Cluster Analysis for Internet Public Sentiment in Universities by Combining Methods

https://doi.org/10.3991/ijes.v6i3.9670

Na Zheng(✉), Jie Yu Wu
Hebei Agricultural University, Hebei, China
zoenalj@163.com

Abstract—A clustering method based on the Latent Dirichlet Allocation and the VSM model to compute the text similarity is presented. The Latent Dirichlet Allocation subject models and the VSM vector space model weights strategy are used respectively to calculate the text similarity. The linear combination of the two results is used to get the text similarity. Then the k-means clustering algorithm is chosen for cluster analysis. It can not only solve the deep semantic information leakage problems of traditional text clustering, but also solve the problem of the LDA that could not distinguish the texts because of too much dimension reduction. So the deep semantic information is mined from the text, and the clustering efficiency is improved. Through the comparisons with the traditional methods, the result shows that this algorithm can improve the performance of text clustering.

Keywords—subject models; vector space model; comment extraction；public opinion；text clustering

1 Introduction

In recent years, with the development of the information technology, the Internet has been widely available in the university, The number of the students who use Internet has increased considerably [1]. By some Internet public affairs, People start to understand the important function of intendancy. The uniqueness of well-educated college students makes them the major and indispensable components of the internet public sentiment. The complicated network public sentiment on Internet brings a challenge which cannot be ignored on the political and ideological work in universities. Therefore, student management departments should reinforce their work on online public opinion collection, research, and assessment, and attach importance to the control and guidance of the internet public sentiment. In Internet public opinion analysis, the student affairs administrators need some intelligent methods to find the exact information in the magnanimous information sources for deeply analysis. Only by using intelligent algorithm to collect and analysis public opinion corpus automatically, an effective, comprehensive and fast monitoring early-warning mechanism can be established.

According to the requirement of analysis of network public opinions at colleges and universities, an online public opinion detection and analysis clustering method has built based on LDA (Latent Dirichlet allocation).This algorithm melts the subject models based on Latent Dirichlet Allocation and the VSM model based on TF-IDF weight to compute text similarity, and the cluster analysis is carried out. So the deep semantic information is mined from the text, and the clustering efficiency is improved. Through the comparisons with the traditional methods, the result shows that this algorithm can improve the performance of text clustering.

2 The subject models

2.1 Latent Dirichlet Allocation

One subject can expressed as a certain distributions of word frequency, And a paragraph or a sentence is regarded as being generated from a probabilistic model. To measure the document similarity, the most common way is to calculate the times of the words that appear at the same time in two documents, TF-IDF algorithm is one of the common methods. The deficiency in this approach is ignoring the implications inherent in the documents. Sometimes No word appears in both documents at the same time, but the two documents are related to each other semantically. So the implied semantic must be considered when judging the similarity between two documents. It is necessary to take the subject model. The Latent Dirichlet Allocation is a kind of the most common model.

The probability of each word in the document can be expressed as the formula below:

$$p(word\,|\,text) = \sum_{topic} p(word\,|\,text) * (topic\,|\,text) \qquad (1)$$

The probability formula can be represented by matrices, see Figure1.

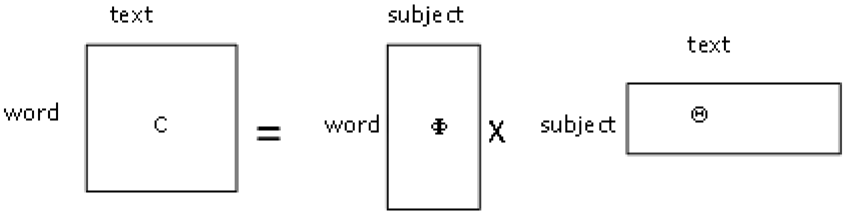

Fig. 1. The text-word-subject matrix

In the figure1, the text-word matrix on the left gives the probability distributions of each word in the text. And the topic-word matrix on the right gives the probability distributions of each word in each topic. The text- topic matrix gives the probability of the appearance of each topic in the text. For a series of text, after the character division, the probability of the appearance of each word in the text is calculated.so the text-word topic on the left can be obtained. The construction of the topic model is to get the text- topic matrix and the topic-word matrix on the right side by training and learning the text-word matrix on the left side. In the topic mode of Latent Dirichlet Allocation, all the documents in a document set can be regarded as a combination of the topics in the latent topic set according to the probability. The structure of the topic model based on Latent Dirichlet Allocation can be described as a three-layer topology, see Figure 2.

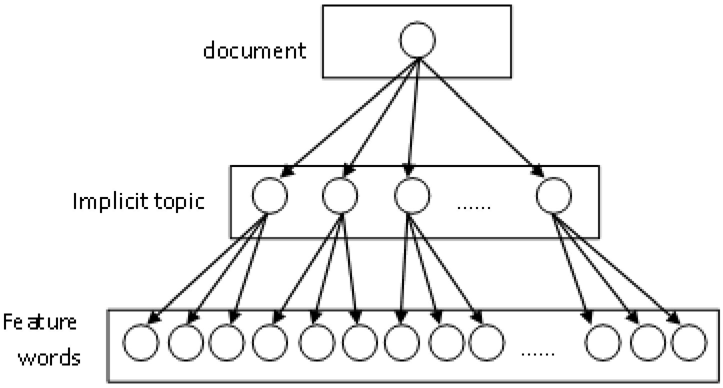

Fig. 2. Three-layer topology of Latent Dirichlet Allocation topic mode

As can be seen from the figure, one document can be viewed as a random combination of all the topics of different probability in the implicit topic set. The generation process of each document in the document set by the topic model based on Latent Dirichlet Allocation is:

Step1: for a document in the document set, randomly select a topic in the corresponding topic set.

Step2: randomly select a word in the corresponding word set in the selected topic.

Step3: Repeat these steps until complete coveraging all the words in the document .the topic model based on Latent Dirichlet Allocation consists of three layers, the document set layer,the document layer and the feature words layer.

The three-tier structure based on Latent Dirichlet Allocation see Figure 3.

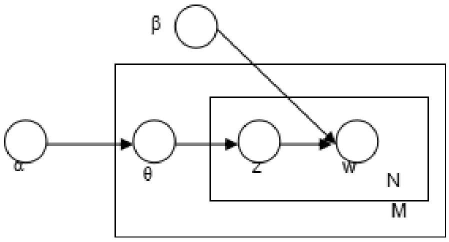

Fig. 3. The three-tier structure based on Latent Dirichlet Allocation

In the figure 3, parameter α and β are used to define the document set layer of the topic model. The vectorαis used to generate the vector θ. The matrix β represents the probability distribution of words which corresponding to potential topics. Parameter α and β show the level of the document set, it's value only need to be set once. The random variable θ defines the document layer of the subject model. θi is the probability distribution of each latent topic in the ith document, it is a vector. θ is a variable with the level of document. Each document maps to a variable named θ, the probability that every document generates the subject named z is different, the value of θ also only need to be set once for each generated document. Z and w are the parameters that indicate the level of the feature words, w is a vector of the feature words in a document. Z shows the distribution of all the feature words in a document. W is a observed variable .Z and θ are hidden variables.α and βare got by learning, w and z are the variables in the word level, z comes from θ and w comes from z.

So the joint probability of Latent Dirichlet Allocation can be expressed as:

$$p(\theta,z,w\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N}p(z_n\mid\theta)p(w_n\mid z_n,\beta \quad (2)$$

The process of generating a document with the subject model based on Latent Dirichlet Allocation should cover the following steps.

Step1: get the scale of the feature words $N = possion(\xi)$ in a document.it can be regarded as selecting the number of the feature words.

Step2: get the parameter $\theta \sim dir(\alpha)$ of the subject distribution in a document. and α is the Dirichlet distribution parameter.

Step3: generate all the feature words for every document.

1) select an implied subject named z. it comes from the polynomial distribution $mulitinomial(\theta)$ of the subject distribution probability vector named e.

2) then select a feature word w.it comes from the polynomial probability distributions $mulitinomial(\phi^{(z)})$ of the latent subject named z.

Using the above steps, the generating probability of the ith feature word named wi in document d can be expressed as:

$$P(w_i)=\sum_{j=1}^{T}P(w_i\mid z_i=j)P(z_i=j) \quad (3)$$

The probability that document d includes the feature word w can be expressed as:

$$P(w\mid d)=\sum_{j=1}^{T}\varphi_w^j*\theta_j^d \quad (4)$$

Then the maximum likelihood estimate is taken to build the three layer model of the Latent Dirichlet Allocation based on the parameters α and β .

$$I(\alpha,\beta)=\sum_{i=1}^{M}\log p(d_i\mid\alpha,\beta) \qquad (5)$$

The Conditional distribution of generating document d is expressed as $p(d\mid\alpha,\beta)$;

$$p(d\mid\alpha,\beta)=\frac{\Gamma(\sum_i a_i)}{\prod_i\Gamma(a_i)}\int(\prod_{i=1}^{T}\theta_i^{\alpha_i-1})(\sum_{n=1}^{N}\sum_{i=1}^{T}\sum_{j=1}^{M}(\theta_j\beta_{ij})^{w_n^j})a \qquad (6)$$

2.2 Estimations of parameters

The Gibbs sampling algorithm is used to estimate the parameters. The subject models based on Gibbs sampling algorithm is：

$$w_i\mid z_i,\phi_{z_i}\sim discrete(\phi^{z_i}) \qquad (7)$$

$$\phi\sim discrete(\beta) \qquad (8)$$

$$z_i\mid z_i,\phi^{d_i}\sim discrete(\phi^{d_i}) \qquad (9)$$

$$\theta\sim discrete(\alpha) \qquad (10)$$

In the formulas, ϕ^d is the prior probability of $dirichlet\ (\alpha)$, is the prior probability of $dirichlet\ (\beta)$.

In the implementation of Latent Dirichlet Allocation, it is only need to assign the words of the subject.so the variable z is made a sample analysis of. The formula of the posterior probability $P(z_i=j\mid z_{-i},w_i)$ is:

$$P(z_i=j\mid z_{-i},w_i)\propto\frac{n^{w_i}_{-i,j}+\beta}{n^{\cdot}_{-i,j}+W\beta}*\frac{n^{d_i}_{-i,j}+\alpha}{n^{d_i}_{-i}+T\alpha} \qquad (11)$$

In the formula, z_{-i} is all the distribution of z_k $n^{w\cdot}_{-i,j}$ is the number that the feature word wi belongs to subject j. $n^{\cdot}_{-i,j}$ is the feature word of subject j. $n^{d_i\cdot}_{-i,j}$ is the number that the feature word wi belongs to subject j in document di. $n^{d_i\cdot}_{-i}$ is the number of the feature words that belongs to subject j in document di.

The following is its basic process based on The Gibbs sampling algorithm.

Step1: give an initial sample set.

Step2: Cyclic sample through multiple iterations.

Step3:after multiple iterations, when the Markov Chain converges to a stable state, estimate the subject of each feature word. the subject distribution and the subject-word distribution are estimated by the formula below.

$$\phi_j^w = \frac{n_j^{\ w} + \beta}{n_j^{\cdot} + W\beta} \qquad (12)$$

$$\theta^d = \frac{n_j^d + \alpha}{n^{d\cdot} + T\alpha} \qquad (13)$$

In the formula, $n_j^{\ w}$ is the number that the feature word belongs to the subject j. $n^{d\cdot}{}_j$ is the number that the feature word in document d belongs to the subject j. n^d is the number of the feature words in document d that belongs to the subject j.

Suppose that There are two texts named di and dj. the text similarity of them are calculated by the formula below based on the TF-IDF weights strategy.

$$S_{TF-IDF}(d_i,d_j) = \frac{d_{i(TF-IDF)} * d_{j(TF-IDF)}}{|d_{i(TF-IDF)}| * |d_{j(TF-IDF)}|} \qquad (14)$$

Based on the subject vector from the Latent Dirichlet Allocation, the text similarity of them is calculated by the formula below.

$$S_{SDA}(d_i,d_j) = \frac{d_{i(LDA)} * d_{j(LDA)}}{|d_{i(LDA)}| * |d_{j(LDA)}|} \qquad (15)$$

This research uses the Latent Dirichlet Allocation subject models and the VSM vector space model of the TF-IDF weights strategy respectively to calculate the text similarity. Then the linear combination of the two results is used to get the text similarity. Then choose the cluster analysis based on k-means clustering algorithm. The formula of linear combination is as below.

$$S\ (d_i,d_j) = \lambda S_{TF-IDF}(d_i,d_j) + (1-\lambda) S_{LDA}(d_i,d_j) \quad (16)$$

In the formula, λ expresses the correlation coefficient. The process is illustrated in figure 4.

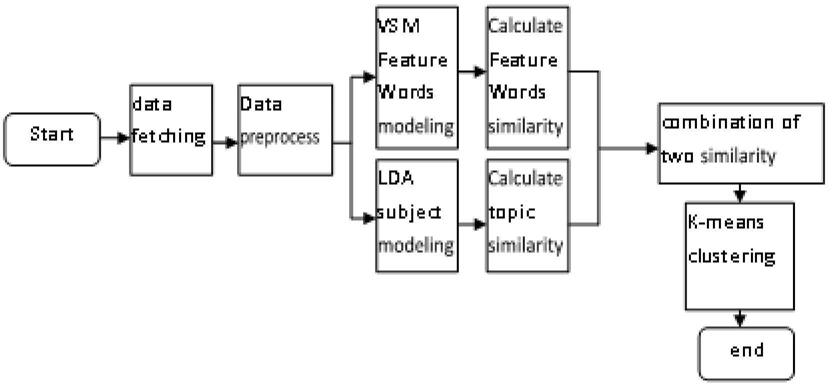
Start
data fetching
Data preprocess
VSM Feature Words modeling
LDA subject modeling
Calculate Feature Words similarity
Calculate topic similarity
combination of two similarity
K-means clustering
end

Fig. 4. Flow diagram of the K-means clustering

3 Experiments and analysis

The research data come from the Chinese corpus categorization database. The Simulations are separately based on the Latent Dirichlet Allocation subject models, the VSM vector space model of the TF-IDF weights strategy and the clustering algorithm of the linear combination of the two. The clustering quality is measured by the F-measure method involving Recall and precision.

The method of F-measure is:

$$F(i,j) = \frac{2*P*R}{P+R} \quad (17)$$

In the formula, P is the Precision and R is the Recall.

$$P = pecision\,(i,j) = \frac{N_{ij}}{N_i} \quad (18)$$

$$R = recall\,(i,j) = \frac{N_{ij}}{N_j} \quad (19)$$

Ni expresses the number of the samples whose class is i in the original set of data. Nj expresses the number of the objects whose class label is j in the clustering results.

Nij expresses the number of the samples in the intersection of Ni and Nj. The evaluation of cluster quality is often based on the weighted average of the value of F among classes. The formula is:

$$F = \frac{\sum_{i}\{[i] * F(i)]}{\sum_{i}[i]} \tag{20}$$

By comparing the three different methods, the results are shown in Table 1.

Table 1. The results of the three different methods

Feature ratio(%)	**F (%)**		
	VSM	***LDA***	***VSM+LDA***
1	83.49	81.83	95.03
10	80.93	85.97	94.46
30	79.50	81.80	94.18
50	78.09	84.99	94.16
70	79.40	78.53	93.63
80	80.15	77.09	92.84
90	80.53	80.08	87.95
100	80.84	82.14	89.84
Average value	80.06	81.17	92.03

The test data show that the Average value of the clustering quality based on the present method is 92.03%. The Value of F of this algorithm is improved by 14.95% of VSM algorithm, and improved by 13.38% of LDA algorithm.it is a noticeable improvement compared with the simple VSM and LDA method. Then through Taking the PKU Weiming BBS forum database as an example, it can be proved that the proposed algorithm can be used for online public opinion analysis.

4 Conclusion

Campus network is an important kind of information resources, and the extraction of its comments is the basic work of public opinion analysis researches and of college student management. The proposed algorithm based on the subject models and the vector space model is used to calculate the similarity grade for the clustering analysis, the results of experiment show that the proposed algorithm can enhance the clustering algorithms' accuracy.

5 Acknowledgment

2017 innovation and Entrepreneurship Education Teaching Reform in Hebei Province project (2017CXCY033)

2018 provincial Department of Human Resources and Social Security cooperation project (JRSHZ-2018-08007)

6 References

[1] J Allan, J Carbonell, G Doddington.1998:Topic Detection and Tracking Pilot Study: Final Report[A].In: Proceeding of the Broadcast News Transcription and Understanding Workshop[C], San Francisco, pp.194-218.

[2] Allan J, Carbonell J, Doddington G, Yamron J and Yang Y.1998: Topic detection and tracking pilot study: Final report[C].Proceedings of the DARPA Broadcase News Transcription and Understanding Workshop, Virginia: Lansdowne, February pp.194-218.

[3] Cao H, Jiang D, Pei J, et al.2009:Towards context-aware search by learning a very large variable length hidden markov model from search logs[C].Proceedings of the 18th international conference on Worldwide web.ACM, pp.191-200.

[4] Chatzopoulou G, Eirinaki M, Koshy S, et al.2011, The QueRIE system for Personalized Query Recommendations[J].IEEE Data Eng. Bull, vol.34(2): pp.55-60.

[5] Du Y, He Y, Tian Y, et al.2011,Microblog bursty topic detection based on user relationship[C].Information Technology and Artificial Intelligence Conference (ITAIC),2011 6th IEEE Joint International.IEEE, vol.1: 260-263.

[6] Guzman J, Poblete B.2013: On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model[C].Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description.ACM, 31-39.

[7] Joachims T, Granka L, Pan B, et al.2005:Accurately interpreting clickthrough data as implicit feedback[C].Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.ACM, pp.154-161.

[8] Kim H N, Rawashdeh M, Alghamdi A, et al.2012,Folksonomy-based personalized search and ranking in social media services[J].Information Systems, vol.37(1): pp.61-76. https://doi.org/10.1016/j.is.2011.07.002

[9] Kumaran G, Allan J.2004:Text classification and named entities for new eventdetection[C].Proceeding of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.ACM New York, NY, USA.pp.297-304.

[10] Lin C, He Y, Everson R, et al.2012,Weakly supervised joint sentiment-topic detection from text[J].Knowledge and Data Engineering, IEEE Transactions on, 2 vol.4(6): pp.1134-1145.

[11] Li S, Lv X, Wang T, et al.2010, The key technology of topic detection based on K-means[C].Future Information Technology and Management Engineering (FITME), 2010International Conference on.IEEE, vol.2: pp.387-390.

[12] Lu D, Li Q.2011: Personalized search on Flickr based on searcher's preference prediction[C].Proceedings of the 20th international conference companion on World wide web.ACM, pp.81-82.

[13] M Ester, Sander.1996:A density-based algorithm for discovering clusters in large spatial databases.Processing of the 2nd Conf.on Knowledge Discovery and Data Mining (KDD'96), pp.226-231.

[14] Naptali W, Tsuchiya M, Nakagawa S.2010,Topic-dependent language model with voting on noun history[J].ACM Transactions on Asian Language Information Processing (TALIP), vol.9(2): pp.7-11.

7 Authors

Na Zheng is with the Academic Affairs Office, Hebei Agricultural University, Baoding, Hebei, China.

Jie Yu Wu is with the College of Information Science & Technology, Hebei Agricultural University, Baoding, Hebei, China.

Article submitted 25 September 2018. Resubmitted 17 October 2018. Final acceptance 20 October 2018. Final version published as submitted by the authors.