Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Short text classification based on strong feature thesaurus*

Bing-kun WANG[†1,2], Yong-feng HUANG[1,2], Wan-xia YANG[1,2], Xing LI[1,2]

([1]*Information Cognitive and Intelligent System Research Institute, Department of Electronic and Engineering, Tsinghua University, Beijing 100084, China*)

([2]*Information Technology National Laboratory, Tsinghua University, Beijing 100084, China*)

[†]E-mail: Wangbingkun77@yahoo.com.cn; wbk10@mails.tsinghua.edu.cn

Received Dec. 19, 2011; Revision accepted June 25, 2012; Crosschecked Aug. 3, 2012

Abstract: Data sparseness, the evident characteristic of short text, has always been regarded as the main cause of the low accuracy in the classification of short texts using statistical methods. Intensive research has been conducted in this area during the past decade. However, most researchers failed to notice that ignoring the semantic importance of certain feature terms might also contribute to low classification accuracy. In this paper we present a new method to tackle the problem by building a strong feature thesaurus (SFT) based on latent Dirichlet allocation (LDA) and information gain (IG) models. By giving larger weights to feature terms in SFT, the classification accuracy can be improved. Specifically, our method appeared to be more effective with more detailed classification. Experiments in two short text datasets demonstrate that our approach achieved improvement compared with the state-of-the-art methods including support vector machine (SVM) and Naïve Bayes Multinomial.

Key words: Short text, Classification, Data sparseness, Semantic, Strong feature thesaurus (SFT), Latent Dirichlet allocation (LDA)

doi:10.1631/jzus.C1100373 **Document code:** A **CLC number:** TP391.4

1 Introduction

With the rapid development of Web 2.0, online publishing in diverse forms such as Twitter, bulletin board system (BBS), social networking services (SNS) and instant communication technology represented by Microsoft service network (MSN), have become the mainstream of information exchange. The online messages classified as short text share some common characteristics, such as short message length and intense user participation. Short texts cover topics of various kinds and are of increasing information importance. It provides a new information resource for the government to monitor the sentimental tendency of the general public. Business firms can track consumer preferences by analyzing these messages. The classification of short texts is the basis of hidden information extraction; thus, it has a wide range of applications including public opinion analysis, topic tracking, and consumer preferences indication.

Short length and poor informative content lead to weak linkage to certain topics. What is more, due to the nature of diversification of language, the same topic can be expressed in totally different ways, reducing the possibility of a feature term's appearing in several different short texts. As a result, short text classification based on feature term co-occurrence often failed to increase its accuracy due to data sparseness.

Intensive research has been conducted to tackle the data sparseness problem and improve the accuracy of short text classification. Sahami and Heilman (2006) tried to solve the problem by introducing a Web-based kernel method. Metzler *et al*. (2007) proposed a solution using similarity measurement. Based on implicit topics, Phan *et al*. (2008) proposed a general framework for short text classification. Other effective approaches include a three-level

* Project (No. 20111081023) supported by the Tsinghua University Initiative Scientific Research Program, China

semantic extension model based on Wikipedia and WordNet (Hu *et al.*, 2009). However, these studies focused mainly on solving the data sparseness problems in short text, without considering the effect of feature terms in determining the category.

To further improve the accuracy of short text classification, we must consider the semantic importance of certain feature terms. Inspired by the 'structure+average' method in probabilistic graphical models (Koller and Friedman, 2009), we propose a new method for short text classification based on both statistical and semantic consideration. First, we establish a strong feature thesaurus (SFT) by mining a large-scale external data collection based on the latent Dirichlet allocation (LDA) and information gain (IG) models. Then, we put larger weights on feature terms in SFT. Experimental results show that our method improves the classification accuracy.

2 Related work

At present there are mainly three types of method in short text classification. The first is calculating the similarity of the feature terms using search engines, and then achieving the short text classification based on similarity. The second is using world knowledge (such as Wikipedia or WordNet) to expand feature terms and map feature terms to the concept, thereby reducing negative impact of data sparseness and improving classification performance in short text. The third is mining the implicit topic of large-scale data collection and short text based on the topic model, and then improving the co-occurrence probability of the feature terms by mapping feature terms to topics.

Some researchers (Sahami and Heilman, 2006; Bollegala *et al.*, 2007; Yih and Meek, 2007) calculated the semantic similarity of feature terms in short text using the results returned by a search engine. Feature terms that have similar meaning but different expressions are linked according to semantic similarity. These methods partly solve the data sparseness problem in short text. For example, without taking semantic similarity into account, the computer will consider 'artillery' and 'tank' as unrelated terms. But in fact, 'artillery' and 'tank' have strong similarity. The semantic similarity has been calculated with snippets returned by a search engine using 'artillery' and 'tank' as the query, and then applied to short text classification. Thus, the accuracy of short text classification is improved to some extent. The disadvantage of this method is repeatedly querying search engines, which is very time-consuming and not good at real-time tasks.

Gabrilovich and Markovitch (2005) demonstrated that world knowledge has a certain effect on eliminating semantic gaps. Gabrilovich and Markovitch (2006) used external knowledge corpus (Wikipedia and Open Directory Project) to map feature terms using the ontology concept, and then expanded text feature representation by the ontology concept. Li *et al.* (2006) calculated the semantic similarity of short text by using WordNet to mine semantic relationships between feature terms. Hu *et al.* (2009) proposed a three-layer structure model based on Wikipedia and WordNet, making full use of extension of internal semantic and external concept to improve the accuracy of clustering short texts. The methods improve the co-occurrence frequency of feature terms by exploiting the semantic of feature terms and mining short text semantic similarity. These methods partly solve the data sparseness problem, but they bring further improvement of data dimensionality and increase complexity in data processing.

Phan *et al.* (2008) mined the implicit topic on a large-scale data collection based on the LDA model, and created a general framework for short text classification using an implication topic. The key idea is building a classifier by a combination of the implicit topic obtained by large-scale data collection and training datasets of short text. Quan *et al.* (2010) measured similarity between different feature terms in short texts with different topics based on the LDA model, and experience on yahoo Q & A datasets showed that the classification accuracy had increased to some extent. Essentially, these methods enhance the accuracy of short text classification by increasing the co-occurrence probability of feature terms in short text.

Similarly, our method uses the LDA model to mine the implicit topic on large-scale datasets. Phan *et al.* (2008) and Quan *et al.* (2010) solved the data sparseness problem by merging feature terms belonging to the same topic. In contrast, we use the LDA model to select a portion of the feature terms that have

the highest probabilities in each hidden topic, and then delete the feature terms that have the highest probabilities in more than one category. Finally, we obtain a strong feature thesaurus (SFT) made up of strong feature terms.

Bollegala *et al.* (2011) proposed a method of constructing a sentiment similarity thesaurus to find the association between words that express similar sentiments in different domains, then using the thesaurus to expand feature vectors during classification. Different from their sentiment similarity thesaurus, our SFT is made up of feature terms that have high category discrimination. The strong feature terms are fetched by applying the LDA model to external large-scale datasets downloaded from the Internet. Also, different from expanding feature vectors using sentiment sensitive thesaurus, we weight feature terms in SFT, which does not increase the length of the feature vector. What is more, the core of Bollegala *et al.* (2011)'s constructing a sentiment sensitive thesaurus is grouping different words that express the same sentiment by finding the relationship between different feature terms based on the co-occurrence of feature terms and sentiment labels of the document. Our method is also constructing an SFT, but the core is excavating the semantic difference of feature terms and comparing the semantic difference of feature terms with statistical methods by weighting feature terms in SFT.

3 General framework for short text classification

There are basically two different approaches to text classification. One is the application of expertise in a certain domain; the other is using statistical methods. Currently, statistical methods are the mainstream, barely requiring human involvement. However, entirely relying on statistical techniques makes it difficult to further improve the classification accuracy beyond a certain level. Particularly, the data sparseness of short texts naturally limits the information obtained via statistical methods. As a result, the classification accuracy of short texts is significantly lower than that of normal texts. To solve the problem, we combine domain knowledge with statistical methods. The general framework of our method is depicted in Fig. 1.

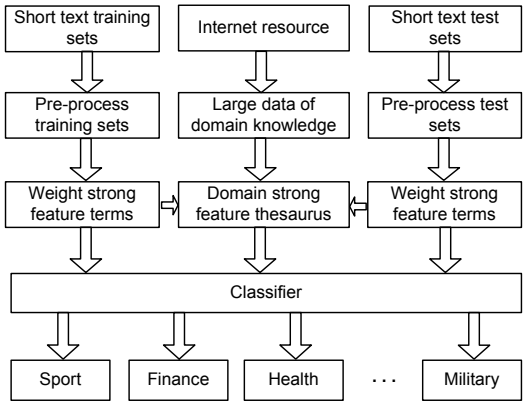

Fig. 1 General framework for short text classification

Basically, our new method is composed of four steps. First, we obtain the datasets of domain knowledge and construct the SFT. The datasets of domain knowledge must be large and sufficiently rich in information to replace the domain character. To achieve balanced domain datasets, each crawling transaction in each catalogue is limited by the maximum depth of hyperlinks (for example, depth=4). To find strong feature terms, we employ LDA to the domain datasets. According to the probability of feature terms under each topic, we choose the feature terms with high probability, and delete some feature terms that have high probability of occurring in more than one category. For example, when establishing SFT we delete 'conditions', which is a high probability feature term belonging to topic 3 and topic 15. We establish a thesaurus made up of all strong feature terms by screening the data of domain knowledge on a large scale. The process is conducted based on the LDA model and IG model.

The second step is the pre-processing of 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, to emphasize the importance of strong feature terms, we typically give heavy weights to those feature terms of the short text that are included in the SFT. We calculate the weights of feature terms with Eq. (5).

Finally, we obtain the classifiers using machine learning technology with the training sets (Witten *et al.*, 2011). Then we safely classify the test sets using the classifiers.

In our method there are two key points: first to construct the SFT, and then to weight the feature vectors by assigning higher weights to those features in SFT. There are two motivations for using our approach instead of the natural one-step approach.

First, we establish a larger and broader SFT by mining a large-scale external data collection based on LDA than by mining only labeled training data. The SFT includes some feature terms that we cannot identify using only labeled training data. We can decrease the adverse impact of shortage of labeled training data based on the SFT. For example, when the words 'destroyers' and 'missiles' do not appear or appear only once in labeled training data, whereas they appear many times in large-scale external data collection, we can put 'destroyers' and 'missiles' into SFT based on LDA, but we do not identify the two feature terms if we use only labeled training data. During the classification, if the words 'destroyers' and 'missiles' appearing in a text belong to test datasets, we can safely put this text to the military category based on SFT, but it is difficult to accurately classify a short text including 'destroyers' and 'missiles' using only traditional methods based on labeled training data.

Second, during the classification, through weighting feature terms in SFT, we improve the importance of strong feature terms in the feature vector and incorporate semantic information of strong feature terms into the vector space model (VSM), and then improve the precision of short text classification. For example, in posts of BBS relating to missile destroyers, the words 'destroyers' and 'missiles' in SFT appear only once, but the words 'computers', 'learning', 'training', 'technology', 'economic', and 'money' appear many times. We can then safely put this text to the military category through weighting 'destroyers' and 'missiles' in SFT, but we indeed classify the text into the economic category using traditional methods. All in all, through our method, we can decrease the adverse impact of shortage of labeled training data and improve the precision of short text classification.

To better describe our method, we define some terminology used in this article:

Definition 1 (Strong feature terms) Strong feature terms are highly semantic-orientated and are vital in determining the category of the specific text, such as 'bank', 'gold', and 'credit' for the finance category.

Definition 2 (Strong feature thesaurus) Strong feature thesaurus is the collection of strong feature terms. The strong feature terms are obtained by mining large-scale data collections belonging to a special domain. It is represented as $T=\{t_i|i\in(1, V)\}$, where V is the number of strong feature terms t_i in strong feature thesaurus T.

Definition 3 (Contribution of category, COC) COC is dividing the IG of a specific strong feature term t_i by the average IG of all strong feature terms. It is represented as

$$\mathrm{COC}(t_i)=\mathrm{IG}(t_i)V\Big/\sum_{i=1}^{V}\mathrm{IG}(t_i). \tag{1}$$

4 Constructing strong feature thesaurus and weighting strong feature terms

4.1 LDA model

LDA is a generative topic model, using three Bayesian probability graphs to achieve topic modeling of the text and information extraction. We can extract a limited number of topics to represent text by the LDA model. This method reduces the space dimension of the text expression, while retaining the essential statistics information of the text. The LDA model includes the text sets layer, the text layer, and the terms layer from the outside. The strength of the relationship between implicit topics in text sets is determined by α, and the probability distribution of the implied topic is determined by β. Parameters (α, β) fully reflect the topic features of text sets. Random vector $\boldsymbol{\theta}$ indicates the probability distribution of the hidden topic in the text layer; z denotes the implicit topic share of which the text distributes to every term; w is the representation of the terms vector in the target document. The generative model is shown in Fig. 2. For a detailed description of the LDA model, refer to Blei *et al*. (2003).

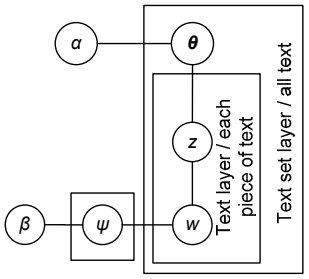

Fig. 2 Latent Dirichlet allocation (LDA), a generative topic model

In this work, we use the LDA model to extract high probability words in each topic as strong feature terms. As an effective method, Gibbs sampling is used to extract hidden topics and feature terms under each topic in LDA models (Griffiths and Steyvers, 2004). The same method is adopted in this work.

Given a set composed of N pieces of text $D=\{d_n|n=1, 2, \ldots, N)\}$ with K unique feature terms $W=\{w_k|k=1, 2, \ldots, K\}$, text set D contains M hidden topics $Z=\{z_m|m=1, 2, \ldots, M)\}$. The probability of topic z_m in Gibbs sampling can be calculated by

$$P(z_m \mid z_{-m}, W) \propto \frac{N_{z_m}^{w_k}+\beta}{\sum_{k=1}^{K} N_{z_m}^{w_k}+K\beta} \cdot \frac{N_{d_n}^{z_m}+\alpha}{\sum_{m=1}^{M} N_{d_n}^{z_m}+M\alpha}, \quad (2)$$

where $N_{z_m}^{w_k}$ represents the number of times by which the word w_k is assigned to topic z_m except the current assignment, $\sum_{k=1}^{K} N_{z_m}^{w_k}$ represents the total number of words assigned to topic z_m except the current assignment, $N_{d_n}^{z_m}$ represents the number of words in document d_n assigned to topic z_m except the current assignment, and $\sum_{m=1}^{M} N_{d_n}^{z_m}$ represents the total number of words in document d_n except the current word w_k. After enough iterations, the probability of word w_k under topic z_m can be calculated using Eq. (3):

$$P(w_k \mid z_m) \propto \frac{N_{z_m}^{w_k}+\beta}{\sum_{k=1}^{K} N_{z_m}^{w_k}+K\beta}. \quad (3)$$

For more details about Gibbs sampling in LDA, readers are referred to Griffiths and Steyvers (2004) and Heinrich (2005).

4.2 Information gain

Given a category set $C=\{C_j|j=1, 2, \ldots, J)\}$, feature term set $W=\{w_k|k=1, 2, \ldots, K)\}$, and text set $D=\{d_n|n=1, 2, \ldots, N)\}$, IG for feature term w_k is defined as follows (Zong, 2008):

$$\begin{aligned}\mathrm{IG}(w_k) &= \mathrm{Entropy}(D) - \mathrm{ExpectedEntropy}(D_{w_k}) \\ &= -\sum_{j=1}^{J} P(C_j)\cdot \log_2 P(C_j) \\ &\quad - P(w_k)\left[-\sum_{j=1}^{J} P(C_j \mid w_k)\cdot \log_2 P(C_j \mid w_k)\right] \\ &\quad - P(\overline{w_k})\left[-\sum_{j=1}^{J} P(C_j \mid \overline{w_k})\cdot \log_2 P(C_j \mid \overline{w_k})\right].\end{aligned} \quad (4)$$

In theory, IG is the best feature selection method. In practice, however, many feature terms have high IG but relatively low frequency. When the feature terms are selected with IG, the data sparseness problem often occurs (Zong, 2008). For this reason, feature selection by only IG cannot guarantee the best choice. Under normal circumstances, we select the feature by combining IG and other feature selection methods to improve the effect of selecting feature terms (Manning *et al.*, 2008).

4.3 Constructing strong feature thesaurus

We extract different topics and corresponding feature terms from data of domain knowledge using the LDA model, and then construct the strong feature thesaurus. The process can be divided into four steps (Fig. 3).

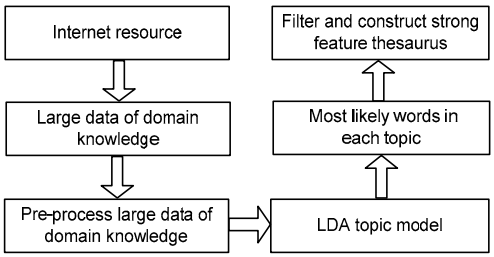

Fig. 3 Process of constructing strong feature thesaurus

The first step is to obtain the datasets of domain knowledge. We download many kinds of Web pages from different websites (for example, sohu.com) using a Web crawler. After filtering out repeated pages and other noise, the Web documents are resolved into pure text (such as the Sogou Category Corpus).

In the second step, the domain datasets are pre-processed. We use the ICTCLAS system developed by Chinese Academy of Sciences (CAS, 2010) to conduct Chinese word segmentation. Only nouns, verbs, and adjectives are left in the text since they convey semantic information. Then we count the frequency of the remaining words. Any word with a frequency less than a threshold would be eliminated from the text.

In the third step, the LDA model is used to extract topics from the text. We use the Gibbs sampling

algorithm to extract diverse topics in processed domain datasets. When extracting topics, we choose parameters according to Griffiths and Steyvers (2004), and then calculate the probabilities of topics and feature terms using Eqs. (2) and (3), respectively.

Finally, given a category set $C=\{C_j|j=1, 2, \ldots, J)\}$ and feature term set $W=\{w_k|k=1, 2, \ldots, K)\}$, we use Eq. (4) to calculate the IG of each feature term w_k, and then filter the same feature terms in different topics. After filtering out some feature terms that are not semantically significant, we obtain a strong feature thesaurus including only strong feature terms.

4.4 Weighting strong feature terms

When classifying short text, the contribution of each feature term in determining categories is different. However, common statistical classification methods do not take this into account. Particularly, our method gives greater weights to feature terms with a larger COC, and thus the accuracy of text classification is improved.

After the pre-processing of the training data and test data, we obtain the VSM expression of short texts. In the vector obtained using VSM, we put heavier weights to those feature terms included in the strong feature thesaurus.

Given a feature term set $W=\{w_k|k=1, 2, \ldots, K)\}$ and text set $D=\{d_n|n=1, 2, \ldots, N)\}$, the weight w_k in d_n is denoted by $\mathrm{tf}(w_k)$. Then VSM of any text d_n can be represented by $\mathrm{VSM}(d_n)=\{\mathrm{tf}(w_k)|k=1, 2, \ldots, K)\}$. Given a strong feature term set $T=\{t_i|i=1, 2, \ldots, V)\}$, if $w_k \in T$ in d_n, then w_k is weighted by

$$\mathrm{NTF}(w)=\mathrm{tf}(w_k)+\mathrm{COC}(w_k). \tag{5}$$

Thus, each piece of short text is made up of two kinds of feature term, common and strong. The VSM of each piece of short text is represented as

$$\mathrm{VSM}(d_n)=\{\mathrm{tf}(w_1),\ldots,\mathrm{tf}(w_L),\mathrm{NTF}(w_{L+1}),\ldots,\mathrm{NTF}(w_k)\}. \tag{6}$$

5 Evaluation

5.1 Data sets

Two datasets were used to evaluate the performance of our proposed method. Due to the absence of standard Chinese short text datasets, we obtained data in two different ways. The first one involves the application of a Web spider. We downloaded comments from BBS and some commercial sites. Second, we extracted short text data from the Sogou Corpus, a well-known Chinese corpus (Sohu Research & Development Center, 2008). In general, it is more difficult to classify the truncated texts, because they are more semantically incomplete. The categories and distribution of the two datasets are given in Fig. 4.

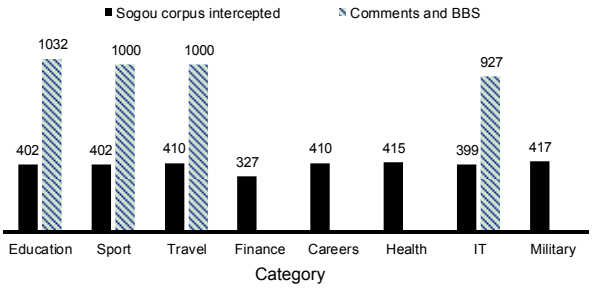

Fig. 4 Categories and distribution of the two datasets

5.2 Evaluation design

There are mainly two factors influencing short text classification, the classification algorithm and the length of short text. The length of text is a very important factor when classifying short texts. Different lengths of short text may lead to different classification accuracies. Our proposed method has two key steps: identifying strong feature terms based on the LDA model, and assigning higher weights to those strong feature terms using the new term frequency (NTF) method (Eq. (5)) when weighting the feature vectors.

We conducted three experiments to study the effect of the length of short text, strong feature selection, and strong feature weighting methods on the performance of our proposed method. First, we compared feature selection methods based on LDA against methods based on *k*-means clustering. Second, we compared NTF against the TFIDF weighting method. Third, we investigated the relationship between classification accuracy and the length of short text.

To evaluate the overall performance of our methods, we compared our methods with some previously proposed text classification methods such as Naïve Bayes, SVM, and Naïve Bayes Multinomial in two different datasets. We used three absolute

indicators, i.e., accuracy, recall rate, and F1 (Zhang *et al*., 2005; Peng *et al*., 2008), and a relative indicator, i.e., rate of error reduction (ROER) (Phan *et al*., 2008), to assess the result of classification. The four text representation models and ROER are defined as follows:

BOW: the traditional 'bag of words' model with the term frequency (TF) weighting method.

BOW+WSF(*k*-means+NTF): an expansion of BOW, including the weighting of strong feature terms. Here, we identify strong feature terms by selecting the most frequent words in each cluster which is obtained by the *k*-means clustering method, and weight strong features in NTF.

BOW+WSF(LDA+TFIDF): an expansion of BOW, including the weighting of strong feature terms. Here, we identify strong features by LDA and weight strong features in term frequency and inverse document frequency (TFIDF).

BOW+WSF(LDA+NTF): our proposed model, an expansion of the BOW model, including the weighting of strong feature terms. Here, we identify strong features by LDA and weight strong features in NTF.

ROER: the difference between the classification error of BOW and the error of BOW+WSF(LDA+NTF) divided by the classification error of BOW.

All texts in the datasets were processed by Python, and then classified in the freely available machine learning software Weka3.6 (Witten *et al*., 2011). Meanwhile, five-fold cross validation was used to verify the performance of short text classification methods.

5.3 Constructing strong feature thesaurus

The kernel of our proposed method is constructing SFT. We downloaded different catalogues of Web pages from yahoo.com.cn and sohu.com using a Web crawler, and the crawling transaction in each catalogue was limited by 10 000 Web documents. After filtering out the repeated pages and other noise, the Web documents were resolved into pure text as domain knowledge datasets. The datasets of domain knowledge were pre-processed by deleting punctuation and words in the stop word list. The ICTCLAS system (http://ictclas.org/) was used to conduct Chinese word segmentation and POS tagging. Then we counted the frequency of those words.

Here, we used GibbsLDA++ (Phan *et al.*, 2011), an open source tool. We used GibbsLDA++ to extract strong feature terms. According to experience (Quan *et al.*, 2010), we set α=50/Z and β=0.1, and chose 10 topics in each category and obtained 80 topics in total. After obtaining the topic-probability distribution of the feature terms, according to the probability distribution of feature terms, we chose 20 feature terms with the maximum probabilities in each topic. Overall, we obtained 1600 feature terms. Some experimental results are shown in Table 1.

As we can see from Table 1, the 20 feature terms with maximum probabilities in each topic have powerful focusing ability. For example, from the 20 feature terms with maximum probabilities in topic 3, which belongs to the military category, we can basically determine that the text containing the feature terms has a larger probability of belonging to the military category than to the other categories. The 20 feature terms with maximum probabilities in other topics also show a strong ability to distinguish categories.

After obtaining 1600 feature terms, we used IG to filter out some feature terms of each topic whose IG is less than the average IG. Then, we deleted the same strong feature terms in different topics such as 'conditions' in Table 1. Finally, we obtained an SFT composed of 1086 strong feature terms.

To evaluate the feature selection method based on LDA, we did a simple *k*-means clustering and then selected the highest frequency words in each cluster as strong feature terms. Here, we used a *k*-means clustering package in weka3.6 and set *k*=55. We selected 20 feature terms with the maximum frequency in each cluster. Totally, we obtained 1100 feature terms. Table 2 shows some experimental results.

From Table 2, the 20 feature terms with maximum frequencies in each cluster included some terms that are inappropriate to be taken as strong feature terms, even though the feature terms had higher frequencies in each cluster. There are two reasons. First, some high frequency terms might be popular in more than one cluster. For example, 'relationship', 'technology', and 'situation' all appeared in three different clusters. Second, some high frequency terms in a cluster have lower IG than average, for example, 'news', 'media', 'suddenly', 'relevant', 'youth' in cluster 5 and 'part', 'field', 'products' in cluster 43.

Table 1 The top 20 words of some topics by latent Dirichlet allocation (LDA) in knowledge sets

Topic 3: Military	Topic 8: Finance	Topic 15: Health
Train 0.04790142401739146	Bank 0.07854848977859517	Treat 0.03766677914474104
Army 0.03882205719818412	Gold 0.030843884841196857	Patient 0.029185324347247174
Equip 0.01911051434521689	Credit 0.02976869208460069	Invalid 0.025537386799937983
Battle 0.01662601960193279	Chinese yuan 0.02801443021857537	Medicine 0.02384109584043921
Military 0.013977109765343125	Insurance 0.025496215604442245	Tumour 0.019189975467619993
Maneuver 0.013410791110623956	Interest rate 0.018224517224305015	Disease 0.019135256404410356
War 0.01200412864567634	Business 0.017007851736577775	Surgery 0.017876717950588688
Battlefield 0.011803176864969537	Market 0.012509018886609607	Mediciner 0.01497660760047788
Officers and soldiers 0.0116387617716	Personal 0.012424135247930962	Symptom 0.014958367912741334
PLA 0.011474346678358407	Risk 0.011914833415859093	Hospital 0.014137581964596767
Soldiers 0.011236858210250368	Foreign fund 0.010726462474358068	Function 0.012149456001313258
Mission 0.009976342494907699	Money 0.010302044280964845	Clinical 0.011036835049383955
Command 0.00941002384018853	Foreign exchange 0.0094815024404	Hypertension 0.010617322231443398
Land force 0.009135998684679254	Operation 0.009340029709273539	Censor 0.008519758141740615
Conditions 0.008313923218151427	Exchange rate 0.00902878970078517	Hypophysis 0.008337361264375155
Tank 0.007966824687839677	Charge 0.008576076961165736	**Conditions 0.008118485011536603**
Military region 0.007784141250833493	Deposit 0.008519487868713305	Prodessor 0.007844889695488414
Fighting capacity 0.0075283844390248	Organ 0.008179953313998727	Transplant 0.007024103747343846
Alliance 0.007510116095324218	Manage money 0.007953596944189	Breast cancer 0.00691466562092457
Tactic 0.007217822596114323	Client 0.007698946028153074	Chinese medicine 0.00678698780676

Boldface represents the term appearing in different topics

Table 2 The top 20 words of some clusters by *k*-means in knowledge sets

Cluster	Top 20 words
Cluster 5 (Health)	**Composition**, pharmaceutical, **relationship**, Ministry of Health, **situation**, **news**, the human body, **the people**, the head and neck, cancer, vitamin, skin, the authentication, **media**, **suddenly**, **relevant**, **youth**, enterprise, **technology**, body
Cluster 21 (Finance)	**Relationship**, **technology**, **situation**, income, the CSRC, product, formal, joint, **Hong Kong**, **relevant**, securities, bull market, expected, the stock market, the customer, view, the fund, **national**, stock, reform
Cluster 43 (Military)	**Relationship**, officials, **technology**, **situation**, **products**, Department of Defense, official, Taiwan, **opinions**, tanks, base, media, scale, conference, minister, **Hong Kong**, relevant, reconnaissance, **part**, **field**

Boldface represents the term appearing in different topics

5.4 Results and analysis

5.4.1 Effect of feature selection and feature weighting methods

To examine the effect of different feature selection and weighting methods, we extracted 56 characters from each text in the Sogou Corpus and constructed a short text dataset. We identified strong feature terms based on LDA and *k*-means clustering methods, weighted strong feature terms in NTF and TFIDF methods, and then classified short text datasets by Naïve Bayesian Multinomial in Weka3.6. The experimental results are shown in Table 3.

In Table 3, the descending order of four text presentence models according to classification performance is BOW+WSF(LDA+NTF), BOW+WSF(LDA+TFIDF), BOW+WSF(*k*-means+NTF), and BOW. We can draw some conclusions from the sequence. First, the method of identifying and weighting strong feature terms is effective on short text classification. Second, LDA is more effective than *k*-means in identifying strong feature terms. Third, weighting strong feature terms with NTF outperforms TFIDF. The classification performance was improved by 0.5% (from 86.5% to 87%) using NTF instead of TFIDF, and by 1.1% (from 85.9% to 87%) using LDA instead of *k*-means. This means the method of identifying strong feature terms is more important than weighting methods in short text classification, which is in accordance with our intuition.

Table 3 Classification performances of four different text presentation models

Category	Precision				Recall				F1			
	BOW	BOW1	BOW2	BOW3	BOW	BOW1	BOW2	BOW3	BOW	BOW1	BOW2	BOW3
Finance	0.890	0.874	0.886	0.875	0.789	0.786	0.804	0.810	0.836	0.828	0.843	0.841
IT	0.805	0.819	0.835	0.846	0.820	0.830	0.835	0.837	0.812	0.824	0.835	0.841
Health	0.811	0.802	0.807	0.831	0.764	0.793	0.798	0.807	0.787	0.798	0.802	0.819
Sport	0.970	0.963	0.966	0.963	0.968	0.970	0.978	0.980	0.969	0.967	0.972	0.972
Travel	0.849	0.857	0.872	0.883	0.866	0.878	0.883	0.880	0.857	0.867	0.878	0.882
Education	0.856	0.880	0.879	0.890	0.799	0.821	0.833	0.826	0.826	0.849	0.856	0.857
Careers	0.729	0.764	0.766	0.753	0.834	0.829	0.829	0.854	0.778	0.795	0.796	0.800
Military	0.912	0.921	0.923	0.932	0.945	0.947	0.950	0.950	0.928	0.934	0.936	0.941
Weighted average	0.852	0.860	0.866	0.871	0.849	0.859	0.865	0.870	0.850	0.859	0.865	0.870

BOW1: BOW+WSF(*k*-means+NTF); BOW2: BOW+WSF(LDA+TFIDF); BOW3: BOW+WSF(LDA+NTF)

On an intuitive level, if we cannot identify strong feature terms which have high discrimination in classification, no method of weighting strong feature terms could achieve better results.

5.4.2 Evaluation of relationship between classification performance and the length of texts

We extracted 56 and 112 terms from each text in the Sogou Corpus respectively and constructed two short text datasets with different lengths. We used BOW and BOW+WSF(LDA+NTF) in the two short text datasets and classified the text by Naïve Bayesian Multinomial to estimate the relationship between classification performance and the length of short text. The experimental results are shown in Fig. 5.

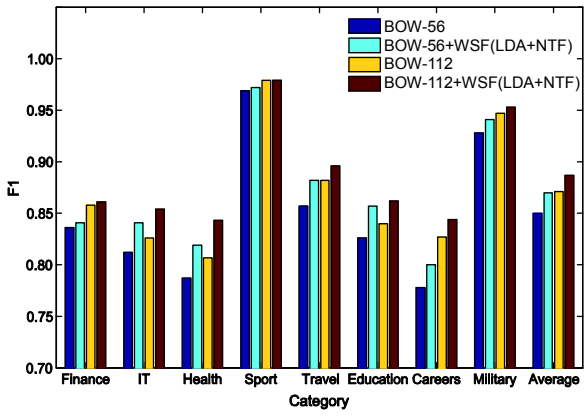

Fig. 5 Results of classification performance vs. the length of text

BOW-56 and BOW-112 mean each piece of text of the dataset has 56 and 112 word terms, respectively

As shown in Fig. 5, BOW-112+WSF(LDA+NTF) and BOW-56+WSF(LDA+NTF) were better than BOW-112 and BOW-56 respectively, which implies that our proposed method can be used for different short text datasets. The performances of BOW-112+WSF(LDA+NTF) and BOW-112 were better than those of BOW-56+WSF(LDA+NTF) and BOW-56 respectively, which means accuracy has positive correlation with the length of short text in classification. Theoretically, word co-occurrences are the basis of the statistical classification method. In general, the longer the text, the higher the probability of the word co-occurrence, and the better the classification performance.

5.4.3 Overall evaluation of classification methods

To evaluate the overall performance of our method, we compared our method with some previously proposed text classification methods such as Naïve Bayes, SVM, and Naïve Bayes Multinomial in two different datasets.

The texts of the Sogou Corpus can be divided into eight categories, and 56 terms were extracted from each piece of text. The experimental results are shown in Fig. 6. The datasets collected from BBS and commercial sites can be divided into four categories. The experimental results are shown in Fig. 7.

Figs. 6 and 7 show that our method outperformed Naïve Bayes, SVM, and Naïve Bayes Multinomial. There are basically two reasons. First, we established a larger and broader strong feature thesaurus by mining a large-scale external data collection based on LDA rather than mining only labeled training data, so feature terms in SFT are more effective and representative. Second, by putting heavier weights on those strong feature terms according to

their semantic importance, the classification accuracy is improved.

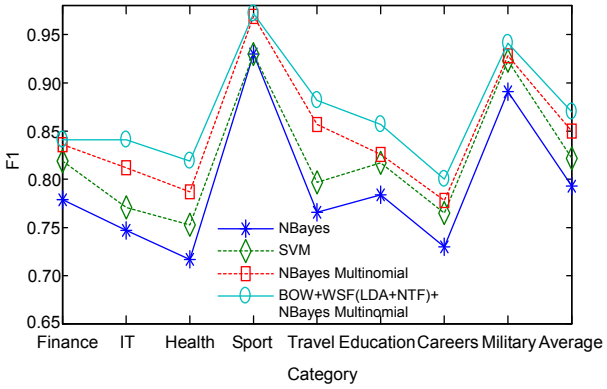

Fig. 6 Classification results for the Sogou Corpus intercepted

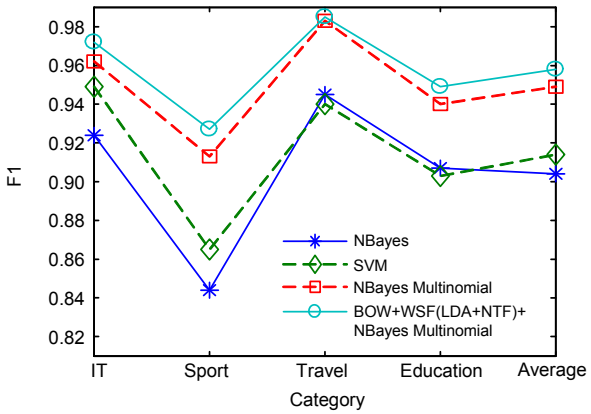

Fig. 7 Classification results for comments and BBS

To evaluate the effectiveness of our method when classification accuracy is different, we compared ROER in two datasets using BOW and BOW+WSF(LDA+NTF) models. The results are presented in Fig. 8.

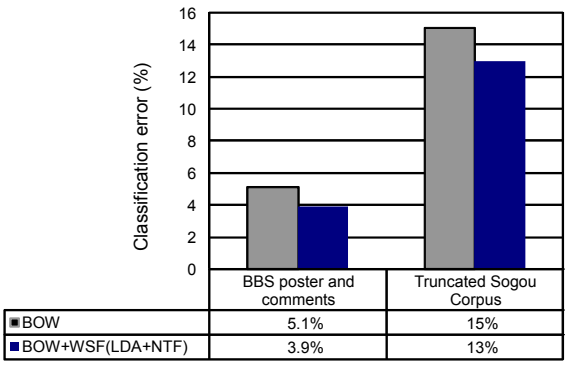

Fig. 8 Results of classification by two text representation models

Fig. 8 shows that, for BBS data, the error rate of the BOW model was 5.1%, while the error rate of BOW+WSF(NTF) was 3.9% in which 23.5% (i.e., (5.1−3.9)/5.1×100%) error was removed. For the data obtained from the Sogou Corpus, the error rate of the BOW model was 15% and the error rate of BOW+WSF(NTF) was 13%. ROER was 13.3% (i.e., (15−13)/15×100%).

Some interesting results can be found. With the improvement of accuracy, ROER increased from 13.3% to 23.5%. That is to say, the higher the classification accuracy achieved, the more the improvement produced by our method. This is due to the consideration of semantic importance by weighting strong feature terms obtained from mining a large-scale domain knowledge dataset. This is the main contribution of our method.

6 Conclusions

This paper deals with a new method to further improve the accuracy of short text classification combining statistical and semantic information. Our innovative investigation basically lies in three aspects. First, unlike previous work, we consider the difference of semantic influence of feature terms on the whole text. Second, we adopt the LDA model to obtain the domain knowledge and furthermore the SFT. Third, we propose a new weighting method for feature terms in SFT.

Theoretical analysis and experimental results show that our proposed method is more effective than previous ones. What is more, our method has better performance when the classification accuracy reaches a higher level. Due to the lack of short text datasets, however, our experimental data is incomplete, and thus there is room for improvement. We will further expand the scale of the experiment and apply the method to various kinds of data.

References

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**(1):993-1022.

Bollegala, D., Matsuo, Y., Ishizuka, M., 2007. Measuring Semantic Similarity Between Terms Using Web Search Engine. Proc. 16th Int. Conf. on World Wide Web, p.757-766. [doi:10.1145/1242572.1242675]

Bollegala, D., Weir, D., Carroll, J., 2011. Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for

Cross-Domain Sentiment Classification. Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, p.132-141.

CAS (Chinese Academy of Sciences), 2010. Chinese Lexical Analysis System of the CAS. Institute of Computing Technology, Chinese Academy of Sciences. Available from http://ictclas.org/ [Accessed on Sept. 20, 2011].

Gabrilovich, E., Markovitch, S., 2005. Feature Generation for Text Categorization Using World Knowledge. Proc. 19th Int. Joint Conf. on Artificial Intelligence, p.1048-1053.

Gabrilovich, E., Markovitch, S., 2006. Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. Proc. 21st National Conf. on Artificial Intelligence, p.1301-1306.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. *PNAS*, **101**(suppl_1):5228-5235. [doi:10.1073/pnas.0307752101]

Heinrich, G., 2005. Parameter Estimation for Text Analysis. Technical Report, University of Leipzig, Germany. Available from http://www.arbylon.net/publications/text-est.pdf

Hu, X., Sun, N., Zhang, C., Chua, T.S., 2009. Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge. Proc. 18th ACM Conf. on Information and Knowledge Management, p.919-928. [doi:10.1145/1645953.1646071]

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge, USA, p.3-6.

Li, Y.H., McLean, D., Bandar, Z.A., O′Shea, J.D., Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.*, **18**(8):1138-1150. [doi:10.1109/TKDE.2006.130]

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK, p.257. [doi:10.1017/CBO9780511809071]

Metzler, D., Dumais, S., Meek, C., 2007. Similarity Measures for Short Segments of Text. 29th European Conf. in Information Retrieval Research, p.16-27.

Peng, T., Zuo, W.L., He, F.L., 2008. SVM based adaptive learning method for text classification from positive and unlabeled documents. *Knowl. Inf. Syst.*, **16**(3):281-301. [doi:10.1007/s10115-007-0107-1]

Phan, X.H., Nguyen, L.M., Horiguchi, S., 2008. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections. Proc. 17th Int. Conf. on World Wide Web, p.91-100. [doi:10.1145/1367497.1367510]

Phan, X.H., Nguyen, C.T., Le, D.T., Nguyen, L.M., Horiguchi, S., Ha, Q.T., 2011. A hidden topic-based framework toward building applications with short Web documents. *IEEE Trans. Knowl. Data Eng.*, **23**(7):961-976. [doi:10.1109/TKDE.2010.27]

Quan, X.J., Liu, G., Lu, Z., Ni, X.L., Liu, W.Y., 2010. Short text similarity based on probabilistic topics. *Knowl. Inf. Syst.*, **25**(3):473-491. [doi:10.1007/s10115-009-0250-y]

Sahami, M., Heilman, T.D., 2006. A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. Proc. 15th Int. Conf. on World Wide Web, p.377-386. [doi:10.1145/1135777.1135834]

Sohu Research & Development Center, 2008. Text Classification Corpus of Sogou Labs. Available from http://www.sogou.com/labs/dl/c.html [Accessed on Sept. 20, 2011].

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Press, San Francisco, USA.

Yih, W.T., Meek, C., 2007. Improving Similarity Measures for Short Segments of Text. Proc. 22nd National Conf. on Artificial Intelligence, p.1489-1494.

Zhang, Y.T., Gong, L., Wang, Y.C., 2005. An improved TF-IDF approach for text classification. *J. Zhejiang Univ.-Sci.*, **6A**(1):49-55. [doi:10.1631/jzus.2005.A0049]

Zong, C.Q., 2008. Statistical Signal Processing. Tsinghua University Press, Beijing, China, p.344 (in Chinese).

Journal Citation Reports® Notices

. . .

2011 JCR Data Update

After JCR data are published in June 2012, this section will be updated weekly with additions or adjustments to the JCR data. All changes will be reflected in the JCR Web interface when the dataset is reloaded and closed in September 2012.

Full Title	JCR Abbreviation	Reason for Update	Total Cites	Journal Impact Factor	5-Year Impact Factor	Immediacy Index
. . .						
Journal of Zhejiang University-SCIENCE C-Computers & Electronics	J ZHEJIANG U-SCI C	Did not appear	37	0.308	0.308	0.048