The Research of Chinese Short-text Classification Based on Domain Keyword Set Extension and HowNet *

Xiangdong Li
School of Information Management, Center for Studies of Information Resources
Wuhan University
Wuhan 430072, China
xli_hotmail@hotmail.com

Fan Gao, Cong Ding
School of Information Management
Wuhan University
Wuhan 430072, China
gaofan_whu@foxmail.com, 302065219@qq.com

***Abstract*—To implement feature extension of short text and improve short text classification performance, this paper extracts the high frequency words and topic core words of each class of the training set as domain keyword set based on two different feature granularity, which are keyword and latent topic, and derives the topic probability distribution of the test text using LDA model, while some topic probability is greater than a certain threshold, extends the keywords of the topic into the testing text. Calculate the semantic similarity of the test text and the domain keyword set for each category by using HowNet. Experimental results show that the method proposed in this paper can effectively improve the short-text classification performance.**

Keywords—short-text classification, keyword set, LDA, feature extension, HowNet

I. INTRODUCTION

Unlike traditional documents, short-text doesn't have enough contextual information, traditional Vector Space Model (VSM) could not be applied properly in short-text classification. In order to solve the problem of sparse feature in short-text, the researchers usually extend features of short-text by using external resource or internal semantic correlation.

This paper proposes a Chinese short-text classification algorithm based on domain keyword set extension and HowNet by combining with the different granularities - keywords and latent topics. The rest of this paper is organized as follows: Section2 reviews related work. Section3 introduces the proposed method. Section4 presents the experimental results before the final section concludes the paper.

II. RESEARCH STATUS

The feature extension method of short-text is generally divided into external resources extension and internal semantic association rules extension. The main external resource extension usually uses WordNet, HowNet, Wikipedia and other semantic dictionary. Zhao etc.[1]used Wikipedia to dig hidden information of short-text, through selecting the words which is strongly associated with the keyword at the semantic level, and extended the keywords to help put the text to be classified in the proper category. Zhang etc.[2]used HowNet to extract text keywords of concept mapping, extended the semantic expression ability. On the other hand, through internal semantic association rules, using the correlation of corpus to build domain keyword sets, also could realize the feature extension of the short-text. Ning[3]extracted field high-frequency words as keywords, with the help of HowNet to extend short-text. Cham etc.[4]used domain high-frequency words as keyword set to extend the feature. Hu etc.[5] extended the high-frequency words in the latent topic to short-text. Sriram etc.[6]took user's information and microblog characteristics as domain keyword sets, broaden the character to implement classification.

It can be found that the existing research in short-text feature extension based on internal relationship mainly in fine-grained or coarse-grained to establish the knowledge base and realize the short-text feature extension. There are few articles considering two kinds of granularity at the same time. Therefore, our method has a certain innovation significance for the improvement of short text classification performance, because it is based on HowNet and domain keyword set extension combined with the keywords and latent topics of two different granularities. [7]

III. SHORT-TEXT CLASSIFICATION BASED ON HOWNET AND DOMAIN KEYWORD SET EXTENSION

A. LDA model

LDA[8] is a probabilistic topic model. In this model, text is expressed as a three-layer probability model, which consists of text, topics and keywords. Each document representation for random mixed distribution on the latent topic sets, and each topic are expressed as the multinomial distribution of keywords, each topic is composed of a series of keywords.

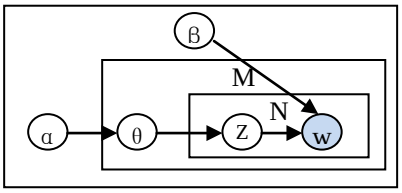

Fig. 1. LDA model

Project "Research on Automatic Classification of Multiple Types of Text Digital Resources" (No. 15BTQ066) supported by National Social Science Foundation of China.

B. Domain keyword set

Domain keyword set is a keyword collection which has strong indicative and good distinguish ability in a field. It could effectively capture domain knowledge and precisely distinguish the other areas. Feature granularity is the extent of the characters of text content [9]. This paper extracts keywords for each category of training set based on different granularities: based on fine granularity, get domain high-frequency words through keyword frequency statistics; based on coarse granularity, get the topic core words for each domain through LDA modeling.

1) domain high-frequency words

Domain high-frequency words refer to the words that have high probability of occurring in a certain field and have small probability of occurring in other areas. The category recognition ability of low-frequency words is low, it will increase the noise of short-text and influence the classification effect. This paper extracts the high-frequency words and removes the low-frequency words. Algorithm is detailed as follows:

Input: training set **D,** categories of training sets **K**, the threshold of the proportion of keywords and categories **Weight**.
Output: domain keywords for each category.

a) Filter the training set according to part of speech, only retain nouns, verbs and adjectives which have greater influence on the classification;
b) Keyword frequency statistics of the filtered training set, obtain the frequency of each keyword in each category, filter the keywords whose frequency is less than 5;
c) Normalize keyword frequency in each category according to the proportion, assume the frequency of keyword w in category C_i is x_i, the proportion of w in C_i is $x_i / \sum_{i=1}^{K} x_i$.

2) Topic core words

Topic core word is a keywords collection which could represent the theme under a certain theme. With categories as the topic, get the core words of each topic in training sets by LDA modeling. Subject core word extraction algorithm is as follows:

Input: training set **D**, categories of training sets **K**, the keyword number threshold **M**
Output: topic core words for each category

a) Filter the training set according to part of speech, only retain nouns, verbs and adjectives which have greater influence on the classification;
b) With categories as the topic, modeling the filtered training sets by LDA, and get the keyword distribution of each topic;
c) On keyword probability distribution of a topic, select probability value top M keywords as the core keywords of the topic.

Combine domain high-frequency words and topic core words, remove duplicate words, then we get the domain keywords set for each category of training set.

C. The feature extension based on domain keyword set

To solve the problems of feature sparse and weak information description in short-text, this paper proposes a domain keyword extension algorithm based on LDA. Specific algorithm is as follows:

Input: domain keyword set of training set **S**, test set **D**, topic probability threshold**C2**, number of keywords extraction threshold **N**
Output: the extended test set

a) Filter the test set according part of speech, only retain nouns, verbs and adjectives which have greater influence on the classification;
b) Modeling the filtered test sets by LDA to get a test text - topic probability distribution;
c) For any text to be classified **d** and topic distribution θ, if the topic probability in text to be classified **d** is greater than the threshold value **C2**, then expand the corresponding domain keywords of topic **i** and the several most relevant keywords of the text to **d** intensively. Relevance computation formula is as follows:

$$Cor(w,d) = \max_{1 \le i \le n} Sim(w, w_i) \quad (1)$$

$Cor(w,d)$ is the correlation between keyword **w** and text **d**, $Sim(w, w_i)$ is the semantic similarity between keyword w and w_i, **n** is the length of text **d**.

d) If all the topics probability in the short-text **d** is less than the threshold **C2**, continue to perform the next text.

Expanding training set associated keywords to test set text could increase the relevance between the test set and training set, thus the effect of classification will be improved.

D. Text similarity caculation based on Hownet

Considering the short text has the characteristics of feature sparse, this paper calculates the semantic similarity between short-texts with the help of HowNet semantic dictionary. Because text is made up of words, the similarity between the short-texts can be converted to the similarity between words.

In HowNet, each word contains several concepts, therefore word similarity calculation could be converted into the similarity calculation between concepts. Meanwhile sememe is the basic unit of concepts, so the similarity between concepts can be converted into the similarity between sememes.

This article uses the formula (2) proposed by Wu etc.[10]to calculate similarity between sememes.

$$Sim(p1, p2) = \frac{\alpha \times \min(depth_{p1}, depth_{p2})}{\alpha \times \min(depth_{p1}, depth_{p2}) + d} \quad (2)$$

The calculation formula of concept similarity is as follows:[11]

$$Sim(s1, s2) = \sum_{i=1}^{4} \beta_i \times sim_i(s1, s2) \quad (3)$$

For the word w_1 and w_2, the similarity of w_1 and w_2 is the max value of all the concept similarities. The formula is as follows:[12]

$$Sim(w_1, w_2) = \max_{1 \le i \le m, 1 \le j \le n} sim_i(s_{1i}, s_{2j}) \quad (4)$$

For short-text d_1 and d_2, The similarity calculation method between d_1 and d_2 is:[13]

$$Sim(d_1, d_2) = \frac{\sum_{i=1}^{m} \max sim(w_{1i}, w_2)}{2m} + \frac{\sum_{i=1}^{n} \max sim(w_{2i}, w_1)}{2n} \quad (5)$$

Please see the original article for the meaning of the symbols in the formula (2)--(5).

E. *Short-text classification based on domain keyword set extension and HowNet*

To solve the feature sparse problem of short-text, this paper proposes a short-text classification algorithm based on HowNet and domain keyword set extension. Classification framework is shown in figure 2:

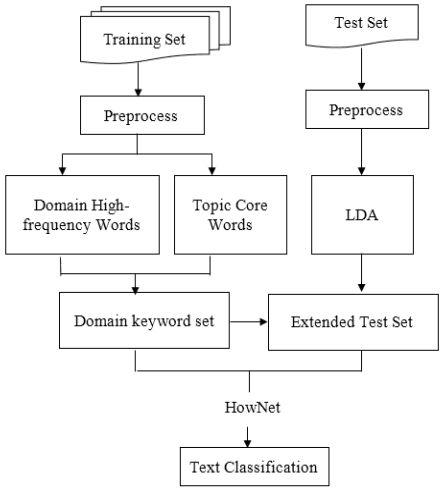

Fig. 2. the framework of short-text classification based on Hownet and domain keyword set extension

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Experimental corpus*

In this paper, the experimental corpus are derived from Fudan University Chinese Corpus[14], Sogou Text Classification Corpus[15] and Sina Weibo Corpus[16]. For two kinds of public corpus, respectively extract headlines from journal articles in Fudan University Chinese Corpus and Sogou news as the experimental data, average length of texts is about twenty words. Fudan University Chinese Corpus, including art, IT, economy, environment, history, aviation, sports 7 classes, a total of 4708 texts, 2979 of them is training set, 1729 of them is test set; Sogou Text Classification Corpus, including IT, economy, education, military, sports, tourism, medicine 7 classes, a total of 3005 texts, 1977 of them is training set, 1028 of them is test set. This article also grabs Weibo texts from the NLPIR Weibo content corpus, which includes tourism, sports, military, science, politics, 5 categories, we randomly selected 2335 texts as the experiment material, 1427 of them is training set, 908 of them is test set. In this paper, all experimental data are random extracted. We repeat the experiment 10 times, take the average as the final result.

B. *Performance evaluation*

To verify our method is effective, on the basis of precision P, recall ratio R and F1 of three traditional indicators, this paper uses Macro_F1 and Micro_F1 to evaluate the classification performance [17].

C. *Experiment result analysis*

For high-frequency words in this article, we had set the categories proportion of weight threshold in training set as 0.8, topic core word M threshold is set to 20, test set text extension topic probability threshold C2 set as 0.5, keyword extraction threshold N is set to 5.

The part of the keywords which extract from domain keyword set in Fudan Corpus training set for each category as shown in table 1:

TABLE I. KEYWORDS SET EXAMPLE IN FUDAN TRAINING SET

Category	Keywords
Art	Poetry, Literature and Art, Aesthetics, Artist, Drama
IT	Computer, Database, System, Machine, Operation, Technology
Economy	Economics, Market economy, Increase, Institution
Environment	Environment, Pollution, Atmosphere, Soil, Geology, Cymolite
History	History, Historic, New, China
Aviation	Damping, Heat transfer, Aero-engine, Axle, Rotation
Sports	Exercise, Athlete, Train, Broad-jump, Body, Basketball

It can be seen from table 1, the method of domain keywords can effectively express a particular category, and distinguish other categories. Noise influence could be effectively avoided by extracting the keywords which has strong instruction ability and good category distinguish ability for each category, the effect of text categorization improved.

To verify the effectiveness and superiority of the method, two experiments were carried out: 1. Compare our method with the method which domain keyword set respectively only contains domain high-frequency words and topic core words. 2. Compare our method with the traditional VSM and LDA model combined with the SVM algorithm.

1) The experiment 1

The VSM and LDA as text representation, combined with the SVM classification algorithm, as two groups of contrast experiment. Three groups of results are shown in table 2:

TABLE II. THE CLASSIFICATION PERFORMANCE COMPARISON OF THREE DIFFERENT CLASSIFICATION METHODS

Corpus	VSM+SVM		LDA+SVM		Our method	
	Macro_F1	Micro_F1	Macro_F1	Micro_F1	Macro_F1	Micro_F1
Fudan	0.68	0.695	0.772	0.79	0.821	0.836
Sogou	0.652	0.674	0.726	0.73	0.785	0.792
Weibo	0.588	0.604	0.676	0.698	0.718	0.726

The experimental results show that classification effect of the classification algorithm based on VSM is not ideal for short-text. Because the ability to describe information of short-text is weak, the accuracy of classification algorithm based on LDA also failed to achieve satisfactory results. Compared with the LDA model combining the SVM classification algorithm, classification algorithm proposed in this paper has been improved on Macro_F1 and Micro_F1, as table 3 shows:

TABLE III. THE IMPROVEMENT OF CLASSIFICATION PERFORMANCE OF OUR METHOD ON THE BASIS OF LDA

Corpus	Macro_F1			Micro_F1		
	Min	Max	Avg	Min	Max	Avg
Fudan	3.1%	5.9%	4.9%	3.8%	6.5%	4.6%
Sogou	4.4%	6.8%	5.9%	4.9%	8.1%	6.2%
Weibo	2.5%	5.4%	4.2%	1.5%	4.3%	2.8%

2) The experiment 2

Compare our method and the extension method respectively using high-frequency words and topic core words. The experimental results are shown in table 4 and figure 3-4:

TABLE IV. THE COMPARISON OF CLASSIFICATION PERFORMANCE OF THREE DIFFERENT EXTENSION METHODS ON THREE CORPUS

Corpus	High-frequency Words		Topic Core Words		Our method	
	Macro_F1	Micro_F1	Macro_F1	Micro_F1	Macro_F1	Micro_F1
Fudan	0.802	0.806	0.786	0.788	0.821	0.836
Sogou	0.716	0.724	0.73	0.754	0.785	0.792
Weibo	0.708	0.711	0.701	0.714	0.718	0.726

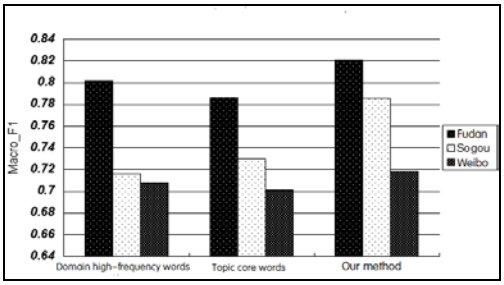

Fig. 3. The comparison of Macro_F1 of different extension methods on the corpora

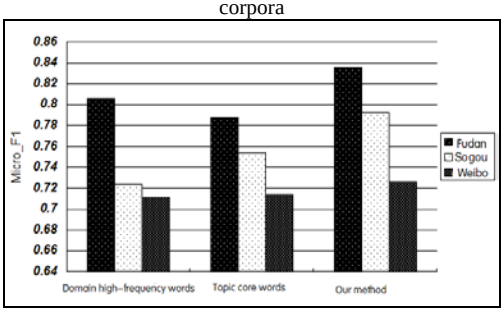

Fig. 4. The comparison of Macro_F1 of different extension methods on the corpora

The experimental results show that on the Fudan corpus, the classification performance of method based on the domain high-frequency words extension is better than that of the method based on extension of topic core words. On the Sogou Corpus, the classification performance of method based on the topic core word extension is better than that of domain high-frequency word extension method. On the Weibo Corpus, the classification performance of the two methods are quite. The method in this paper is superior to the method which domain keyword sets respectively only contains domain high-frequency words and topic core words.

Based on the above experiments, we can draw the conclusion that using the domain keyword sets to mine and supplement short-text implied information can effectively improve the effect of short-text classification. Classification performance of the method combined with the domain high-frequency words and topic core words is better than that of only using high-frequency words or topic core words of extension method.

V. CONCLUSIONS

This paper proposes a short-text classification algorithm based on domain keyword set and HowNet, the algorithm extract the training set keywords for each category by combining with different feature granularity. On fine-grained word, get training set high-frequency words by keyword frequency statistics; on coarse-grained latent topic, using LDA model to extract all kinds of topic core words. Combine domain high-frequency words and topic core words, delete the repeat keywords, then we get domain keyword set. Using the domain keyword set, combined with LDA, we can extend text feature of test text. Calculate text semantic similarity with the help of HowNet to, then forecast the category of the classification of text. The experimental results show that the method achieves good classification effect. But there are also some shortcomings, because the HowNet contains about 60000 words only, many keywords that not included in HowNet can't be calculated similarity, it will affect classification result. Therefore the next step of our work is to use HowNet combined with other algorithms (such as matching algorithms), improve the accuracy of short-text semantic similarity computation, and further improve the classification effect.

REFERENCES

[1] Zhao Hui, Liu Huailiang. Classification Algorithm of Chinese Short Texts Based on Wikipedia[J]. Library and Information Service, 2013,57(11):120-124.

[2] Zhang Suzhi, Liu Jingjiao. A Short Text KNN Classification Algorithm Based on Semantic[J]. Journal of Zhengzhou University of Light Industry(Natural Science),2012,27(6):1-4.

[3] Ning Yahui, Fan Xinghua, Wu Yu. Short Text Classification Based on Domain Word Ontology[J]. Computer Science, 2009,36(3):142-145.

[4] Zhan Yan, Chen Hao. Short Text Classification Based on Theme Ontology Features Extended[J]. Journal of Hebei University (Natural Science Edition), 2014,34(3):307-311.

[5] Hu Yongjun, Jiang Jiaxin, Chang Huiyou. A New Method of Keywords Extension for Chinese Short-Text Classification[J]. New Technology of Library and Information Service,2013(6):42-48.

[6] Sriram B, Fuhry D, Demir E, et al. Short Text Classification in Twitter to Improve Information Filtering[C]. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010: 841-842.

[7] Li Xiangdong, Cao Huan, Ding Cong, Huang Li. Short-text Classification Based on HowNet and Domain Keyword Set Extension[J]. New Technology of Library and Information Service, 2015,02:31-38.

[8] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research,2003,3: 993-1022.

[9] Si Xiance. Content-based Recommendation and Analysis of Social Tags[D]. Beijing: Tsinghua University,2010.

[10] Wu Jian, Wu Zhaohui, Li Ying, et al. Web Service Discovery Based on Ontology and Similarity of Words[J]. Chinese Journal of Computers,2005,28（4）: 595-602.

[11] Li Shengqi, Tian Qiaoyan, Tang Cheng. Disambiguating Method for Computing Relevancy Based on HowNet Semantic Knowledge[J].Journal of the China Society for Scientific and Technical Information,2009,28（5）: 706-711.

[12] Sun Jianwang, Lv Xueqiang, Zhang Leihan. Short Text Classification Based on Semantics and Maximum Matching Degree[J]. Computer Engineering and Design, 2013,34（10）: 3613-3618.

[13] Zhou Yun, Zhu Dingju, Bo Jia'ning. Sentence Similarity Calculation Based on Hownet[J]. Bulletin of Advanced Technology Research, 2010,4(8):32-37.

[14] Fudan University Chinese corpus [DB/OL]. [2014-06-20].http://www.datatang.com/data/43318.

[15] Sogou: Sogou Classification Corpus[DB/OL].[2014-06-20] . http://www.Sogou.com/labs/dl/c.html.

[16] Sina Corpus [DB/OL].[2014-06-20]. http://www.nlpir.org/?action-viewnews-itemid-231.

[17] Feng Guohe. Review of Performance Evaluation of Text Classification[J]. Journal of Intelligence, 2011,30（8）: 66-70.