2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Analysis of Public Opinion on Weibo and Construction of Knowledge Gragh

Dandan Wei,Chunxiao Zhao,Boxuan Zhang,Jialin Ma,Quanyin Zhu*,Xinxin Xu,Shengbiao Wang

Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huaian, China
*Corresponding author's email: hyitzqy@126.com

***Abstract*—Weibo is gradually replacing traditional media, becoming the main way for netizens and youth groups to obtain and exchange information. Weibo public opinion will have a great impact on society. This system is mainly based on the topic of public opinion mining and sentiment analysis, constructing a topic and sentiment knowledge graph on Weibo public opinion to provide a basis for guiding the direction of Weibo public opinion.**

Keywords: Knowledge Graph, Weibo Public Opinion, Sentiment Analysis

I. INTRODUCTION

With the development of the times, the concept of online public opinion has become more and more popular. Generally speaking, public opinion refers to the masses of people living in their own social environment. In the general environment, people hold opinions, attitudes, and personal beliefs about other people, events, things, and their development status. With sentiment, etc[1].these opinions can also quickly spread to all corners of the network platform through comments, reposts, etc., which has enabled the network public opinion to gain a huge spread, extremely fast transmission speed, and strong social influence.

As a new type of online communication media, Weibo allows users to share short texts, pictures, or short videos in a very convenient way. It has attracted the attention and love of a large number of netizens. The number of Weibo users and the amount of data have exploded. Growth. Therefore, in the current information age, Weibo has become the most convenient channel for public opinion, and negative public opinion is extremely likely to have a great impact on society.

This project is to use the Single-Pass algorithm to perform text clustering on the cleaned and preprocessed data, model LDA topics, and build a knowledge graph with topic clusters. On the basis of the knowledge graph of Weibo topics, we find the topic of Weibo public opinion, and use the convolutional neural network technology to construct the emotional knowledge graph of Weibo public opinion.

II. RELATED WORK

A Single-Pass algorithm clustering

Weibo public opinion analysis is to filter out hot event related information from many data on Weibo for cluster analysis. The classic Single-Pass algorithm based on incremental is more suitable, and in order to achieve better clustering effect, Gesandoji [2] Et al. proposed an improved method based on the difference in the importance of different location feature items in semi-structured web pages, and achieved good experimental results. Based on this, this paper selects the improved Single-Pass clustering method for clustering, and then makes up for the deficiencies of the classic Single-Pass algorithm analyzed above. The improvement mainly comes from two aspects. One is to reduce the randomness of the cluster center selection. The main idea is that if there are many vector points around vector point A, it is considered that the vector points are dense in this area, and vector point A has a strong representative Sex. The operation step is to first give two data, namely the neighborhood radius eps, and the minimum number of objects in the neighborhood MinPts, also known as the minimum density threshold or the minimum number of neighborhood points, and then calculate the number of documents T in each document eps, when calculated When the number of documents T is greater than the minimum density threshold MinPts, this document can be used as the centroid of the initial clustering. The second is the process of optimizing the time required to compare the similarity one by one. The main idea is that the centroid can be used as a representative of a set to a certain extent, so when comparing the similarity, it can be compared with the centroid. The main step is to compare the new document with the clusters of topic clusters that have been partially clustered. Only need to compare the new document with the existing centroid data, select the highest similarity data and the similarity threshold size However, there is no need to compare the text with other documents, thereby reducing the calculation process and improving the efficiency of the algorithm.

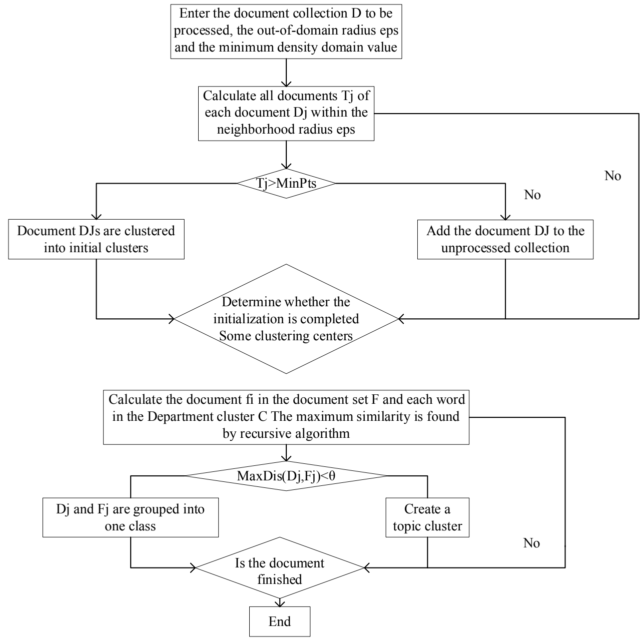
Enter the document collection D to be processed, the out-of-domain radius eps and the minimum density domain value
Calculate all documents Tj of each document Dj within the neighborhood radius eps
Tj>MinPts
No
No
Document DJs are clustered into initial clusters
Add the document DJ to the unprocessed collection
Determine whether the initialization is completed Some clustering centers
Calculate the document fi in the document set F and each word in the Department cluster C The maximum similarity is found by recursive algorithm
MaxDis(Dj,Fj)<θ
Dj and Fj are grouped into one class
Create a topic cluster
No
Is the document finished
End

Figure 1. Improved Single-Pass algorithm flow chart

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00082

The weight of a feature item represents the ability of a feature word to be different from other texts, and how much it can play in the text representation. The weight calculation method used in this paper is the TF-IDF weight function [3]. The core criterion of the category discrimination ability of a word is that the frequency of occurrence in a text is higher but the frequency of occurrence in other texts is lower. TF-IDF's TF represents the frequency of occurrence of a word in the text d, IDF represents the frequency of inverse documents, used to measure the ability of a word to distinguish a certain category, which means that if a feature appears in fewer documents , The higher his IDF value, the more obvious the corresponding regional component force, calculation method: $w_{ij} = tf_{ij} \times idf_{ij}$

Among them, idf$_{ij}$ represents the inverse document frequency of feature tij, and tfij represents the frequency of feature t$_{ij}$ in the jth text Dj. The frequency of the inverse document is calculated as:

$$idf_{ij} = \log\left(\frac{N}{n_{ij}}\right) + 0.01 \tag{1}$$

N represents the total number of text sets, and nij represents the normalized TF-IDF function calculation method that includes the feature t$_{ij}$ as:

$$w_{ij} = \frac{tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right)}{\sqrt{\sum_{j=1}^{m}\left(tf_{ij} \times \log\left(\frac{N}{n_{ij}} + 0.01\right)\right)^2}} \tag{2}$$

tf$_{ij}$ represents the frequency of the j-th word in the i-th document, and w$_{ij}$ represents the weight of the j-th word in the i-th document.

For text similarity calculation, if the computer determines whether the content expressed in the text belongs to the same topic, it is necessary to calculate the similarity between different texts to achieve the purpose of putting documents with higher similarity in the same topic, which is convenient for the computer to perform operations . The distance between the vectors can be used to express [4], generally in the following way:.

1. Inner product: $sim(d_i, d_j = \sum_{k=1}^{n} w_{ik}, w_{jk}$ (3)

2. Absolute distance： $sim(d_i, d_j \sum_{k=1}^{n} \left|w_{ik} - w_{jk}\right|$ (4)

3.Euclidean distance:

$$sim(d_i, d_j) = \sqrt{\sum_{k=1}^{n}(w_{ik} - w_{jk})^2} \tag{5}$$

4. Chebyshev distance:

$$sim\left(d_i, d_j\right) = \max\ |w_{ik} - w_{jk}| \tag{6}$$

5. Angle cosine：

$$sim(D_m, D_n) = \cos\theta = \frac{\sum_{i=1}^{n} w_{mi} \times w_{ni}}{\sqrt{(\sum_{i=1}^{n} w_{mi}^2)(\sum_{i=1}^{n} w_{ni}^2)}} \tag{7}$$

Among the above methods, cosine similarity is the most widely used method in the application field, with simple operation and high feasibility. In this paper, this method is used to calculate the similarity. The degree of similarity between Dm and Dn is expressed as sim(D$_m$, D$_n$), wmi and document Dm and Dn i-th feature weight, N represents the sum of two vector feature items, and 1≤i≤N.

B LDA topic modeling

As a type of unstructured character data, microblog text is something that a computer cannot easily "understand", so it is necessary to convert semi-structured or unstructured text objects into structured data that is easy for the computer to understand for subsequent follow-up Cluster analysis. At this stage, the mainstream text representation methods are divided into two types: probability model and vector space model (VSM). The LDA model in the probability model is a 3-layer Bayesian probability model. Due to its certain advantages in text clustering, it has become a popular direction for document topic generation models [5]. Therefore, this project uses the LDA model for text representation of Weibo data. Set the text set to have a total of k topics, and each topic z is expressed as a polynomial distribution of terms. The process of LDA's probability graph model is shown in Figure 2.

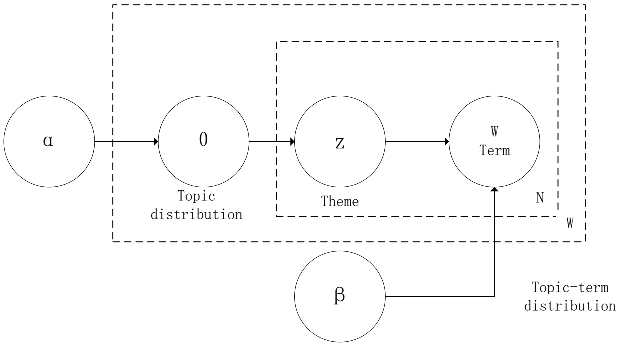

Figure 2. LDA probability graph model

In Figure 2, box W represents the text set, box N represents the set of topics z and keywords w in the text set, α is the prior parameter of the Dirichlet distribution, β is the estimated matrix parameter, and θ is all The probability distribution of topics.

The probability density distribution of a Dirichlet random variable θ (k dimension) can be calculated by equation (8):

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i}{\prod_{i=1}^{k}\Gamma(\alpha_i)} {\theta_1}^{\alpha_1 - 1} \dots {\theta_k}^{\alpha_k - 1} \tag{8}$$

Among them, $\Gamma(.)$ represents a gamma function. Let z denote a theme vector with N elements, then the joint distribution can be calculated by equation (9):

$$P(\theta, z, w \mid \alpha, \beta) = P(\theta \mid \alpha) \prod_{n=1}^{N} P(z_n \mid \theta) P(w_n \mid z_n, \beta) \tag{9}$$

$$P(w \mid \alpha, \beta) = \int P(\theta|\alpha)\left(\prod_{n=1}^{N} \sum_{z_n} P(z_n \mid \theta) P(w_n \mid z_n, \beta)\right) d\theta \tag{10}$$

Where P(Zn|θ) is the value of θ that satisfies the condition z_n^i=1.

C CNN

In the theoretical and practical research of sentiment classification, clustering technology, text keyword extraction, semantic understanding, and machine learning technology have been continuously adopted, and have achieved certain research results [6]. Sentiment-based dictionaries are the basis of traditional sentiment analysis, but often sentiment dictionaries have higher construction costs, insufficient coverage, difficulty in updating, and low accuracy. Therefore, scholars have adopted machine learning and artificial intelligence techniques to make up for the shortcomings of traditional research, and convolutional neural networks have powerful feature learning capabilities, which can overcome the difficulty of artificial feature extraction, thereby making emotion classification and learning easier and more feasible. It can greatly improve the efficiency and accuracy of emotion classification, and it is gradually favored by experts and scholars, and it is widely used at the practical level [7]

According to research needs, combining Weibo users' forwarding, commenting, and liking behaviors of Weibo information, the emotions of Weibo users are divided into 7 categories: happy, appreciation, surprise, sadness, disgust, fear, and anger. The "seven emotions" classification is relatively consistent. The emotion classification and typical microblog content emoticons and example words are shown in Table I.

The main principle of CNN Weibo public opinion sentiment classification model algorithm is to use a large number of Weibo blog posts or comment data training sets that have been manually marked with emotional polarity to train the model, and input the trained model into the test set data to start testing to meet the accuracy requirements. After that, sentiment classification is performed on microblogs to be classified. Compared with sentiment classification models such as semantic dictionary, it has the advantages of high classification accuracy and strong noise data analysis ability.

CNN Weibo public opinion sentiment classification model training algorithm includes 4 steps and 2 stages. The first stage is the forward propagation stage. First, we extract from the training set (train_ x^i， train_ y^k)-a sample train_ x input network, and calculate the actual output value opts_ y^k .At this stage, the information is gradually transformed from the input layer to the output layer, and the operation of formula 9 is performed. The second stage is the backward propagation stage. First, the difference between the actual output opts_y^k and the target output train_ y^k is calculated, and then the matrix is adjusted according to the accuracy control requirements. The calculation of the first and second stages must be within the range of sample error, the sample set error $rL = \sum rL^k$, and the error rL^k of the kth sample is measured according to Equation (12).

$opts\_y^k =$

backward by layer based on the error, so the second stage is also called the backward propagation stage or the error

$$F_n\left(\dots\left(F_2\left(F_1\left(train_x{}^i W(1)\right)W(2)\right)\dots\right)W(n)\right) \qquad (11)$$

$$rL^k = \frac{1}{2}\sum_{j=1}^{n}(train_{y_j}{}^k - opts_{y_j}{}^k)^2 \qquad (12)$$

TBALE I. WEIBO SENTIMENT CLASSIFICATION

Emotion classification	Typical emoji	Typical examples
happy	🙂	be full of joy, be all smiles, be delighted that things are better than one expected, a smile has driven all the hard lines in his face and brightened his countenance
appreciate	👍	unanimously praise, enjoy great popularity among the people, diligently, be conscientious and do one's best
surprised	😲	be struck dumb, terrified, be startled at, can't believe it
sadness	😭	choke with sobs, Swallow one's voice and endure tears, be overcome with grief, wring one 's heart to the very core
hate	😣	disgust, despise, nausea, abhor, detest and detest, antipathy
fear	😨	jittery, shiver all over though not cold, have a lingering fear， tremble with fear in one 's boots
anger	😡	bristle with anger, be unable to contain knew no bounds, cleft canthus, as mad as a hornet

From the perspective of the propagation direction of Weibo public opinion information, the second stage of the solution process is the opposite of Weibo's chain propagation path for Weibo public opinion information. When the neuron connection weight is adjusted, the output layer error can be solved, but other The layer error needs to be derived

propagation stage. In the training process, let the number of input layer units be N, the number of middle layer units be Z, the number of output layer units be C, the input vector X =($x_0, x_1, \dots, x_n$), and the middle layer output vector M = ($m_0, m_1, \dots, m_z$), the output layer vector Y=($y_0, y_1, \dots, y_c$), the target output vector O=($O_0, O_1, \dots, O_C$), the weight of the output unit and the hidden unit is V_{ij}, implied The unit-to-output unit weight is W_{jk}, the implicit unit threshold ∂_j, and the output unit threshold. Then, the algorithms of the middle layer and the output layer are shown in Equations 13 and 14, respectively, where f(x) =

$$m_j = f\left[\sum v_{ij} x_i + \partial_j\right] \tag{13}$$

$$y_k = f\left[\sum w_{jk} m_j + \emptyset_k\right] \tag{14}$$

Based on the above conditions, the training process of Weibo public opinion sentiment classification network is as follows:

(1) Determine the training samples. Randomly select a certain amount of Weibo public opinion data from the training set as the training sample X.

(2) Set the weights V_{ij}, W_{jk}, and the threshold ∂_j to be random numbers close to 0, and initialize the learning rate α and the error control parameter β.

(3) Select a sample X_i from the training samples X, given the target output vector O. The intermediate layer output vector M is solved based on the formula 15 algorithm, and then the actual output vector Y is solved based on the formula 16 algorithm. The output vector y_k and the target vector O_k. Compare and calculate the output error τ_k, and calculate the hidden layer unit error τ_j in the same way. among them

$$\tau_k = (O_k - y_k) y_k (1 - y_k)、\tau_j (1 - m_j) \sum \tau_k W_{jk} \tag{15}$$

(4) Calculate the weight adjustment amount according to equations 15 and 16, and calculate the threshold adjustment amount according to equations 17 and 18.

$$\Delta V_{ij}(n) = (\alpha/(1+N) \times (\Delta V_{ij}(n-1) + 1) \times \tau_j \times m_j) \tag{16}$$

$$\Delta V_{jk}(m) = (\alpha/(1+M) \times (\Delta V_{ij}(m-1) + 1) \times \tau_k \times m_j) \tag{17}$$

$$\Delta\phi_k(n) = \left(\frac{\alpha}{1+N}\right) \times (\Delta\phi_k(n-1) + 1) \times \tau_j \tag{18}$$

$$\Delta\partial_j(m) = (\alpha/(1+N)) \times (\Delta\partial_j(n-1) + 1) \times \tau_j \tag{19}$$

(5) When k traverses 1 to C, if the sample set error rL is less than or equal to the error control parameter β, then enter the next step, otherwise return to step (3) Iterative calculation until the conditions are met.

(6) Save the weight and threshold data, and the classifier training is completed.

III. Conclusion

This project uses Single-Pass algorithm for text clustering, LDA topic modeling, and topic clusters to build knowledge graphs. Based on the knowledge graph of Weibo topics, the topic of Weibo public opinion was found, and the convolutional neural network technology was used to complete the construction of the emotional knowledge graph of Weibo public opinion.

References

[1] Yan Lu. Research on early warning mechanism of network public opinion crisis under big data environment [D]. Jilin: Jilin University, 2018

[2] Ge Sangduoji, Qiao Shaojie, Han Nan, etc. Single-Pass-based network public opinion hot spot discovery algorithm [J].Journal of University of Electronic Science and Technology of China, 201 5(4): 599-60

[3] Lin Yongmin, Lu Zhenyu, Zhao Shuang, Zhu Weidong. Analysis and improvement of text feature weighting method TF-IDF [J]. Computer Engineering and Design, 2008, (11): 2923-2925

[4] Ma Junhong. Text similarity calculation theory and application research [D]. Northwest University, 2011

[5] Xu Weilin, Zhu Zong, Gao Li, et al. Analysis of public opinion on microblog based on topic model . Software Guide, 2016, 15(5): 153-154.

[6] Li Jie, Li Huan. Research on feature extraction and sentiment classification of short text review products based on deep learning [J]. Information Theory and Practice, 2018, (2): 143-148.

[7] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521 (7553) :436.