The Microblog Public Opinion Analysis Based on the SVM and the LDA Model Combining

Weilin Xu, Zong Zhu, Li Gao and Jinling Liu
Computer and Software Faculty, Huaiyin Institute of Technology, Huaian, jiangsu, 223003.China

Keywords: SVM model; LDA model; Microblog; Public opinion; Clustering.

Abstract. In view of the LDA model, the superiority of long text clustering to Microblog about the use of the user and time to build a long text. According to implied rich semantic information in text and traditional text clustering makes because of the high dimension calculation results inaccurate faults, the SVM model is given and the LDA model combining the similarity, the last is the use of K - Means algorithm for clustering. Combining the experimental results show that the SVM and the LDA model significantly improves the clustering quality and stability.

1. Introduction

With the rapid development of Web 2.0 Internet technology, people's lives are increasingly inseparable from the microblogging. Since the microblogging has such a large user base and a huge amount of information, and have a profound impact on all aspects of people's lives, it is urgent and significant to find valuable hot information through data mining on the microblogging information.

Compared with the traditional text mining method, the LDA model can efficiently complete some basic work such as mining texts to find the potential relationship, to determine the relevance, to do classification and so on. However, it is difficult to use LDA model for microblogging text mining because of two main reasons. First, usually the microblogging text is composed by short text less than 140 words and contains less information. Second, the semantic structure of network language is not standardized resulting in a large noise, which leads the text matrix formed by text mining to be extremely sparse with a very high dimension. In this paper, in order to make the LDA model better to deal with microblogging text, the microblogging information is classified into the long text set with an explicit subject; Then the space vector of SVM feature words are obtained through data preprocessing, feature words extraction and the TF-IDF strategy; Finally using LDA the hot topic of microblogging is clustered. In this paper, the microblogging set is composed into a long text set according to certain rules, and then text are clustered from two aspects of the feature word and the theme using the LDA theme model combined with the SVM which can make up for the deficiency of these two methods and improve the accuracy of clustering.

2. Obtain Public Opinion based on the Combination of LDA and SVM

2.1 Build Long Text Vector Set for Microblogging Text

This paper collected 10000 microblogging information about Jiangsu Huai'an from the website Sina as experimental data. The long text set is constructed according to the microblogging issued by the same user within a certain period of time (four hours in this paper) in chronological order. The specific algorithm is described as follow:

Algorithm 1 construct the long text set for microblogging

Step1 assuming that 10000 microblogging collected in a chronological order uses the method proposed in paper[5] to divide the word, eliminate ambiguity of word, remove stop using words and conjunction, pronoun and so on. After the dimension reduction, the transferred feature vector, is represented as $TS=\{M_i|M_i=(W_{i1},t_{i1};Wi2,t_{i2};\square, Wim,t_{im})|\ i=1,2,\square,n\}$,where W_{ij} is the characteristic word of the text, and t_{ij} is the weight of W_{ij}.

Step2 $TXT=\{MT_{ik}=\Phi|i=1,2,\square,m;k=1,2,\ldots,n\}$ is defined as the long text set constructed, and $TXT=\Phi$,the time interval is T_0,the user set is $USER=\{U_i|i=1,2,\ldots,m\}$.

Step3 DO WHILE TS≠Φ

Step3-1 Select the first vector Mi on left of Ts, judge which users issued M_j, assuming Mi the microblogging issued by the user U_j.

Step3-2 if(the difference of issue time between Mi and the first issued microblogging among MTj is less thanT_0)

Step3-2-1 $MT_{jk}=MT_{jk} \cup \{M_i\}$

Step3-2-2 else

Step3-2-3 $TXT=TXT \cup \{ MT_{jk} \}$

Step3-2-3 $MT_{jk}+1=MT_{jk}+1 \cup \{M_i\}$

Stept3-2-4 endif

Step3-3 $TS=TS-\{M_i\}$

Step3-4 ENDDO

This algorithm constructs a long text vector orderly for microblogging issued by each user at time interval T_0. This is based on the principle that in a certain period of time microblogging published by users only have a small number of themes, and even most of them only have one theme, furthermore assuming that two adjacent microblogging issued by users are more likely to have the same subject.

2.2 Cluster Algorithm based on LDA and SVM

The key of text clustering is the calculation of similarity degree. In this paper, LDA similarity and SVM similarity combined to construct the similarity of clustering algorithm. Assuming text similarity of LDA is SimL, text similarity SVM is SimV, define the similarity of text as:

$$Sim(MT_{ij},MT_{kl})=\lambda SimV+(1-\lambda)SimL \quad (1)$$

Where $0 <\lambda<1$ is the linear correlation coefficient.

The LDA model has ability of dimensionality reduction and can represent text semantic relations. However, as the number of thematic vectors' dimensionality is too low, the text discrimination of the LDA model is weak. It can improve the clustering stability and quality by combining the LDA model with the SVM model. The specific algorithm is described as follows:

Algorithm 2 clustering algorithm combining LDA with SVM

Step1 uses the long text set TXT obtained in Algorithm 1

Step2 uses TF-IDF for SVM modeling [7], computes SimV

Step3 constructs LDA theme modeling [8], calculates SimL based on the subject text similarity

Step4 using formula(1), caculates $Sim(MT_{ij},MT_{kl})=\lambda SimV+(1-\lambda)SimL$

Step5 uses K-Means algorithm to cluster

3. Experimental Results and Analysis

3.1 Linear Correlation Coefficient Λ

The value of λ is set from 0.1 to 0.9, and the missed judgment rate(MJ), the misjudgment rate (EJ) and cost function (CF) value [4] are calculated respectively.C_{EJ} is the missed coefficient,

CEJ is the misjudgment coefficient, Ptarg is the prior probability of the text belonging to a certain category. In order to make a comparison in the same figure, the value of CF and MJ are enlarged 10 times and the value of EJ is magnified 100 times respectively. The value of MJ, EJ and CF are changed with the correlation coefficient λ regularly as shown in Figure 1.

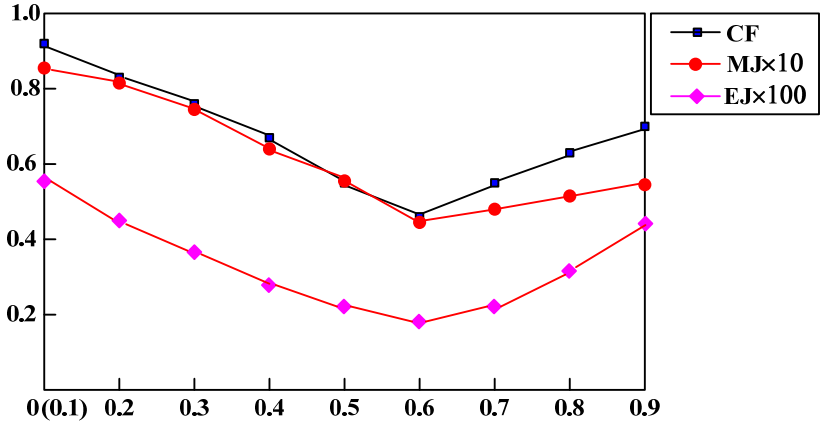

Figure 1. The diagram showing value changes of MJ, EJ and CF according to the correlation coefficient λ

It can be seen from Fig. 1 that the value of the missed judgment rate (MJ) 、 the misjudgment rate (EJ) and cost function (CF) first decrease along with the increasing of the value of λ, and reach the lowest point at 0.6, then increase when the value of λ decreases. Therefore, this experiment takes λ= 0.6 which can achieve the best clustering effect.

3.2 Cluster Quality Testing

Based on the distribution of data sets, the F value is commonly chosen to evaluate the cluster quality by using the integrity rate and the precision rate. The comparison of experiment results represented by F value [9] between SVM model, LDA model and the model combining SVM and LDA (abbreviated as SVM+LDA) are shown in Figure 2.

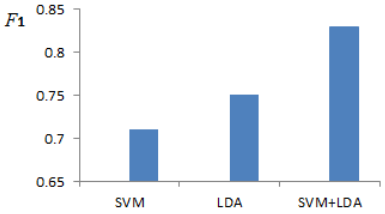

Fig.2 Comparison of clustering results of SVM, LDA and SVM+LDA

It can be seen from Fig. 2 that the clustering results of the SVM+LDA model not only achieves a high level of stability, but also improve the clustering quality. This is because the SVM+LDA model applies the powerful dimensionality reduction ability of the LDA model, strengthens the semantic relation between text and make use of advantages of the SVM model on the characteristics word extraction.

4. Conclusion

The potential semantic relation between texts is an important index of text similarity. LDA model is a probability generation model to solve the potential theme of text. In order to adopt the advantage of LDA model for long text clustering, this paper constructs the long set using microblogging text based on users and time at first, then clusters by the similarity in the SVM+LDA model. By this way, shortcomings of LDA can be overcame, such as the low dimensionality of theme vector and the weak text discrimination, which improves the stability and accuracy of text clustering.

References

[1]. D. M. Blei, “Probabilistic topic models,” Communications of the ACM, vol. 55(4), 2012, pp. 77-84.

[2]. B.Schölkopf, J. Platt, and T. Hofmann, "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model," Advances in Neural Information Processing System, Proceedings of the Twentieth Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December, 2006, pp.241-248..

[3]. T. Hofmann, " Probabilistic latent semantic indexing," Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.
[4]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, 2003,pp. 022.
[5]. L. Hong, and B. D. Davison, "Empirical study of topic modeling in Twitter," Proceedings of the Sigkdd Workshop on Social Media Analytics, 2010, pp. 80-88.
[6]. W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," Paper presented at the In ECIR, 2011.