Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 22 (2013) 78 – 86

Procedia Computer Science

17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013

Combining Lexical and Semantic Features for Short Text Classification

Lili Yang[a,*], Chunping Li[a], Qiang Ding[b], Li Li[b]

[a] *Tsinghua National Laboratory for Information Science and Technology School of Software,Tsinghua University,Beijing 100084, China*
[b] *Shannon Lab, Huawei Technologies Co.LTD, Beijing 100095, China*

Abstract

In this paper, we propose a novel approach to classify short texts by combining both their lexical and semantic features. We present an improved measurement method for lexical feature selection and furthermore obtain the semantic features with the background knowledge repository which covers target category domains. The combination of lexical and semantic features is achieved by mapping words to topics with different weights. In this way, the dimensionality of feature space is reduced to the number of topics. We here use Wikipedia as background knowledge and employ Support Vector Machine (SVM) as classifier. The experiment results show that our approach has better effectiveness compared with existing methods for classifying short texts.

Keywords: Short text; Topic model; Wikipedia; Feature selection

1. Introduction

Text classification plays a very important role in many application domains. With the widespread of web applications such as social networks and online review systems, etc., we are now confronting much more short texts and news every day. Traditional text mining methods have their limitations for automatic classification of short texts, as the word sparseness in short texts, the lack of context information and informal sentence expressiveness.

A common method to overcome these problems for classifying short texts is to enrich the original texts with additional information. One way is to employ search engines and utilize the search results to expand related contextual content [1, 2, 3]. The other way is to utilize external repositories (e.g., Wikipedia and Open Directory Project, etc) as background knowledge [4, 5, 6, 7]. Although both of these two methods achieve improvement of short text classification to some extents, there is handicap to deal with amount of unrelated and noisy information if we naively expand original texts.

Probabilistic latent topic models [6, 8, 9, 10] have effectively been used in text mining. The basic idea of these kinds of models is to learn the topics from domain related datasets and assume each text is a multinomial

*Corresponding author. Lili Yang, School of Software,Tsinghua University,Beijing 100084, China.
E-mail address: yangll11@mails.tsinghua.edu.cn.

doi:10.1016/j.procs.2013.09.083

distribution over these topics. As the number of topics is relatively small, the dimensionality of each text thus becomes lower and the vector space of texts is no longer sparse. We observe that the probabilities of all topics are none-zero because these models must assure each text to have a probability to be generated by any of topics. This means that any a text has more or less relations with every topic. In real applications, however, a text may be bound up with a small number of topics and often has no relations with others at all. Only applying topic distribution has obvious limitations especially for dealing with short texts.

In this paper, we propose a topic model based approach which combines both lexical and semantic features to avoid the aforementioned limitations for short text classification. Like some existing methods, we also employ a background knowledge repository to learn topics with respect to all target categories. After we obtain all topics from the repository, we assign each word of short texts to the learned topics by making use of a Gibbs sampling method. That is to say, we would map each word occurrence to a particular topic and then we can represent a short text with these mapped topics instead. In this way, we can notice words in a short text may be mapped to a few but not all topics. Additionally, with respect to discriminative capacity of words, we adopt different mapping weights. For words coherent with a particular category, we assume that the topic which words assigned to is more closely related to the target category. We thus present the method of expected cross entropy based on lexical evidence for measuring the discriminative capacity of words in short texts. We evaluate the performance and effect of our proposed approach on both GoogleSnippet and Ohsumed datasets using Wikipedia as background knowledge. The experiment results show that our approach achieve better effectiveness compared to existing methods.

The remainder of this paper is organized as follows. In Section 2 the background and related works are introduce. In Section 3 we present our proposed approach in details. In Section 4, we show experiments and result analysis, respectively on two real-world datasets. We have a discussion in Section 5 and the concluding remarks in Section 6.

2. Related Work

One of the main challenges for text classification is the high dimensionality of feature space which not only would lead to high computational complexity but also is prone to overfitting problems. Plenty of feature selection measures have been put forward to reduce dimensionality in the past years, such as term frequency-inverse document frequency (TF-IDF), information gain (IG), mutual information (MI) and expected cross entropy (ECE), etc [11]. Documents are then represented by these selected features. By applying a classification model (K-Nearest Neighbor, Naive Bayes and Support Vector Machine) to the training set, we can obtain a classifier which could be employed to predict the category labels for future unseen documents. This type of classification methods is called lexical-based classification.

Semantic-based text classification springs up after topic models become popular for semantic analysis. [8] and [9] reduced the dimensionality of the feature space of a document to the number of topics by using topic distribution parameters for each document and then combined with traditional classifier to achieve classification. [12] associated each document with a single categorical label corresponding to a topic. [13] proposed a novel cross-domain text classification algorithm which extends traditional PLSA algorithm to integrate both labeled and unlabeled data into a unified probabilistic model. [14] designed one-to-one mapping between topics and labels and could be applied for multi-label classification.

While both types of aforementioned lexical-based and semantic-based classification are comparatively suitable for classifying long texts, problems for dealing with short text rise as short texts are constantly emerging nowadays. [15] proposed a method for extracting important topic words from a blog by measuring whether the blog includes rich content, which is achieved by comparing web search results of the candidate words with the content of the blog. [1] presented a contextual vector to represent each short text by using the L_2 normalization of the centroid of all results returned by a search engine. [16] utilized TAGME, a powerful tool to identify meaningful phrases for annotating short and poorly composed texts so as to help the understanding of short texts. [17] proposed a topics based similarity measurement method to select feature words based on both of the lexical weight and relationships of topics that words belonged to. [10] analyzed short texts by assuming each short text is associated with one certain topic.

Another line of short text mining techniques is to cooperate with external repository by combining with topic models. [4] presented a novel approach to cluster short text messages via transfer learning from auxiliary longer

textual data and applied the topic model, assumed that short texts and auxiliary texts have different generative processes. [5] presented a "universal dataset" based hidden topic analysis method which integrates topics and words by appending the words with respect to topics into feature vectors for building a classifier.

In this paper, we propose a novel way to combine the lexical and semantic features for classifying short texts, meanwhile retaining the dimensionality of feature space to be low. In our approach, short texts are often considered to be related with only a small number of topics and to overcome the limitation of traditional semantic-based classification methods.

3. Proposed Approach

Here we present our approach in details. The main process of our approach is as follows.

(1) Choose an external credible repository and extract some longer documents related to target categories as background knowledge.
(2) Apply topic model to these longer documents to learn a certain number of topics.
(3) Select discriminative feature words using our improved expected cross entropy as the measure.
(4) Map the weighted words of short texts to corresponding topics as the vector representations of short texts.
(5) Training the classification model on labeled data.

3.1. Category Topic Learning

In our approach, the topics related to the target categories are learned from a background knowledge repository. The choice of the repository is of great importance because its content should be abundant enough to cover topics related to categories as much as possible. After collecting related long texts, we apply topic model to learn topics from the background knowledge dataset.

Latent Dirichlet Allocation (LDA) [8] is a generative probabilistic model to learn the semantic topics from a corpus. The basic idea is that documents are represented as multinomial distributions over latent topics; meanwhile each topic is characterized by a multinomial distribution over words. The generative process of LDA is as follows.

- For each of the K topics k

  Draw word distribution of topic $\boldsymbol{\phi}_k \sim Dirichlet(\beta)$

- For each of the M documents m

  1 Draw topic distribution $\boldsymbol{\theta} \sim Dirichlet(\alpha)$
  2 For each of the N words w_n in document m
    2.1 Draw a topic $z_n \sim multinomial(\boldsymbol{\theta})$
    2.2 Draw the word $w_n|z_n \sim multinomial(\boldsymbol{\phi}_{z_n})$

In the generative process, α and β are called hyper-parameters determining the parameters of document-topics distribution and topic-words distribution separately. The graphic model of LDA is shown in Fig 1.

A Gibbs sampling approach is used to acquire the document-topics distribution θ and the topic-words distribution ϕ in this model. In our approach, we are interested in ϕ_{kt} (the probability word t assigned to topic k) as later the mapping of all the words in short texts to topics is based on this distribution. ϕ_{kt} could be acquired after training the background knowledge:

$$\phi_{kt} = \frac{n_{kt} + \beta}{n_k + V\beta} \tag{1}$$

where n_{kt} is the times word t assigned to topic k, n_k is the total number of words assigned to topic k and V is the vocabulary size. β is the hyper-parameter as explained in the generative process.

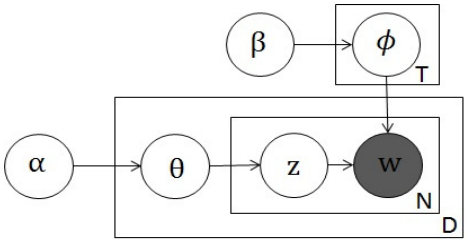


Fig. 1: Generative Process of LDA

3.2. Feature Selection

Expected cross entropy (ECE) is a kind of feature selection measure which considers both word frequency and the relationship between word and category. The bigger the ECE value, the larger impact the corresponding word has for the purpose classification. ECE value of word w is usually calculated as:

$$f(w) = p(w)\sum_i p(C_i|w)log\frac{p(C_i|w)}{p(C_i)} \tag{2}$$

where w represents for a word and C_i represents for category i.

Here, a further improvement consisting of two steps is made for this feature selection measure. First, based on the observation that in most cases, a representative word of category A may be not of great importance in category B, we think a word should have its different weights in different categories, while in Equation (2), a word has an overall weight in all the categories. Therefore, we could calculate the weight of a word in different categories separately as following:

$$f(w,C_i) = p(w|C_i)p(C_i|w)log\frac{p(C_i|w)}{p(C_i)} \tag{3}$$

From Equation (3), we can find if w has a strong relationship with category i, or the category is of small size, the word has higher possibility to have high weight with regard to category i.

Second, we think that most of distinctive words would have strong relationship with one category and less coherent with others. Equation (4) is used to measure the final weight of a word with regard to a category which we name it as **M-ECE** value.

$$F(w,C_i) = f(w,C_i) - \sum_{j\neq i} f(w,C_j) \tag{4}$$

The top-N distinctive words for each category are selected to represent lexical features. For these feature words, we use different mapping weights when combining with semantic features in our subsequent procedure.

3.3. Words Mapping with Weight

We here present the way to obtain the semantics from a text and reduce the dimensionality of feature spaces to the number of topics. Then we show how to combine lexical and semantic features together to represent a short text while retaining the dimensionality of feature space unchanged.

We first map the words of short texts to a learned topics. The Gibbs sampling approach is adopted. For every word in each text, we iteratively use Formula (5) to assign a topic to it.

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}, \bullet) \propto \frac{n_{mk}^{\neg i} + \alpha}{n_m + K\alpha} \cdot \phi_{kt} \tag{5}$$

where n_{mk} represents the number of words assigned to topic k in document m, and n_m represents the length of document m. $\neg i$ here means not including the current processing word and t represents the word in i-th position of document m, ϕ_{kt} is obtained from the topic learning of background knowledge and calculated by (1). K is

the number of topics and α is hyper-parameters as explained in LDA generative process. As we can see, the assignment of a word to a corresponding topic is influenced by both the context information of a text and the topic-words distribution ϕ obtained from background knowledge.

After finishing the word-topic assignment, we could get semantic representation of short text by using corresponding topics to take place of original words. Therefore all the texts could be represented with a vector, in which each element represents the times of the corresponding topic assigned to. Thus the dimensionality of feature space is reduced to the number of topics. As we noticed, several words in a text may be assigned to the same topic while some topics may not be assigned by any words within the text. Hence, the vectors we get have some zero elements and a small number of elements are much bigger than others.

Now we are going to combine the semantic features of a text with those lexical feature obtained by using the feature selection method we proposed. We assume if words are more related to some a category, the corresponding topics are more related to the category. Therefore, we came up with an idea by increasing the mapping weights of these words to assigned topics in order to put more emphasis on them. Here, mapping weight η refers to how many times a topic appear in the text if a word is assigned to it. For example, word *currency* is a lexical feature of category *Business*. If it is assigned to *topic 1*, then *topic 1* is considered to appear η times owing to an occurrence of word *currency* in the text. The calculation of η could be followed by Formula (6).

$$\eta \propto F(w, C_i), \eta > 1 \tag{6}$$

where $F(w, C_i)$ is the M-ECE value of word w with regard to category C_i. That is to say, the more important the word is in the category, the higher mapping weight it will get. Hence the corresponding topic is emphasized when representing a short text with topics.

After mapping words to topics with different weights, short texts could still be represented by all these learned topics as the elements of the vector are different comparing with only considering semantics as we have put more emphasis on topics.

4. Experiment and Analysis

4.1. Data Set

In order to evaluate our approach, we conduct experiments on two data sets. GoogleSnippet [1] Dataset contains the web search results related to 8 different domains. We choose 5 categories of the original dataset as our experimental dataset, details of this data set are shown in Table 1. Ohsumed [2] includes medical abstracts from the Medical Subject Headings (MeSH) categories of the year 1991, which is categorized into 23 categories. We also choose 5 categories of the original dataset and part of the abstracts of them as our second dataset. The statistics of Ohsumed are shown in Table 2.

Table 1: GoogleSnippet Dataset

Category	# Train	# Test	AveLen
Business	1200	300	16.34
Computer	1200	300	16.21
Health	880	300	15.96
Politics-Society	1200	300	15.53
Sports	1120	300	15.98

Table 2: Ohsumed Dataset

Category	# Train	# Test	AveLen
Cardiovascular	2000	500	153.98
Digestive	2000	500	109.57
Immunology	2000	500	118.36
Neoplasms	2000	500	105.20
Respiratory Tract	2000	500	100.98

As we can see, the average length (AveLen) of texts in GoogleSnippet dataset is only ca. 16 after preprocessing. Although abstracts of Ohsumed dataset are much longer, they still contain less word co-occurrence.

[1] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz
[2] http://disi.unitn.it/moschitti/corpora.htm

4.2. Background Knowledge Preparation

We chose Wikipedia as background knowledge as it is well organized, content abundant and with high quality. We crawled the background data from Wikipedia by taking the web page describing each category name as seed. We collected 1000 web pages for each category of the two datasets and removed the duplicated pages and the pages with less than 100 words after preprocessing.

The statistics of the background knowledge for all the categories are shown in Table 3 and Table 4. As we can see, the average length of the background dataset for almost all categories are more than 1000 thus could be considered as long texts and are suitable to learn topics from the them.

Table 3: Background Dataset for GoogleSnippet

Category	# webPages	AveLen
Business	642	1169.91
Computer	639	1058.13
Health	555	1265.29
Politics-Society	561	1326.02
Sports	458	1249.83

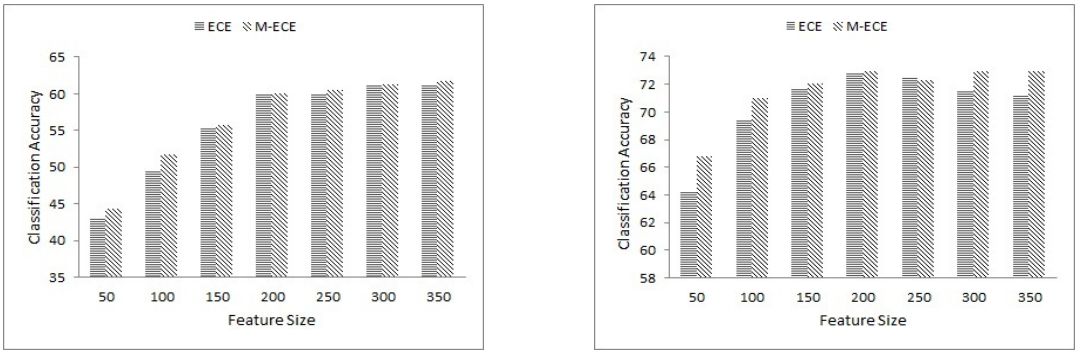
Table 4: Background Dataset for Ohsumed

Category	# webPages	AveLen
Cardiovascular	554	789.17
Digestive	457	1130.01
Immunology	570	1259.98
Neoplasms	607	986.40
Respiratory Tract	638	1003.05

4.3. Results and Analysis

4.3.1. Evaluation of M-ECE

We first did some experiments to show the effectiveness of improving the traditional ECE measure for feature selection. The evaluation was conducted on both GoogleSnippet and Ohsumed datasets. We selected different size of features ranging from 50 to 350 for both traditional ECE and our modified measure M-ECE and then applied support vector machine (SVM) for text classification. The classification accuracy is shown in Fig 2.



(a) GoogleSnippet (b) Ohsumed

Fig. 2: Classification Accuracy of Traditional ECE and M-ECE

We can see in Fig 2(a), in almost all cases of different feature sizes, M-ECE performs better than traditional ECE, except when feature size is 300, they perform the same. Specifically, when the feature size is small, this superiority is more obvious while traditional ECE and M-ECE have approximate performance when feature size becomes larger. Besides, we have noticed that when the number of features is more than 200, the accuracy keeps almost stable and we may infer 200 is the best feature size for lexical classification of this dataset. Fig 2(b) demonstrates similar results on Ohsumed dataset as in most cases M-ECE performs better than traditional ECE although the exception exists when the feature size is 250. Besides, the accuracy begins to decrease when feature size is larger than 200 if we use traditional ECE as selection measure. However, M-ECE achieves stable accuracy with the growing size of features. Hence, we can conclude that M-ECE is a more effective and stable than ECE measure for selecting features for text classification.

4.3.2. Evaluation of Proposed Approach

In order to evaluate the classification effect of the proposed approach, we did further experiments to compare our approach with existing methods. We first introduce these methods we used as well as how we did the experiments. Results are shown later.

ECE: Traditional approach using expected cross entropy as the measure for feature selection (according to Equation (2)). In our experiment, we selected 100 words with highest expected cross entropy value as features.

M-ECE: Our method based on feature selection with Equation (4) proposed in this paper. In our experiments, we selected the top 20 words from each category as both of our datasets have 5 different categories.

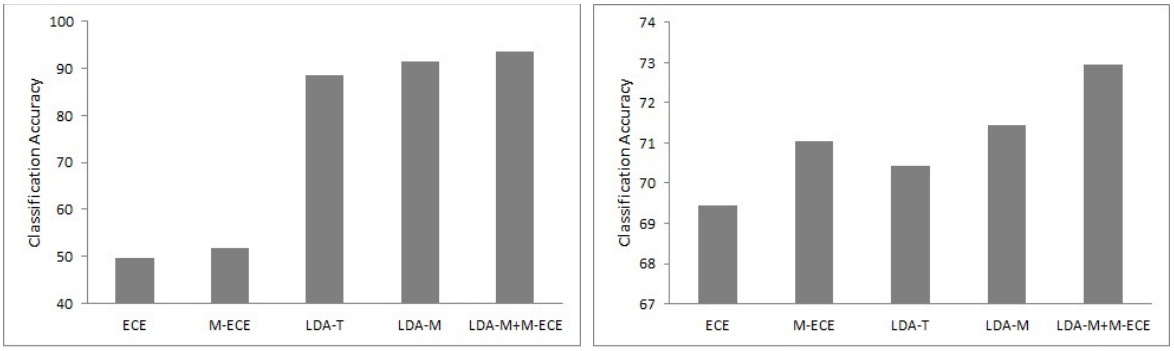
LDA-T: A semantic-based approach which first applied LDA to background knowledge to learn 100 topics with regard to all the target categories. Topic distribution of each short text was then inferred and used the distribution parameters to represent a short text as described in [8, 9]. Hence, in this way, each dimension is none-zero.

LDA-M: The semantic-based classification approach using our proposed words-topics mapping mechanism without considering mapping weights. After learning 100 topics from background knowledge, topic assignment of any words in texts was inferred and the assignment to some topics was considered as one appearance of the corresponding topics in the text. In this way, each text is associated with a small number of topics.

LDA-M+M-ECE: Our proposed approach by combining M-ECE and LDA-M. We first selected 100 lexical features according to Equation (4) and learned 100 topics with regard to all the target categories. Then we assigned words to topics at the same time considering weight η. In our experiments, we adopted a simplification by treating $\eta = 3$ for top 50 lexical important words and for those importance between 51 and 100, we set their weight as $\eta = 2$, others having weight of 1.

Of all the methods we mentioned above, the dimensionality of feature space was 100. After constructing a feature vector for each text, SVM was used to achieve classification. For those methods used LDA to obtain topics, we applied the empirical value of hyper-parameters as $\alpha = K/50$ and $\beta = 0.01$ in the model.

The classification results on GoogleSnippet and Ohsumed's in Fig. 3(a) and Fig. 3(b) respectively. Of both the datasets, we can notice an improvement of M-ECE over traditional ECE just as we have demonstrated in Subsection 4.3.1. Besides, we noticed that only considering the semantic features are not always helpful as we can see LDA-T achieves much better results in GoogleSnippet while performs worse in Ohsumed dataset compared to M-ECE. LDA-M, which uses our proposed words mapping mechanism, achieves better results than LDA-T although both of them are semantic-based methods using LDA to obtain topics with the help of background knowledge. The best performance of LDA-M+M-ECE shows the benefit of the combination of semantic features obtained by our mapping mechanism and lexical features obtained by the improved feature selection measure.



(a) GoogleSnippet (b) Ohsumed

Fig. 3: Classification Accuracy of Different Methods

4.3.3. Impact of Different Training Size

We keep experiments to see what impact the training size would have on classification results in our approach. So we re-divided both of the datasets randomly to form new training and testing sets of different sizes. The ratio of

training and testing size ranges from 0.25 to 4 and the classification results are shown in Fig. 4. It shows that even with a small labeled training set, our method could achieve higher accuracy. On GoogleSnippet dataset, even the training set is a quarter of testing set, the accuracy is 92.38% and the accuracy almost keeps stable with training sets becoming larger. On Ohsumed dataset, the accuracy is increasing very slowly with the size of training set becoming larger. Hence, we can conclude that the classifier built by the measure method we proposed has higher predictive capability and scales well along with the variation of training and testing size.

4.3.4. Impacts of Number of Topics

We here demonstrate how the number of topics would affect the classification accuracy in our approach. We repeatedly applied LDA to the background dataset and changed the number of topics from 40 to 160. The mapping step remained unchanged, and thus we constructed feature spaces of different dimensionality. The classification results are shown in Fig. 5.

As we can see, the accuracy is almost stable on GoogleSnippet, achieving the highest accuracy of 93.87% when the topic number is 60, and the lowest of 92.73% when topic number is 160. On Ohsumed dataset, the accuracy has a bit fluctuation with the change of topic number, especially when the number is small. Yet it keeps almost stable when topic number is bigger than 100. As a result, we conclude that accuracy is quite stable with respect to topic number in our approach.

Fig. 4: The Effect of Different Training Sizes

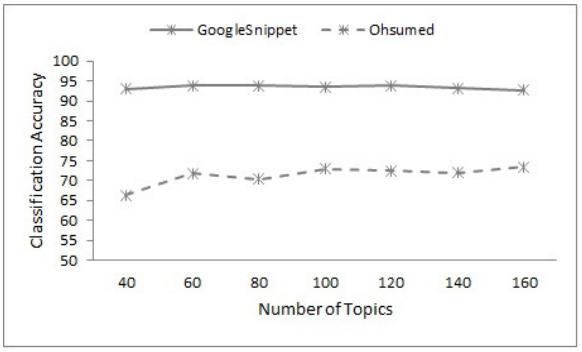


Fig. 5: The Effect of Topic Numbers

5. Discussion

Our proposed approach in this paper relies heavily on the right topics learned for the given corpus of short texts, implicating the importance of choosing the right background knowledge as well as how we acquire high-quality topics from it.

For the sake of simplicity, we constructed the background knowledge for both the experimental datasets by crawling web pages from Wikipedia, known as the richest online encyclopedia. However, other collections may be appropriate to act as the background knowledge for particular short text corpus, for example, related medical journals for dataset Ohsumed. Another issue is the quality of topics learned from the background knowledge. As far as we known, the HDP [18] model probably is more adaptable to learn high-quality topics as it can learn suitable number of topics automatically according to the nature of corpus without manually settings. Both of the issues are worth of further researches.

6. Conclusion

In this paper, we present a novel approach to combine lexical and semantic features for short text classification and also put forward a new measure method to select lexical features from short texts. Experimental results indicate both the improvement of the feature selection and classification for short texts. Our future work lies

in applying our proposed approach to other related text analysis and mining tasks. We are also interested in extending basic LDA model to the correlated topic models to extract not only some semantic features, but also the correlations between these features for mining short texts.

References

[1] Mehran Sahami , Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. Proceedings of the 15th international conference on World Wide Web, 2006.
[2] D Bollegala, Y Matsuo, M Ishizuka. Measuring semantic similarity between words using web search engines. Proceedings of the 16th international conference on World Wide Web, 2007.
[3] W. Yih and C. Meek. Improving similarity measures for short segments of text. Proceedings of the 22nd National Conference on Artificial Intelligence, 2007.
[4] Ou Jin , Nathan N. Liu , Kai Zhao , Yong Yu , Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011.
[5] Xuan-Hieu Phan , Le-Minh Nguyen , Susumu 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, 2008.
[6] Mengen Chen , Xiaoming Jin , Dou Shen. Short text classification improved by learning multi-granularity topics. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, p.1776-1781, 2011.
[7] Somnath Banerjee , Krishnan Ramanathan , Ajay Gupta. Feature Selection for Unbalanced Class Distribution and Naive Bayes. Proceedings of the Sixteenth International Conference on Machine Learning, p.258-267,1999.
[8] D. Blei , A. Ng , M. Jordan and J. Lafferty. Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, p.993-1022,2003.
[9] Yue Lu , Qiaozhu Mei , Chengxiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, v.14 n.2, p.178-203, 2011.
[10] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Volume 1, p. 536C544. 2012.
[11] Dunja Mladenic , Marko Grobelnik. Feature Selection for Unbalanced Class Distribution and Naive Bayes. Proceedings of the Sixteenth International Conference on Machine Learning, p.258-267, 1999.
[12] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In NIPS, volume 22, 2008.
[13] Gui-Rong Xue , Wenyuan Dai , Qiang Yang , Yong Yu. Topic-bridged PLSA for cross-domain text classification. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008.
[14] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In EMNLP '09: Proceedings of the Conference on Empirical Methods in Natural Language Processing, p.248-256, 2009.
[15] Jinhee Park,Sungwoo Lee, Hye-Wuk Jung and Jee-Hyong Lee. Topic word selection for blogs by topic richness using web search result clustering. Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication, 2012.
[16] Paolo Ferragina , Ugo Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). Proceedings of the 19th ACM international conference on Information and knowledge management, 2010.
[17] Xiaojun Quan , Gang Liu , Zhi Lu , Xingliang Ni , Liu Wenyin. Short text similarity based on probabilistic topics. Knowledge and Information Systems, v.25 n.3, p.473-491,2010.
[18] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. Technical Report, Department of Statistics, UC Berkeley, 2004.