Short text semantic feature extension and classification based on LDA

Liu-tong Xu, Zhao XUE* and Hai Huang

Beijing University of Posts and Telecommunications, Cloud Computing and Service Computing Center, 100876, Beijing, China

*Email: susannahermione@163.com

Abstract. To solve the problem of feature sparseness of short texts, we studied the application of LDA Topic Model on feature extension and classification of short texts. Training LDA on external long texts related to short texts, and achieving the inference and extension of short texts’ topics based on LDA solves the feature sparseness of short texts and improves the accuracy of classification effectively. Latent semantic information in LDA can also effectively improve the interpretability of short texts.

1. Introduction

With the rapid development of information technology, we receive more and more information in daily life. At the same time, information technology has also brought some negative issues, such as “information overload” phenomenon-A lot of redundant information makes it inefficient in choosing valuable information. In order to better solve this problem, search engine technology is used to filter information, which makes the time required to obtain useful information greatly reduced. However, search engine is not perfect, it still has many problems, for example, one of the obvious ones is that users themselves are not very clear about what information they need in many situations, and they also can’t accurately describe the information they need by keywords or phrases. At the same time, users’ needs for information are constantly changing, even the same keyword may correspond to different information needs, so it is necessary to classify and identify the information needs which are described by texts. In this way, it can greatly narrow the scope of information retrieval. However, the efficiency of manual labelling is too low, and the accuracy often fails to meet usage requirements, therefore classifying texts by computer algorithms has been proposed.

With the rapid development of Internet and mobile terminal technology, the increasing number of varied types of information represented by short texts containing rich information have emerged, such as short messages. Its application has also gradually expanded to different types of information systems. Therefore, it is of great practical significance to study the short text classification methods which can lead to scientifically obtaining more and more valuable information and providing users with better information services.

Text classification mainly focused on the processing of long texts in the early stage of research. As the research further develops, some algorithms have achieved good results. However, compared with long texts, short texts are different from them, the differences specifically embody in the following aspects: firstly, short text has less characters, so its feature is sparse and it may contain more noise information; secondly, the dynamic characteristics of short texts are more obvious, which leads to
frequent update. Therefore, it is difficult to effectively process short texts using the classification methods which are currently applied to long texts. The technologies that have emerged are still not mature enough, which makes the application subject to certain restrictions.

2. Related work

Text classification has a long history. Back in the 1950s, Luhn H. P [1] proposed the concept of word frequency, which treats a text as a collection of words, then uses the word frequency to measure its importance. The earliest text classification methods mainly use customized classification rules, whose classification accuracy depends on the familiarity with the classification interval, and this makes classification difficult and its efficiency can’t be guaranteed.

With the popularity of Internet applications, there have been massive text data, algorithms for classifying large-scale text data have also appeared and achieved good results. Among them, machine learning algorithms have been widely used because of high accuracy and intelligence. Salton [2] et al. designed a vector space model and applied it to text classification; Lewis [3] made a survey on text classification; Yang Yiming [4] studied the feature selection through experiments and compared different methods, such as Information Gain. At the end of the last century, Cortes [5] et al. proposed the Support Vector Machine (SVM) algorithm based on statistical theory, this method gradually attracted people’s attention because of its good performance. On the basis of SVM, some experts have improved it so that its application has been further expanded. For example, the SVM-KNN proposed by Li Rong [6], successfully combined SVM with KNN; Liu Chunwei [7] made more improvements on SVM, and designed the PSO-SVM algorithm which used Particle Swarm Optimization to determine the parameters of SVM.

In the process of the rapid development of the Internet, there are more and more short text messages such as news, blogs, and chat records, how to process them and provide more valuable information has become a hot research in the field of text mining, experts in this field have conducted a lot of research work, and some algorithms or systems have also been applied to the processing of short text information. Rafiei [8] proposed a feature-based short text classification method, which makes information filtering effective; Sahami [9] proposed a search engine-based method to calculate the similarity between short texts; Xuan-Hieu Phan [10] et al. mainly adopted a method for obtaining hidden topics in the classification of short texts and achieved good results.

The current research and application of short text classification is mainly based on English texts, which can’t be directly applied to Chinese language materials. Therefore, Chinese experts are also conducting research in this area. Fan Xinghua [11] proposed a short text feature extension method based on co-occurrence and achieved a certain degree of effect in short text classification; Wang Xiwei [12] proposed another short text feature extension method based on context semantics.

As experts are continuing to deepen research on short text classification, there are more and more new methods. The LDA topic model is the typical one among them. Blei [13] first proposed LDA in 2003 and applied it to text classification. It was originally mainly used in the field of mathematics, however, with the deepening of its research, its application has gradually expanded to more areas, and currently it has been applied to data mining processing. Lu Y [14] designed a topic- emotion mixed LDA model which divides words into ordinary words and words with emotional tendencies. The latter can be further subdivided into positive, negative, and middle categories, then it can generate the document’s implicit semantic vector representation.

3. Our work

Compared with long text, short text has fewer characters and contains more noise information. Dynamic changes are obvious which leads short text to be updated frequently. All of these make short text classification still need more researches.

For long texts, after segmenting a large amount of text, the corpus will form a large dictionary. Then after filtering the stopwords, the dictionary and long texts still contain many words, and the text feature vector based on word frequency is still in high dimension. However, short text often contains
only dozens of characters, even a dictionary generated by a large amount of short text, its scale is much smaller. After filtering the stopwords, only a few feature words remain, this will make fewer words in the dictionary and short texts which leads to “dimension disaster” phenomenon. Feature sparseness is a very obvious problem for short texts.

Feature extension can solve the sparseness problem. Text feature extension methods are divided into knowledge bases based [15] and LDA based [16-17]. The former mainly uses the currently widely used search engines such as Baidu to search and introduce relevant information of feature words. This method can increase short text features, and make its meaning clearer. However, its disadvantage is that the effect is too dependent on the rationality of the knowledge base. Currently, with the development of LDA, how to apply it to text feature extension has become a popular research. Our paper will conduct related research work in LDA, using the latent semantic information of long texts related with short texts to implement the feature extension and classification of short text.

3.1. LDA
In recent years, with the development and improvement of LDA, its application in text classification is also increasing. Parameter estimation of LDA usually adopts Gibbs sampling, then gets a text-topic probability distribution vector $\theta$ and a topic-word probability distribution vector $\varphi$, $\theta_i$ represents the probability that the text implies a topic. The larger the value, the better the match between the text and the topic; $\varphi_j$ represents the probability that a word appears on a topic. The larger the value, the greater the probability that the word appears on this topic, the top $M$ words are usually selected according to the order of the probability to represent the topic. The generative graphical model of LDA is shown in figure 1.

![Figure 1. LDA: a generative graphical model.](image)

From the generative graphical model depicted in figure 1, we can write the joint distribution of all known and hidden variables given the Dirichlet parameters which is shown in table 1.

| Table 1. Generation process for LDA. |
|-------------------------------------|
| ---                                |
| for all topics $\in [1, K]$ do      |
| sample mixture components $\varphi_k \sim Dir(\beta)$ |
| end for                             |
| for all documents $m \in [1, M]$ do |
| sample mixture proportion $\beta_m \sim Dir(\alpha)$ |
| sample document length $N_m \sim Poiss(\varsigma)$ |
| for all words $n \in [1, N_m]$ do   |
| sample topic index $z_{m,n} \sim Mult(\beta_m)$ |
| sample term for word $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$ |
| end for                             |
end for

**Parameters and variables:**
- \( M \): the total number of documents
- \( K \): the number of (hidden/latent) topics
- \( V \): vocabulary size
- \( \alpha, \beta \): Dirichlet parameters
- \( \vartheta_m \): topic distribution for document \( m \)
  \( \theta = \{ \vartheta_m \}_{m=1}^M \): a \( M \times K \) matrix
- \( \varphi \): word distribution for topic
  \( \phi = \{ \varphi_k \}_{k=1}^K \): a \( K \times V \) matrix
- \( N_m \): the length of document \( m \)
- \( z_{m,n} \): topic index of the \( n \)th word in document \( m \)
- \( w_{m,n} \): a particular word for word placeholder \([m,n]\)

LDA uses the text-topic probability distribution vector \( \theta \) as the feature vector of the text, and implements classification based on it. But for short text, the length of the text itself is small, some of the topics could not reflect specific meaning. So, some methods [18] suggest that a large number of external related texts can be directly used to train LDA, then implement the feature extension of short text according to the topic information of this model. However, it is difficult to ensure that the extended added topic is completely consistent with the meaning of the original text. Some other methods [19-20] are applied to train LDA using a mixed topic model, but this makes the training time increase and the efficiency decline, and the number of topics \( K \) is manually set, corresponding optimal \( K \) of different corpus is often different, so it is difficult to determine the optimal number of mixed multi-topics, and these deficiencies have made the application limited. The above methods of short text classification based on LDA still have obvious deficiencies, summarized as the following aspects:

1) There is no filtering on unwanted topics. Redundant noise information affects classification accuracy.

2) Direct modeling of short texts at the level of implicit semantic information but ignoring the influence of word features on the statistical level, which causes the accuracy of classification to be excessively concentrated on the effects of LDA. Factors that influence the results are lopsided.

3) For multi-topic LDA models, more training time is required. Especially with the increase in the size of corpus, the time spent on the training will also increase significantly.

4) How to determine a reasonable number of topics is also an important issue. It takes a constant trial to determine the optimal value, which makes its application conditions too harsh.

### 3.2. Feature extension and classification

The improvement of the short text classification algorithm in our work is mainly to use the LDA built on the external related long texts to achieve the topic extension of short texts. The details are as follows: Infer the related topics of short texts using the LDA obtained on external related long texts, filter and remove unwanted topics, then calculate the text-topic probability distribution vector \( \theta \) and the topic-word probability distribution vector \( \varphi \) for short texts, and then introduce topics on external related texts that meet the condition by comparing with the threshold and introduce them to the short text representation vector. In this way, the semantic features of short texts are extended. As described above, the steps of the Chinese short text classification algorithm based on LDA Topic Model proposed in our work are as follows:

**Step 1:** A pre-processed external related long text set is used to obtain a LDA topic model;

**Step 2:** After the short text is preprocessed, infer the relevant topics of the short texts including training and testing set based on the model obtained in the first step, remove unwanted topics, and calculate the probability distribution \( \theta \) and \( \varphi \);
Step 3: Extend features based on the result of comparing $\theta_i$ with the threshold, and determine the final text-topic probability distribution vector $\theta$ of short text;

Step 4: Train the parameters of the classifier with the training short texts and obtain the classifier, then verify the performance of the classifier on the testing set.

The corresponding flow chart of the above steps is shown in figure 2.

Figure 2. Flow chart of short text feature extension and classification algorithm based on LDA.

In our work, we select the classifier based on neural network [21], it uses short text representation vector $\theta$ as input, and gets the probability that the short text belongs to a certain class. The class with the highest probability is the one which the short text is determined belong to.

4. Experiments

4.1. Data

In order to verify the performance of algorithm, we conducted related experiments. The short texts data used in our work is mainly a short news corpus crawled from China News Online, including eight classes of politics, sports, finance, culture, education, military, technology and entertainment from 2016.12-2017.12. The average number of texts per category in training and testing sets is 10,000, and the total amount of texts is about 160,000. The external related long texts set is Sogou network-wide news data (SogouCA).

4.2. Evaluation

Evaluation metrics is an important basis for measuring the rationality of the algorithm which reflects the performance of the classification model. Through quantitative analysis of evaluation metrics, the advantages, disadvantages, and scope of application of the model can be directly derived, and establish a baseline for further improvement. At present, accuracy, recall, and precision are often used. By quantitatively analyzing the evaluation metrics, the experimental results can be accurately described. The definitions of each evaluation metrics are as follows.

(1) Accuracy

Accuracy is defined as the ratio of the number of correctly classified texts to the total number of texts, its calculation formula is as follows, $TP$, $TN$, $FP$, and $FN$ respectively mean the number of correct positive samples, the number of correct negative samples, the number of wrong positive samples, and the number of wrong negative samples. For classification models, it should have a high accuracy.

---

1 http://www.chinanews.com
2 http://www.sogou.com/labs/resource/ca.php
Recall & precision

Recall indicates the ratio of the number of documents that are classified into the class $c_i$ by the classifier to the number of real documents within the class $c_i$. Its calculation formula is as follows.

$$ R = \frac{TP}{TP+FN} \quad (2) $$

Precision indicates the ratio of the number of documents correctly classified into the class $c_i$ and the number of documents determined to be $c_i$ by the classifier. Its calculation formula is as follows.

$$ P = \frac{TP}{TP+FP} \quad (3) $$

$F1$- Measure

Recall and precision are complementary relationships. One increases, the other may decrease. So, using these two metrics to evaluate the overall performance of the classification model is often difficult to achieve good results. At this point, we can choose $F1$- Measure which comprehensively considers rate and precision. Its calculation formula is as follows.

$$ F1 = \frac{2 \times P \times R}{P+R} \quad (4) $$

Micro-R, Micro-P, Micro-$F1$

The above three metrics mainly evaluate the performance of the classification model on each class in the corpus, but does not evaluate its overall classification effect. At this point, we can choose to use Micro-R, Micro-P, Micro-$F1$ metrics to analyse the performance of the classification model.

The calculation formula for Micro-$R$ is as follows.

$$ Micro - R = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i+FN_i)} \quad (5) $$

The calculation formula for Micro-$P$ is as follows.

$$ Micro - P = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i+FP_i)} \quad (6) $$

The calculation formula for Micro-$F1$ is as follows.

$$ Micro - F1 = \frac{2 \times (Micro - R) \times (Micro - P)}{(Micro - R) + (Micro - P)} \quad (7) $$

4.3. Results

Our work mainly extends topic information in the previous short text classification algorithm. It can better solve the problem of sparse semantic features of short texts and improve the interpretability of original texts. In order to verify the performance of the improved algorithm, the following experiment was conducted.

Previous research has confirmed that when LDA is not used, DF feature selection method has been able to achieve good results. We compare the performance of DF feature selection method, the document-topic probability distribution method, and the improved feature extension method in this paper. The specific experimental results are shown in figure 3.

![Figure 3. Comparison histogram of classification results of different algorithms.](image-url)
5. Conclusion
In this paper, we improved the existing short text classification algorithm in feature extension by applying the LDA Topic Model to the feature extension and selection. First, LDA is trained on the external related long texts, and then the short text is represented as a text-topic probability distribution vector θ based on it. Then, by comparing θ_i with the threshold, the topic obtained on the long texts satisfying the condition is introduced into the short text representation vector θ. In this way, the semantic information of short text is enriched, the feature extension of short text is realized, and it can get better results in Chinese short text classification.

References
[1] Luhn H P. The automatic creation of literature abstracts[M]. IBM Corp. (1958).
[2] Salton G. A vector space model for automatic indexing[J]. Communications of the ACM, (1974), 18(11):613-620.
[3] Lewis D D. Feature selection and feature extraction for text categorization[C]// The Workshop on Speech & Natural Language. Association for Computational Linguistics, (2013):212–217.
[4] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]// Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. (1997):412-420.
[5] Cortes C, Vapnik V. Support-Vector Networks. [J]. Machine Learning, (1995), 20(3):273-297.
[6] Rong L I, Sun Y. Application of SVM-KNN Classifier into Web Page Classification[J]. Science Technology & Engineering, (2009).
[7] Liu C W, Luo J X. A PSO-SVM Classifier Based on Hybrid Kernel Function[J]. Journal of East China University of Science & Technology, (2014), 40(1):96-101.
[8] Rafiei D, Bharat K, Shukla A. Diversifying web search results[C]// International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April. DBLP, (2010):781-790.
[9] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets[C]// International Conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May. DBLP, (2006):377-386.
[10] Xuan-Hieu Phan, Le-Minh Nguyen, et al., Learning to classify short and sparse text & web with hidden topics from large-scale data collections [C]. In Proceedings of the 17th International Conference on World Wide Web, New York, ACM Press, (2008),91-100.
[11] Fan X, Hu H. A New Model for Chinese Short-text Classification Considering Feature Extension[C]// International Conference on Artificial Intelligence and Computational Intelligence. IEEE, (2010):7-11.
[12] Wang X W, Fan X H, Zhao J. Method for Chinese short text classification based on feature extension: Method for Chinese short text classification based on feature extension[J]. Journal of Computer Applications, (2009), 29(3):843-845.
[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. J Machine Learning Research Archive, (2003), 3:993-1022.
[14] Lu Y, Mei Q, Zhai C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA[J]. Information Retrieval, (2011), 14(2):178-203.
[15] Zhao H, Liu H. Classification Algorithm of Chinese Short Texts Based on Wikipedia[J]. Library & Information Service, (2013), 57(11):120-124.
[16] Hu Y. A New Method of Key Words Extraction for Chinese Short-text Classification[J]. New Technology of Library & Information Service, (2013).
[17] Zhang Peijing, Song Lei. Overview on Topic Modeling of Microblogs Text Based on LDA [J]. Library and Information Service, (2012), 56(24): 120-126.
[18] L Guo, Y Li, D Mu, T Yang, Z Li. A LDA Model Based Topic Detection Method[J]. Journal of Northwestern Polytechnical University, (2016),34(4):698-702.
[19] Chen M, Shen D, Shen D. Short text classification improved by learning multi-granularity
[20] Guan P, Wang Y. Identifying Optimal Topic Numbers from Sci-Tech Information with LDA Model[J]. New Technology of Library & Information Service, (2016).

[21] H.K. Lam, Udeme Ekong, Hongbin Liu. A study of neural-network-based classifiers for material classification[J]. Neurocomputing, (2014), 144(1):367-377.