Application of Computer Data Mining Technology in the Analysis of Emotional Law

Shuyue Zhang1,*

1Guangzhou College of Technology and Business, Guangzhou city Guangdong Province 510850, China

*Corresponding author email: 287190712@qq.com

Abstract. In recent years, with the in-depth development of the Internet, computer data mining technology has gradually matured, social networks have been integrated into people's daily lives, and more and more Internet users have expressed their personal emotions and commented on social hotspots on social networks. In this regard, social network public opinion analysis plays a vital role, and has attracted widespread attention from academia and industry. In order to better excavate Weibo users' opinions and ideas on social events, and track the development of the events, it is more conducive to the work of public opinion monitoring and rumor control. This paper expounds the research status of data mining and the current situation of psychological problems of college students. It introduces computer data mining technology into the analysis of college students' emotional regularity, and introduces the concepts, functions, technologies, methods and processes of data mining. This paper proposes a clustering method based on the characteristics of the semantic recessive sentiment of frequent itemsets, which fully considers the implicit sentiment regularity of semantics in Weibo. Firstly, we define the frequent feature word set of Weibo based on the characteristics of dominant sentiment, and use the maximum frequent items. Clusters to obtain initial clusters of regular emotions; for text overlap between initial clusters, an inter-cluster overlap reduction algorithm based on short text extended semantic membership is proposed to obtain completely separated initial clusters; based on cluster semantic similarity matrix, agglomerated emotions are given Regular clustering method. Finally, the validity of the text method is verified by the training corpus data provided by the NLP & CC2013 evaluation. It can be known from the test results that when the minimum support degree θ of the cluster is 0.5-0.6, the clustering effect of the analysis of sentiment regularity is better.

Keywords: Computer Data Mining Technology, Sentiment Law, Data Mining Clustering Algorithm, Sentiment Analysis
1. Introduction

With the deepening of the research on people's emotion laws, there are also related researches on the application of information technology and computer technology in the evaluation and early warning system of people's emotion laws. Based on system dynamics, the study of university students 'emotion laws early warning technology, the analysis and analysis of university students' emotion laws and early warning systems Design and implementation. Data mining technology has its unique advantages over other technologies, that is, it can mine information that is difficult to find between people's emotional problems and statistical data. Data mining can also find the statistical information between people's emotional problems and their basic information. Relationships and the main factors that cause emotional problems.

Weibo also contains a large number of sentence texts that contain users 'subjective feelings. The short text often reflects these users' preferences for a particular product, brand, or personal opinions on popular events in today's society. The number of Weibo users is huge, and the opinions and opinions expressed by the vast majority of users are universal and persuasive. In addition, because the user's starting point is different, their views on the same event are very different. Therefore, timely analysis of the emotional situation of the majority of Weibo users will have a certain reference value for the government to correctly guide public opinion and enterprises to further improve their products.

This article combines the goals of Weibo sentiment regularity analysis and uses computer data mining technology to conduct fine-grained analysis of emotions, which are divided into seven emotions: joy, preferences, anger, sadness, disgust, surprise, and fear, and the short text of Weibo for data mining emotion Regular analysis is more conducive to social network public opinion analysis. This paper analyzes the shortcomings of the traditional frequent-item-based clustering algorithm on the short text clustering problem on Weibo. This paper proposes a clustering method based on the frequent itemset semantic recessive emotion regular feature clustering. The item set constructs the initial clusters; then the semantic library is introduced to achieve overlap reduction between clusters based on the semantic hierarchy; finally, the microblog sentiment regularity clusters are merged by agglomerative hierarchical clustering to obtain the final emotional regularity clusters. At the same time, this paper verifies the effectiveness of the algorithm through experimental comparison.

2. Proposed Method

2.1. Computer Text Sentiment Analysis Data Mining Method

As a special text classification problem, sentiment analysis has the common problems of traditional pattern classification. In the process of text sentiment classification, it is also necessary to choose an appropriate method.

(1) Text classification method

Computer text classifier methods are currently more popular mining algorithm methods: K-nearest neighbor mining algorithm classifier, Naive Bayes mining algorithm classifier, support vector machine mining algorithm classifier, Rocchio mining algorithm classifier, and so on. Most of these commonly used computer text classifiers can be directly applied to computer text sentiment classification.
1) K-nearest neighbor mining algorithm classifier

K-Nearest Neighbor classification algorithm, k represents k nearest neighbors, meaning that each sample can use its k nearest neighbors as a representative. The core idea is: for the feature space, among the k samples of the nearest neighbor of a sample d, when most of the nearest neighbor samples belong to a certain category c, it means that the sample d contains category c features, and the category of the sample d is determined as c, that is, the category determination is based on the k sample categories of the nearest neighbors. In this algorithm, the similarity calculation between samples is measured using the cosine value.

The KNN algorithm is simple to implement, convenient to implement, and does not require training samples. It is suitable for the needs of classification standards that change at any time. It has a great advantage in dealing with multi-classification problems (a document belongs to multiple categories). The disadvantage is that the amount of calculation is large. When determining the category of the test file, it needs to be compared with all the training files. It is difficult to control the time complexity and space complexity, and it is not suitable for processing unbalanced data.

2) Naive Bayes Mining Algorithm Classifier

Naïve Bayesian classification (NB) is a simple classification model, and has achieved good results in computer text classification. The algorithm idea is to solve the computer text to be classified, find the probability of the various categories appearing under the condition of the feature item, and finally divide the computer text into a category based on which category has the highest probability.

In computer text classification, given document d and a fixed set of categories c. The document is document with class class. The document space is a high-dimensional space. The training set for a given category is a category corresponding to a document. The probability of the document d belonging to the category i c is calculated as follows:

\[ p(c_i) = \frac{p(d|c_i) \cdot p(c_i)}{p(d)} \]  

(1)

In computer text classification, our goal is to find the category to which a document most likely belongs. For the NB score, the most likely class is the result of \( C_{map} \) with a maximum a posteriori (MAP) estimate.

\[ C_{map} = \arg \max P(c | d) \arg \max P(c) \prod_{t \in \text{train}} P(t \mid c) \]  

(2)

Where \( P(t_k \mid c) \) is the conditional probability that \( t_k \) appears in the class c document, and P (c) is the prior probability that the document appears in class c. In order to get rid of the zero probability, the simpler method uses the add one smoothing method, which adds 1 to each number.

\[ P(t \mid c) = \frac{T_{ct} + 1}{\sum_{c'} T_{c't} + 1} = \frac{T_{ct} + 1}{\sum_{c'} T_{c't} + B} \]  

(3)

Where \( B = | V | \) is the number of all words in the vocabulary. Plus one smoothing can be considered as the result of using a uniform distribution as the prior distribution (each term appears only once in each class), and then updated according to the training data.
2.2. Computer Text Data Mining Clustering Algorithm

The clustering algorithm is an unsupervised machine learning algorithm that directly divides similar data objects into different categories or clusters. The text clustering technology is based on the assumption that texts in the same category are more similar, and texts in different categories are more different. The clustering algorithm can be divided into hierarchical clustering methods, partitioning clustering methods, density-based clustering and other methods.

1) Hierarchical clustering method

Hierarchical method uses levels to decompose the data set, including two top-down and bottom-up operation modes, namely Agglomeration and Division Hierarchical Clustering.

The agglomeration hierarchy algorithm treats each data object as an independent cluster in the initial stage, and then iterates and merges atom clusters until it reaches the expected number of clusters or other termination conditions. The representative method is the AGENES method. The split-level algorithm is just the opposite of the AGENES method. It treats all data objects as the same type of cluster, and then performs cluster decomposition according to the rules until the expected number of clusters or other termination conditions are reached. The representative method is the DIANA method.

2) Division clustering method

Partitioning method first creates k partitions (k ≤ n), given n samples or data tuples, where each partition corresponds to each cluster one-by-one, and the object is processed by loop iteration Division transfer to improve the efficiency of division.

The main division methods of the partition clustering method include Kmeans, kmedoids, CLARA (Clustering LARge Application), CLARANS (Clustering Large Application Based Upon Randomized Search), etc. The most classic of the partition clustering algorithms is the Kmeans algorithm.

Kmeans algorithm, given a data set containing n data and the number of clusters k, first randomly select k data objects as the cluster center; then calculate the distance between the remaining samples and the k cluster centers, At the same time, assign the sample to the cluster with the smallest distance; then recalculate the center of each adjusted cluster and generate a new cluster center; finally, the median iteration of the algorithm continuously decreases and converges to a fixed value to achieve lateral alignment. The function converges.

Dividing and clustering methods generally have low computational complexity and are widely used, especially in relatively large data.

2.3. Weibo Sentiment Analysis

In terms of Chinese microblogs, experts and scholars have also conducted a lot of research. In Chinese microblogs, users participate in emotional discussions of events or express their views on certain things and certain products, including the personal emotional color of users. Currently, the research is mainly carried out from the following two aspects:

1) Classification of emotion categories
The division of human emotion laws is a complicated process. So far, psychologists have no recognized standard for the division of human emotion laws. Human emotions are complex and diverse, and different classification results can be obtained from different perspectives. Ancient sages often categorized emotions into six or seven. The "personal feelings" recorded in the Book of Rites are divided into joy, anger, sadness, fear, love, evil, and desire. Psychologists divide emotions into 18 types: joy, sadness, hatred, and shame.

2) Emotional law in this article

Emotional law is a person's attitude and experience on whether objective things meet personal needs. In the short text of Weibo, we found that Weibo emojis are widely used, and emoticons have great expressiveness and expression. More simple, free, and visual. These emojis can intuitively express the user's current mood. In addition, emotional regular words are words or phrases that have a tendency of emotional regularity in the text. They have a clear color of emotional regularity. Words can intuitively analyze the user's emotion law, so emojis and emotion law words are intuitive and obvious characteristics of emotion law.

However, the analysis of sentiment regularity based on the characteristics of explicit sentiment regularity is insufficient, because the content semantics of short text on Weibo is not considered. In contrast, the content semantics of short text on Weibo is a hidden feature of emotional regularity. Semantic analysis helps to optimize the process of sentiment analysis. Therefore, this article proposes a microblog sentiment analysis method that combines two characteristic factors (dominant and implicit sentiment regularity features).

3. Experiments

3.1. Identify Research Objects and Mining Goals

In this study, the basic personal information of the students provided by the psychological counseling center of a university and the survey results of the students in the school using the Self-evaluation Scale for Sentiment Law SCL-90 were selected as mining data. Randomly select some symptoms of emotion law, hoping to establish a model of university students' emotion law through classification and mining, and provide new ideas and approaches for the analysis of university students' emotion law.

3.2. Dominant Feature Selection Based on Weibo Sentiment Regularity

Through the analysis of Weibo text, we found that if a feature item in Weibo has emotional characteristics, it is considered that the feature item reflects the emotional trend of the Weibo text to some extent, and this feature needs to be retained The characteristics of these words are the dominant emotional characteristics in Weibo. In Weibo, Weibo emojis and sentiment words are intuitive and explicit emotional features. Emoticons and sentiment words on Weibo reflect the emotional characteristics of Weibo more accurately. Dominant sentimental features of Weibo play a vital role in sentiment analysis. From the perspective of feature selection, these dominant sentimental features of Weibo need to be retained to better serve subsequent Weibo sentiment analysis.

4. Discussion
4.1. Performance Evaluation of Sentiment Law Analysis Algorithm

This experiment uses the second Natural Language Processing and Chinese Computing Conference (NLP & CC2013) hosted by the Computer Society to publicly mark the sentiment regular microblog data. The file is an XML file with a total of 14,000 microblogs, of which 4000 are training data and 10,000 are test data. The test data is divided into happy happiness, surprise, disgust, angry anger, sad sadness, like, fear, and no emotion as shown in Table 1 below.

| Emotion classification | Test data set | Training dataset |
|------------------------|---------------|------------------|
| Happy                  | 1112          | 360              |
| Preference             | 1539          | 361              |
| Sad                    | 862           | 321              |
| Disgust                | 343           | 127              |
| Anger                  | 457           | 284              |
| Surprised              | 295           | 130              |
| Fear                   | 195           | 94               |
| No                     | 6173          | 1950             |

As shown in Table 1 above, it is particularly noted that since a single Weibo may contain multiple different emotions, the analysis of the sentiment regularity in this article is mainly based on the main emotional rule of the blogger. From the table, we can see that the fusion of dominant and implicit in this article. The semantic semantic clustering method can effectively analyze the sentiment regularity of Weibo.

4.2. Selection of Minimum Support for Frequent Item Clusters

In the process of agglomerative hierarchical clustering based on semantic similarity, we set the minimum threshold $\lambda$ of the similarity between clusters of each emotional rule. When the similarity between clusters is less than $\lambda$, the clustering of emotional regular clusters ends. In order to analyze the selection effect of different lambdas, our cluster minimum support $\theta$ value is 0.5, and the minimum threshold $\lambda$ of similarity of different emotion rules clusters is selected. Different lambda values affect the regularity of the emotional law clustering results. Shown in Figure 1.
Figure 1. Effect of minimum support $\theta$ of different clusters on F-measure value

As shown in Figure 1 above, the minimum support degree $\theta$ for obtaining frequent item clusters directly affects the number of initial clusters obtained during the feature extraction stage, and thus affects the experimental results. In order to analyze the selection effect of the minimum cluster support $\theta$, we choose different cluster minimum support $\theta$, and the effects of different $\theta$ on the affection rule clustering result macro average F-Measure average and micro average F-measure. The test results show When the minimum support degree $\theta$ of the cluster is 0.5-0.6, the clustering effect of emotion rule analysis is better.

5. Conclusion

This article analyzes and compares the existing Weibo collection methods. This article uses the API provided by Sina Weibo to collect Weibo. This provides a guarantee for the data source and also provides resources for subsequent sentiment analysis. Based on the characteristics of Weibo, Short text performs a series of pre-processing tasks, including Chinese word segmentation, stop word removal, etc., to reduce noise interference.

Acknowledgements

Academy-level scientific research projects of Guangzhou College of Technology and Business in 2019 (No.: KA201929)

References

[1] Youxue Jiang. Research on E-commerce Consumer Factors of Agricultural Products in Agritourism Based on Data Mining Technology[J]. International Journal of Multimedia & Ubiquitous Engineering, 2016, 11(11):297-308.

[2] Borroni S, Fossati A, Gratz K L, et al. The Mediating Role of Emotion Dysregulation in the Relations Between Childhood Trauma History and Adult Attachment and Borderline
Personality Disorder Features: A Study of Italian Nonclinical Participants[J]. 2016, 30(5):653-676.

[3] Borrasmorell J E. Data mining for pulsing the emotion on the web.[J]. Methods Mol Biol, 2015, 1246(1246):123-130.

[4] Kose, Eiji, Uno, Kana, Hayashi, Hiroyuki. Evaluation of the Expression Profile of Extrapyramidal Symptoms Due to Antipsychotics by Data Mining of Japanese Adverse Drug Event Report (JADER) Database[J]. YakugakuZasshi Journal of the Pharmaceutical Society of Japan, 137(1):111-120.

[5] MI Yunlong, MI Chunqiao, LIU Wenqi. Research Advance on Related Technology of Massive Data Mining Process[J]. Journal of Frontiers of Computer Science & Technology, 2015, 9(6):641-659.

[6] Youxue Jiang. Research on E-commerce Consumer Factors of Agricultural Products in Agritourism Based on Data Mining Technology[J]. International Journal of Multimedia & Ubiquitous Engineering, 2016, 11(11):297-308.

[7] Yang K, Shi-Zhu L I. Application of Big Data Mining Technology in Monitoring and Early-warning of Schistosomiasis[J]. Chinese Journal of Parasitology & Parasitic Diseases, 2015, 33(6):461.

[8] Royce Kimmons, George Veletsianos. Public Internet Data Mining Methods in Instructional Design, Educational Technology, and Online Learning Research[J]. Techtrends, 2018, 62(5):492-500.

[9] Amit Kumar, Ajay Kumar. Data Mining Technology and its Applications to Power Systems[J]. International Journal of Computer Applications, 2015, 131(8):1-6.