Short text model based on Strong feature thesaurus

Wentao Lu[1],Yongfeng Huang[1], Xing Li[1], Zhuo Zhang[2], Yingkun Li[2]

[1]Department of Electronic and Engineering, Tsinghua University, Beijing, China

[2]Senate Branch, Section 65053 of The PLA Troops, Dalian, China

luwt11@mails.tsinghua.edu.cn, yfhuang@tsinghua.edu.cn, xing@cenet.edu.cn
820009339@qq.com, luwt12@163.com

Keywords: Short Text Model; Data Sparseness; Strong Feature; Latent Dirichlet Allocation; Clustering

Abstract. Data Sparseness, the evident characteristic of short text, is caused by the diversity of language expression and the short text length. The previous text models represented by Bag of Word (BOW) only considers the statistical feature of words, and thus always underperformed when it comes to short texts. To tackle this problem, we introduced a new text model by combining the statistical method and semantic estimation. Specifically, we managed to obtain the “Strong Feature Thesaurus” through mining process with Latent Dirichlet allocation (LDA) model, and then the semantic information is incorporated in the BOW by weighting those strong feature terms. To assess the performance of this model, we conduct two experiments of the clustering of short text corpuses. The results have shown that our model outperform the prevailing text models such as BOW.

Introduction

With the rapid development of network technology, more and more users want to share their interested information on the network, typical application forms such as blogs, Twitter, social networking services(SNS). The user can communicate more convenient, timely information and express their opinions, resulting in a large number of comments and opinions with personal emotion. Those online messages, which are classified as short texts, all share some common characteristics namely the short message length and intense user participation. Short texts can reach topics of all kinds and are of increasing informational importance.

Modeling method of short text is through the core of all the possible operation on the short text. The name of it can list long classification, similarity computation, short text data mining. Therefore, analysis and application of it has a wide range of public opinion, topic tracking and consumer preference indication.

The information content of short length difference is characteristic of short text, leading to some topic chain is weak. More importantly, because of the diversity of languages, the same theme can be in completely different ways of expression, thus reducing the possibility of the feature in the short text of several different.Therefore, the occurrence of long-term cooperative modeling often fail to improve its accuracy due to sparse data based on short text.

Intensive research has been conducted to solve the data sparseness problem and improve the modeling accuracy of short text. The implicit themes based on X - H phan forward the "bag of words (bow) + modeling method for short text classification [1]theme". The United States tries to short text clustering [introduction "this one concept modeling method and application of arch and the wiki" 2] to solve the problem. Other effective methods including Hu, X "simknow" modeling method is based on clustering Wikipedia and the world [3 article]. Based on the LDA model, a proposed biterm topic model (BTM) of short text topic [4] modeling. Although these studies and consider the semantic information hidden in the feature words, they cannot distinguish between them. We know that the different contribution of different feature often in the themes identified.

In order to further improve the accuracy of short text modeling, we must take into account the semantic importance of certain feature terms. Inspired by the “structure+ average” method in

probabilistic graphical models [5], we managed to propose a new model combining both statistical and semantic information of feature terms. We managed to discriminate the feature terms by putting them into different groups according to their influence on the semantic information of the whole piece of text. Firstly, we established a "strong feature thesaurus" on the basis of Latent Dirichlet Allocation (LDA) model. Then, we put larger weight on feature terms which have significant semantic importance. Thus, the discriminative power of strong feature terms is strengthened. Experimental results suggest that our model improved the purity of short text clustering.

The outline of the paper is as follows. Section II describes the general framework of our new method for short text modeling; Section III discusses the establishment of the "Strong Feature Thesaurus" as well as the procedure of weighting them; Section IV presents our main experimental process and corresponding analysis of results; finally, section V concludes the paper.

The General Framework

There're basically two different approaches in text modeling. One is the traditional BOW, and the other is expands BOW such as "BOW + WordNet". These two approaches are primarily based on the analysis of feature terms' statistical information and their literal meaning. However, the diversity of the language expressions makes it especially difficult to determine semantic meaning of words within context. As a result, these methods share a common problem that the accuracy of modeling is often limited. The sparseness of short text makes it worse that the accuracy of short text modeling is basically lower than that of common texts. In our model, we incorporate the domain knowledge, which is obtained through mining on large datasets. With the help of domain knowledge, we treat the strong feature terms respectively by giving them larger weight. The general framework is depicted in Fig.1.

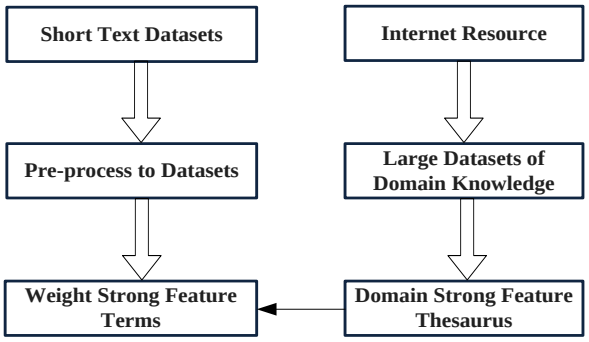

Figure 1. General framework of short text classification

Basically, our new method is composed of three steps. Firstly, we established a "Strong Feature Thesaurus" by screening the datasets of domain knowledge on a large scale. The process is conducted on the basis of LDA model [1][4].

The second procedure is the pre-processing of the short text data, which includes terms segmentation, part-of-speech tagging (POS tagging), part-of-speech choice, frequency statistics, frequency selection and feature selection.

In the third step, we typically gave heavy weight to those feature terms that are included in the "Strong Feature Thesaurus".

There're two core innovations in our method, one is the construction of "Strong Feature Thesaurus", and the other is the weighting process. Strong feature terms are highly semantically-orientated and can be used to determine the category feature of the whole text. Given an example of short text classification, when words like "destroyers", "artillery" and "missiles" appear in a text, we can safely put this text in the military class. In order to apply this thought in statistical method, heavy weights must be given to those strong feature terms in text classification. The terminologies used in this article are presented below:

Definition1: Category set is the collection of predefined classes in classifying short text. It's represented by $C= \{C_j \mid j \in (1, J)\}$.

Definition2: Text set is the collection of short texts to be processed, which is represented by $D= \{Doc_l \mid l \in (1, L)\}$.

Definition3: Feature terms set is the collection of feature terms extracted from Text sets. Expression with T= $\{t_k \mid k \in (1, K)\}$

Definition4: Strong feature term set is the collection of feature terms, which are highly semantic-orientated and are vital in determining the category of the specific short text. Those feature terms are obtained by data mining using domain knowledge. It's represented by F= $\{f_i \mid i \in (1, N)\}$.

Definition5: Topic set is the collection of implicit topics obtained by data mining using domain background knowledge. Represented by TP= $\{topic_m \mid m \in (1, M)\}$.

Definition 6: Contribution of category is dividing the IG of a specific strong feature term by the average IG of all strong feature term. It's represented by

$$Contri(f_i) = IG(f_i) / \sum_{i=1}^{N} IG(f_i) \tag{1}$$

Construct domain strong feature thesaurus and Weighting strong feature terms

A. Construct domain strong feature thesaurus

We extracted different topics and corresponding feature terms from datasets of domain knowledge through LDA model, and then we construct "Strong Feature Thesaurus". The process can be divided into four steps, which is shown in Figure 2.

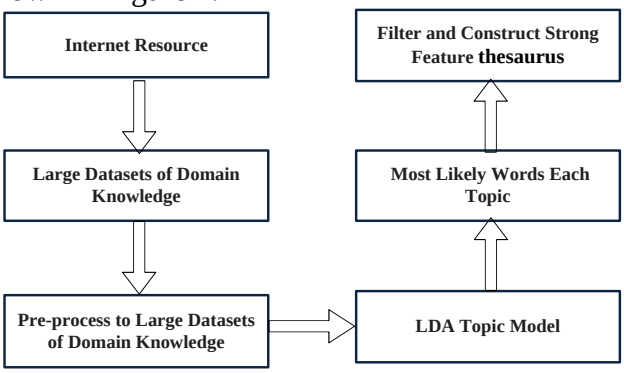

Figure 2. The process of constructing strong feature thesaurus

The first step is to obtain the datasets of domain knowledge. We downloaded different catalogue of web pages from yahoo.com.cn and sohu.com using web crawler, and crawling transaction in each catalogue is limited by 10000 web documents. After filtering out the repeating pages and other noises, the web documents are resolved into pure texts as domain knowledge datasets.

In the second step, we did some pre-processing to the datasets of domain knowledge. We use the ICTCLAS system (developed by Chinese Academy of Science) to conduct Chinese word segmentation and POS tagging. Then we count the frequency of those words.

In the third step, LDA model is used to extract topics from the text. Here, we use "GibbsLDA++" [1], an open source tool. We find 20 feature terms with the largest probability in each topic.

Finally, for feature terms got in step 3, only nouns, verbs and adjectives are left in the text because they convey semantic information. Then any word that appears less than three times would be eliminated from the text. After filtering out some repeated feature terms, we get "Strong Feature Thesaurus".

B. Weighting strong feature terms

Different feature terms might not have the same importance and should be treated respectively. While common statistical methods didn't take this into account, our method gives greater weight to feature terms with greatest discriminative power, and thus the accuracy of text modeling is improved.

After the pre-processing of the short text datasets, we obtained the Vector space model (VSM) expression of short text. In the vector obtained by VSM, we put heavier weight to those feature terms included in the "Strong Feature Thesaurus".

Given Feature terms sets T=$\{t_k \mid k \in (1,K)\}$, Text sets D= $\{Doc_l \mid l \in (1,L)\}$, the weight of t_k in Doc_l is set w_{kl}, Then VSM expression of any text can be represented by Doc_l=$\{w_{kl} \mid k \in (1,K), l \in$

(1,L)}. Given strong feature terms sets F= {f_i | $i \in (1, N)$}, if $t_k \in$ F in Doc_l, then t_k is weighted by 1+Contri (f_i). The weighting formula is

$$New(w_{kl}) = (1 + Contri(f_i))w_{kl} \quad (2)$$

Evaluation

In order to evaluate effect of our proposed method, we conduct clustering experiments in two short text datasets. Experiments results prove that our method is more effective than baseline method.

A. Data Sets

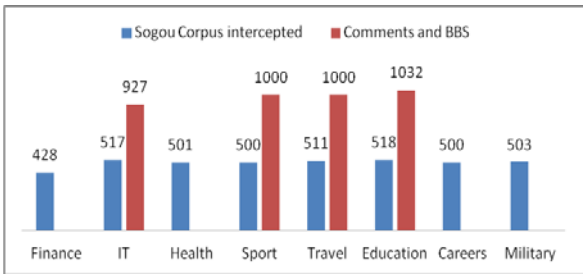

Figure 3. categories and distribution of the two datasets

Due to the absence of standard Chinese short text datasets, we managed to obtain the data in two different ways. The first one involves the application of web spider; we downloaded many comments from BBS and some commercial sites. However, these texts are not representative enough. Secondly, we extracted short text data from Sogou Corpus [8] (a well-known Chinese corpus). Generally, the truncated texts are more difficult in text modeling, because they are more semantically incomplete. The categories and distribution of the two datasets are given in Fig.3.

B.Clustering Methods

The web documents are downloaded using Wget (a stable web spider), parsed by Htmlparser, processed by Python, and then clustered by Expectation maximization (EM) interface functions in Weka [7]. Two text models, BOW and BOW+WSF, are used in the text processing step. The two text models are defined as follows:

BOW: it refers to the "bag of words" model with the TF weighting method.

BOW+WSF: an expansion of BOW model, which includes the weighting of strong feature terms. Here, "WSF" refers to "Weighting of Strong Feature Term".

To facilitate evaluation, a confusion matrix can be constructed from the resulting clusters. From the matrix, various measurements can be computed. We choose "Cluster Purity" as evaluation indicator.

Cluster Purity [6]: This measures the extent that a cluster contains only one class of data. For C= {C_j | $j \in$ (1, J)}, the clustering method also produces J clusters, which partition D= {Doc_l | $l \in$ (1, L)} into J disjoint subsets, D_1, D_2, …, D_J. The purity of each cluster is computed with

$$purity(D_l) = \max_j(\Pr_l(C_j)) \quad (3)$$

Where $\Pr_l(C_j)$ is the proportion of class C_j data points in cluster l or D_l[6].

The total purity of the whole clustering is

$$purity_{total}(D_l) = \sum_{l=1}^{k} \frac{D_l}{D} purity(D_l) \quad (4)$$

In order to construct strong feature terms, as mentioned method in III, we choose corresponding eight class data, 2000 text are selected in each class data, to form a domain knowledge sets. We apply "GibbsLDA++" to extract strong feature terms. According to experience [4], we set α=50/Z

and β=0.1, and choose 10 topics in each class and got 80 topics in sum. After getting topic-probability distribution of feature terms, we choose the 20 feature terms with maximum probability in each topic. Totally, we got 1600 feature terms.

After getting 1600 feature terms, according to POS tagging, only nouns, verbs and adjectives are left since they convey semantic information. Then, any word that appears less than three times would be eliminated. At last, we get a "Strong Feature Thesaurus" composed of 1086 strong feature terms.

C. Result and Analysis

The dataset collected from BBS and commercial sites can be divided into four categories. The texts of Sogou Corpus can be divided into 8 categories, and 56 characters are extracted from each passage. We tried to cluster these short texts using the two text models. The experimental results are shown in Fig.5.

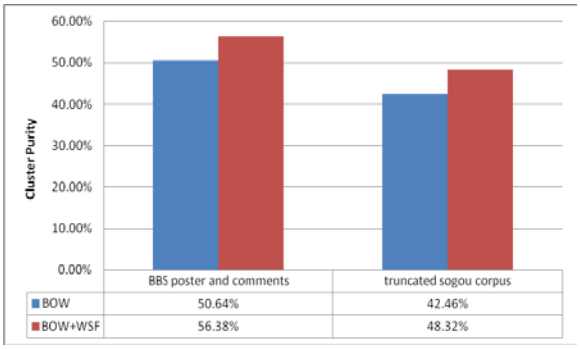

Figure 4. Cluster result for Comments and BBS

As we can see from Fig.4, BOW+WSF perform better than BOW in short text clustering. For BBS datasets, cluster purity of BOW model is 50.64%, the cluster purity of BOW + WSF model is 56.38%, our method increase clustering purity from 50.64% to 56.38%, increased 5.74% of clustering purity. As for the datasets obtained from the Sogou Corpus, clustering purity of BOW model is 42.46% and clustering purity of BOW+WSF is 48.32%, and 5.86% of clustering purity is increased in our methods.

While the clustering purity varied in two different short text corpuses, the improvement of the purity by BOW+WSF stayed the same. That is to say, our model can still provide great improvement even when the clustering purity reaches a rather high level. This is because the "Strong Feature Thesaurus" obtained through operations on the domain knowledge has greater discriminative power. Thus, the incorporation of semantic information by weighting those feature terms improve the efficiency of clustering regardless of the purity.

Cluster number is a parameter of great importance in the clustering process, and should be chosen carefully to get the best result. In the experiment, we use the number of categories as center and diameter. The influence of cluster number on purity is tested; eventually we selected the cluster numbers which have the highest purity as clustering parameters. The experimental results are shown in Fig.5

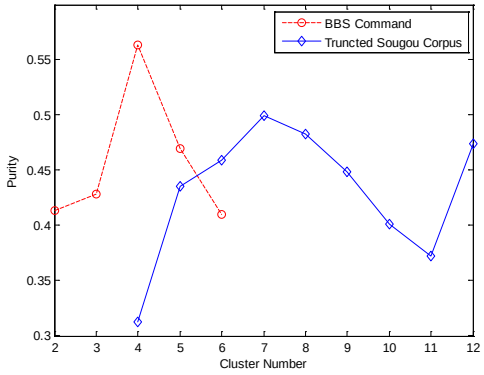

Figure 5. Classification result for Sogou Corpus intercepted

From the Fig.5 shown above, as for the four categories of BBS comment data, the clustering purity is maximized when the cluster number equals to 4. And for the 8 categories of text data from Sougou Corpus, the clustering purity is maximized when the cluster number is equal to 7. We can deduce from above that the clustering purity reaches highest point when the cluster number is close to the number of categories. And by choosing the cluster number respectively, we can further improve the clustering result.

Conclusion

In order to solve the sparseness problem in short text, this paper presents a new text modeling method combining both statistical and semantic information. Our innovative investigation basically lies in two aspects. Firstly, unlike previous work, we consider the difference of sematic influence of feature terms on the whole text. Secondly, we adopt the LDA model to obtain the domain knowledge and furthermore the "Strong Feature Thesaurus".

Theoretical analysis has proved that our text model can reach higher accuracy, and it's confirmed by the experiment of short text clustering. Specifically, when the purity of clustering changed, the efficiency of our model kept steady. However, due to the lack of training data, our experimental dataset is incomplete, thus there's still room for improvement. We will further expand the scale of the experiment, and apply the method to data of all kind. This model, however, can be applied to short text classification and similarity calculation, which might also appear in our future research.

Acknowledgements

The work was fully supported by the grant No.2007CB310806 and No.2008BAH37B07.

Reference

[1] X.-H. Phan, Le-Minh Nguyen and S. Horiguchi, "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections", Proceeding of the 17th international conference on World Wide Web, ACM, New York, USA, pp.91-100, April 2008.

[2] S. Banerjee, K. Ramanathan, and A. Gupta. "Clustering short texts using Wikipedia", Proceedings of the 30th ACM SIGIR, pages 787–788, 2007.

[3] Hu, X., Sun, N., Zhang, C., and Chua, T.-S, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," Proceeding of the 18th ACM conference on Information and knowledge management, ACM, New York, USA, pp. 919-928, Nov. 2009.

[4] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo, "BTM: Topic Modeling over Short Texts", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 26(12):2928-2941, 2014

[5] Daphne Koller,Nir Friedman. "Probabilistic Graphical Models-Principles and Techniques", The MIT Press.2009-8-31

[6] Bing Liu, "web Data Mining: Exploring Hyperlinks, Contents, and Usage Data" Springer. July 2011.

[7] I. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, January 2011.

[8] http://www.sogou.com/labs/dl/c.html.