. 6, no. 1, pp. 3–28, 2020.

[2] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, 2018.

[3] E. Kross, P. Verduyn, M. Boyer et al., "Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook," *Emotion*, vol. 19, no. 1, pp. 97–107, 2019.

[4] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019.

[5] M. Moessner, J. Feldhege, M. Wolf, and S. Bauer, "Analyzing big data in social media: text and network analyses of an eating disorder forum," *International Journal of Eating Disorders*, vol. 51, no. 7, pp. 656–667, 2018.

[6] M. Balaanand, N. Karthikeyan, and S. Karthik, "Designing a framework for communal software: based on the assessment using relation modelling," *International Journal of Parallel Programming*, vol. 48, no. 2, pp. 329–343, 2020.

[7] F. Ali, S. El-Sappagh, S. R. Islam et al., "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Generation Computer Systems*, vol. 114, pp. 23–43, 2021.

[8] J. R. Ragini, P. R. Anand, and V. Bhaskar, "Big data analytics for disaster response and recovery through sentiment analysis," *International Journal of Information Management*, vol. 42, pp. 13–24, 2018.

[9] J. Kim and M. Hastak, "Social network analysis," *International Journal of Information Management*, vol. 38, no. 1, pp. 86–96, 2018.

[10] M. Klöwer, D. Hopkins, M. Allen, and J. Higham, "An analysis of ways to decarbonize conference travel after COVID-19," *Nature*, vol. 583, no. 7816, pp. 356–359, 2020.

[11] L. Liao, X. He, H. Zhang, and T.-S. Chua, "Attributed social network embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2257–2270, 2018.

[12] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services," *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020.

[13] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: past, present and future," *Information Systems Frontiers*, vol. 20, no. 3, pp. 531–558, 2018.

[14] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852–1872, 2018.

[15] L. Bode and E. K. Vraga, "See something, say something: correction of global health misinformation on social media," *Health Communication*, vol. 33, no. 9, pp. 1131–1140, 2018.

[16] M. E. J. Newman, "Network structure from rich but noisy data," *Nature Physics*, vol. 14, no. 6, pp. 542–545, 2018.

[17] Y. Dong, Q. Zha, H. Zhang et al., "Consensus reaching in social network group decision making: research paradigms and challenges," *Knowledge Based Systems*, vol. 162, pp. 3–13, 2018.

[18] X. Zheng, J. Han, and A. Sun, "A survey of location prediction on Twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1652–1671, 2018.

[19] C. Brell, C. Dustmann, and I. Preston, "The labor market integration of refugee migrants in high-income countries," *Journal of Economic Perspectives*, vol. 34, no. 1, pp. 94–121, 2020.

[20] D. Holtz, M. Zhao, S. G. Benzell et al., "Interdependence and the cost of uncoordinated responses to COVID-19," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 33, pp. 19837–19843, 2020.

[21] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *IEEE Computer*, vol. 52, no. 1, pp. 36–45, 2019.