Short Text Feature Enrichment Using Link Analysis on Topic-Keyword Graph

Peng Wang, Heng Zhang, Bo Xu, Chenglin Liu, and Hongwei Hao

Institute of Automation of the Chinese Academy of Sciences
95 Zhongguancun East Rd.
Beijing, 100190 P.R. China
{peng.wang,heng.zhang,boxu,chenglin.liu,hongwei.hao}@ia.ac.cn

Abstract. In this paper, we propose a novel feature enrichment method for short text classification based on the link analysis on topic-keyword graph. After topic modeling, we re-rank the keywords distribution extracted by biterm topic model (BTM) to make the topics more salient. Then a topic-keyword graph is constructed and link analysis is conducted. For complement, the K-L divergence is integrated with the structural similarity to discover the most related keywords. At last, the short text is expanded by appending these related keywords for classification. Experimental results on two open datasets validate the effectiveness of the proposed method.

Keywords: topic Model, Short Text, Feature Enrichment, Link Analysis, SimRank, K-L divergence.

1 Introduction

With the advent of the era of big data, mass of short texts have been generated on the web and mobile applications, including search snippets, micro-blog, products review, and short messages. The classification of these short texts plays an important role in understanding user intent, question answering and intelligent information retrieval [1, 2]. Currently, how to acquire the effective representation of short text and enhance the categorization performance have been an active research issue and have drawn much attentions [3, 4].

Short text bring about data sparsity and ambiguity problems because they cannot provide enough word co-occurrence or contextual information [3]. Thus, the general text mining methods based on bag of words cannot apply directly to short texts because they ignore the semantic relations between words [2]. Moreover, some essentially related short texts may have very little overlapping keywords, which seriously affects the similarity measurement and the categorization performance [5].

To solve these problems, some methods have been proposed to expand the representation of short text using latent semantics or related words. The expansion information is derived from the training datasets internally [6], or from external corpus such as Wikipedia [3]. Zhang et al. [5] proposed a graph-based text similarity measurement and exploited background knowledge from Wikipedia to find semantic

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 79–90, 2014.

affinity between documents. Phan et al. [4] presented a general framework to expand the short and sparse text by appending topic names discovered using latent Dirichlet allocation (LDA) [7].

With Search Engine, Sun A. [1] proposed a simple method for short text categorization by selecting the most representative words as query to search a few of labeled samples, and the majority vote of the search results is the predictable category. Sahami and Heilman [8] enrich the text representation by web search results using the short text segment as a query.

In this paper, we propose a method using link analysis on topic-keyword graph for enriching the feature representation of short text to overcome the sparsity and semantic sensitive problems. First, we apply the biterm topic model (BTM) [9] to extract topics. Then, we re-rank the keywords distribution under each topic according to an improved TFIDF-like score [10]. Finally, a topic-keyword graph is constructed to prepare for link analysis.

On the topic-keyword graph, we use the link analysis algorithm—SimRank [11,12] to compute the structural similarity between every two nodes. Further, the K-L divergence of topics distribution is integrated with the structural similarity to discern the most similar keywords. The short text is expanded by appending these similar keywords for classification. Our method can avoid noise (class irrelevant) words and extract salient keywords by synthesizing the semantic knowledge and link structure information. Experiments are conducted on search snippets and 20Newgroups to validate the effectiveness of the method proposed.

The rest of this paper is organized as follows. Section 2 briefly reviews the relevant works. Section 3 introduces our method of short text representation enrichment. Section 4 presents the experimental results on two open datasets. Finally, concluding remarks are offered in Section 5.

2 Related Work

Short text classification is a challenge due to its noise words, lacking of sufficient contextual information and the semantic sensitive problem [3]. Thus, traditional statistical methods usually fail to achieve satisfactory classification performance [2].

In recent years, some research focuses on how to utilize large-scale external data to explore semantic information, and enrich the original text to help text mining [13]. Gabrilovich and Markovitch [14] proposed a method to improve text classification performance by enriching document representation with Wikipedia concepts. Zhang et al. [5] presented a graph-based text similarity measurement using SimRank, which is the most related work to our study. Based on the background knowledge from Wikipedia, they build a document-concept bipartite graph to compute the affinity of different documents and then perform text classification.

SimRank [11] is applicable to any domain with node-to-node relationships for measuring the similarity between nodes. The motivation of SimRank is that two nodes are similar if they are related by similar nodes. Due to the latent semantics and link structure information embedded in the graph representation, the SimRank algorithm can be used as a similarity measurement on the semantic graph.

Based on consistent Wikipedia corpus, Phan et al. proposed a method to discover hidden topics using LDA and expand the original text [4]. Zhu et al. [15] proposed to use multi-original external corpus to model topics for better categorization performance. This method can draw more broad and accurate topics compared to that based on one external corpus. Chen et al. [3] pointed out that leveraging topics at multiple granularity can model short texts more precisely.

In order to overcome the insufficiency of LDA in modeling short text in terms of the document-level word co-occurrence, Yan et al. [9] presented a new variant of topic model—bi-term topic model (BTM), for extracting topics using bi-terms existing in the whole training dataset instead of document-level, which can well alleviate the problem of sparsity.

Unlike the probabilistic formulation topic model [7, 16] with the constraints that the admixture proportions of topics and likelihood function should be normalized, Zhu and Xing [17] presented a non-probabilistic one named sparse topical coding (STC), which can control the sparsity of inferred result directly. STC model achieved the state-of-the-art classification accuracy on 20Newgroups dataset.

3 Short Text Feature Enrichment Method

Our method is aimed to identify the topical-indicative words automatically from training dataset, and expand the short text for classification. Based on BTM and SimRank, a unified feature enrichment method named SRBTM is proposed.

3.1 Overview

After topic modeling, the keywords under each topic are used to produce a thesaurus by re-ranking algorithm. Then, we construct a topic-keyword bipartite graph and compute the similarity between keywords using link analysis combined with K-L divergence. Finally, these most related keywords are selected to expand the original short text. The detailed framework is shown in Figure 1.

According to the thesauri produced in re-rank stage, we search seed words appeared in the short text, and all keywords mentioned in the thesauri are used as candidate words. Then, we present a novel similarity measurement in Equation (1) to compute the affinity between seed words and candidate words. For each seed word, we select top- v candidate words with maximum $CScore$ to enrich the short text.

$$CScore(sw_i, cw_j) = \frac{SR(sw_i, cw_j)}{KL(sw_i, cw_j)}, \tag{1}$$

$SR(sw_i, cw_j)$ is the SimRank score between seed word i and candidate word j, and $KL(sw_i, cw_j)$ is the K-L divergence.

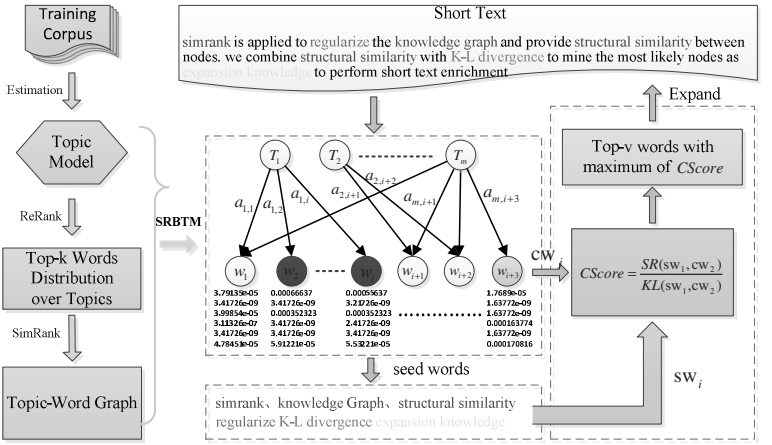


Fig. 1. The framework based on topic-keyword graph and link analysis

The merits of our unified framework are: (1) adding the most related topic keywords to short text can resolve the sparsity problem, and we can alleviate the synonymy and polysemy problems meanwhile; (2) the added keywords are new features for original short text which can be leveraged for categorization. Thus, topic related texts which do not share common seed words literally may be enriched with the same keywords. Then, the computation of similarity on these texts can be improved and classification performance will be enhanced; (3) BTM model short text at corpus-level, which reflects the global information. However, the keywords are added according to the content of each short text, which exploit the local information. So we synthesize the global information and local information to guarantee the effective enrichment.

3.2 Biterm Topic Model

With respect to topic extraction, documents are modeled as mixtures of latent topics and each topic can be further represented by a set of keywords. To overcome the data sparsity problem, Yan et al. proposed the biterm topic model (BTM) [9] especially for short text, which directly model the word co-occurrences in the whole corpus to make full use of the global information. The parameters and variables used in BTM are listed in Table 1.

Briefly, BTM as one of the improved variants of LDA, can effectively alleviate the data sparsity in modeling short text. Moreover, as demonstrated in Table 2 that the re-ranked keywords from each topic are high related, BTM make the representation of short text more topic-focused.

Table 1. Variables in BTM

Para.	Details
M	number of bi-terms
α,β	Para. for Dirichlet
$\vec{\theta}$	topic distribution
z	index of a topic
$\overrightarrow{\varphi_{iz}}$	i th word distribution
V	vocabulary size
K	the number of topics
$\mathbf{\Phi}$	a $K \times V$ matrix
BT	corpus with M biterms

Table 2. Most likely words of some topics

Topic0:music band rock album song songs released
Topic1:species food animals animal plants humans
Topic2: energy mass field quantum particles force
Topic3: india indian hindu pakistan sanskrit century
Topic4: blood body brain heart cells muscle syndrome
Topic5: water carbon oil chemical gas process oxygen
Topic6: government party president constitution
Topic7: power energy solar electric electrical
Topic8: ystem data code software computer
Topic9: horse opponent horses body hand match
Topic10: south africa united country islands world

3.3 Re-ranking Method

The keywords extracted by BTM are denoted as $\mathbf{\Phi} = \{\overrightarrow{\varphi_z}\}$. $\overrightarrow{\varphi_z} = [\varphi_{z,1}, \varphi_{z,1}, \cdots, \varphi_{z,V}]$ is the word distribution vector with the length of V under topic z. Song et al. [10] presented a keyword re-ranking algorithm for LDA-based topic modeling. Word distribution $\mathbf{\Phi} = \{\overrightarrow{\varphi_z}\}$ is applied to compute a TFIDF-like score for each keyword, and the original order of keywords is re-ranked.

To better identify salient information, we improve the re-ranking method in [10] for ranking BTM-derived keywords to refine the topic definitions. Different from the method in [10], we propose to use $\exp(\varphi_{z,i})$ as frequency-like term to compute the saliency score in (2),

$$SAS = \frac{e^{\varphi_{z,i}}}{\sum_{m=1}^{M} e^{\varphi_{z,i}}}, \tag{2}$$

where $\varphi_{z,i}$ is the probability distribution of the i th word under topic k. Then, the re-ranked results demonstrate in Table 2. Each line is a thesaurus corresponding to a specific topic which forms a clique in Figure 2, which show that our proposed method can make each topic more salient.

However, Figure 2(b) shows that there are still some keywords keep relevant to more than one topic. For short text expansion, these keywords may be noisy rather than be useful in providing discriminative information.

3.4 Topic-Keyword Graph Construction

After re-ranking the keywords for each topic, we construct a topic-keyword semantic graph as shown in Figure 2 and Figure 3. The semantic graph is built using topics and keywords as nodes, which are derived from BTM. Specifically, the hub node (or parent node) of each clique refers to a topic name. The top-k keywords in the thesaurus under a specific topic are selected as leaf nodes to form the clique. Then, all the

(a)native re-ranking method (b) our improved re-rank method

Fig. 2. Topic-Keyword Graph

cliques constitute the topic-keyword graph, which synthesize the semantic knowledge and link structure information.

The semantic graph is weighted by the score computed previously for keywords re-ranking using Equation (2). The weights are put on the edge from the hub nodes to leaf nodes as shown in Figure 3, which will be used to carry out link analysis and compute the structural similarity.

3.5 Link Analysis

Inspired by the underlying idea of link analysis algorithm—SimRank [11] that two objects are similar if they are related to similar objects, we propose a graph-based method to select the related keywords for short text enrichment. As in Figure 3, for a node w in the bipartite graph, we denote the set of its in-neighbors by $I(w)$, and individual in-neighbors are denoted as $I_i(w)$, for $1 \le i \le |I(w)|$. The SimRank score between node w_a and w_b is computed by

$$s(w_a, w_b) = \begin{cases} 1, & \text{if } w_a = w_b \\ \dfrac{C}{|I(w_a)||I(w_b)|} \sum_{i=1}^{|I(w_a)|} \sum_{j=1}^{|I(w_b)|} s(I_i(w_a), I_j(w_b)), & \text{if } w_a \neq w_b \end{cases}, \tag{3}$$

Where $C \in (0,1)$ is a decay factor. Specifically, the SimRank score is defined to be 0 when $|I(w_a)| = \phi$ or $|I(w_b)| = \phi$. According to (3), SimRank is symmetrical that $s(w_a, w_b) = s(w_b, w_a)$. Additionally, SimRank is an iterative fix-point algorithm, and its time complexity is $O(knd)$, where k is the number of iterations, n is the number of nodes, and d is the average of the in-degree of leaf nodes.

In our case, the weights on the edges in Figure 3 indicate the salient level of the corresponding keyword under the specific topic. However, the native SimRank algorithm fails to properly utilize the weights to enhance the possibility of selecting the

most representative keywords to enrich short text for classification. So we propose to use (6) to compute the topical SimRank score,

$$SR(w_a, w_b) = SAS(w_a)SAS(w_b)s(w_a, w_b)\,, \tag{4}$$

In Figure 3, using the modified SimRank, we can obtain that w_2 is more similar to w_i than w_1, because w_1 is shared by more than one topic and with low salient level. This merit ensures that we can expand short text and try our best to avoid introducing noise.

As Figure 4 shows that the keywords distribution under topics follow the long-tail distribution. Some keywords may shared by many topics because they are relevant to all of the topics. Fortunately, the modified SimRank can be used to further purify these shared keywords.

3.6 Short Text Expansion

With the aim of resolving the sparsity problem in short text feature representation and avoiding noise, we propose to discover topic keywords to enrich the original short text. For classification task, the keywords shared by many topics are considered to be noise or with little discriminability. Thus, the most likely candidate keywords selected as expanding information are these that with few topics.

As in Figure 1, each leaf node is corresponding to a topic distribution, which is a column vector in the matrix $\mathbf{\Phi} = \{\overline{\varphi_z}\}$. For achieving high reliability, the K-L divergence of the topics distribution on the candidate keywords is integrated with SimRank score using Equation (1) to discern the most similar keywords.

In order to obtain $CScore(sw_i, cw_j) = CScore(cw_j, sw_i)$, we apply symmetrical-version of K-L divergence as in (5),

$$KL(sw_i, cw_j) = \frac{1}{2}[D(p_{sw_i}^{(z)} \,\|\, \frac{p_{sw_i}^{(z)} + p_{cw_j}^{(z)}}{2}) + D(p_{cw_j}^{(z)} \,\|\, \frac{p_{cw_j}^{(z)} + p_{sw_i}^{(z)}}{2})] \tag{5}$$

Where $D(p \,\|\, q) = \sum_k p_k \log \frac{p_k}{q_k}$.

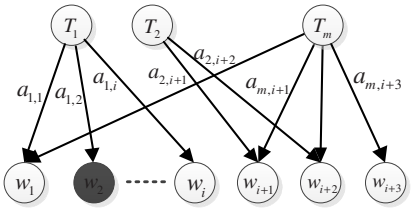


Fig. 3. Topic-keyword bipartite subgraph

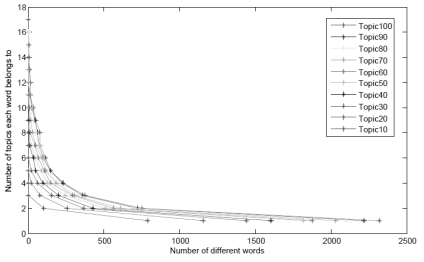


Fig. 4. The keywords distribution under topics

Finally, for each seed word, all the likely candidate keywords discovered from the topic-keyword graph are ranked in decreasing order according to Equation (1) and the top-v candidates are selected for short text enrichment. Our proposed method is a graph-based model that exploits topics as background knowledge and synthesizes both semantic structure representation and similarity computation. Experimental results show that our method is effective and can outperform the state-of-the-art techniques.

4 Experiments

To validate the effectiveness of our proposed method, we conducted experiments on two real-world datasets: Search snippets and 20Newsgroup.

Search snippets dataset, collected by Phan X. H. [4], consists of 10,060 training snippets and 2,280 test snippets from 8 categories, as shown in Table 3. On average, each snippet has 18.07 words.

20Newsgroups is a standard corpus including 18,846 messages from 20 different Usenet newsgroups. Each newsgroup is corresponding to a different topic.

4.1 Evaluation on Search Snippets

Based on search snippets dataset, we choose MaxEnt and LibSVM as classifiers to evaluate the qualities of our methods for feature representation. Among various machine learning methods, MaxEnt and SVM have been successfully applied in many text mining tasks [18], which proves that MaxEnt is much faster in both training and inference while SVM is more robust. For comparisons, we employ LDA as the usage in [4] as baseline.

In baseline, we firstly learn LDA model based on external Wikipedia corpus as [4] to estimate its parameters. Then, the latent topics drawn by LDA are used to expand original short text. For evaluation, the topics distribution corresponding to each snippet is fed to LibSVM classifier, and the expanded snippets with latent topics are leveraged as new features for MaxEnt classifier. Meanwhile, BTM is trained directly on the search snippets, and our method SRBTM as shown in Figure 1 is applied to discover related keywords to enrich the short snippets.

In the proposed method SRBTM, we select the top-v candidate keywords as enriching features when topic number is a constant k. In order to provide substantial results to prove the effectiveness of our method, we assign $v = [0,1,2,\cdots 10]$ for a fixed topic number.

Then, we compute the average and variance of classification accuracy when v changes, and the result are shown in Figure 6(a). We can find that our method outperforms the baselines obviously, and obtain the highest accuracy of 0.8678 when k =10, v =9, which reduce classification error by 10.01% compared to [3] and by 25.52% compared to [4].

Table 3. Details of search snippets

Domain	Tr_snippets	Te_snippets
Business	1200	300
Computers	1200	300
Cult.-arts-ente.	1880	330
Edu.-Science	2360	300
Engineering	220	150
Health	880	300
Politics-Society	1200	300
Sports	1120	300
Total	10060	2280

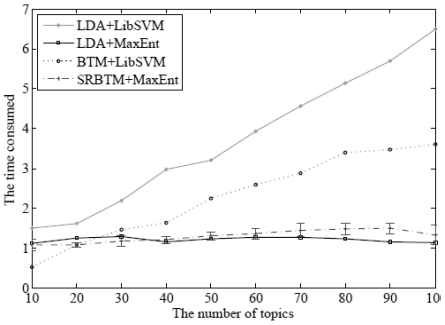
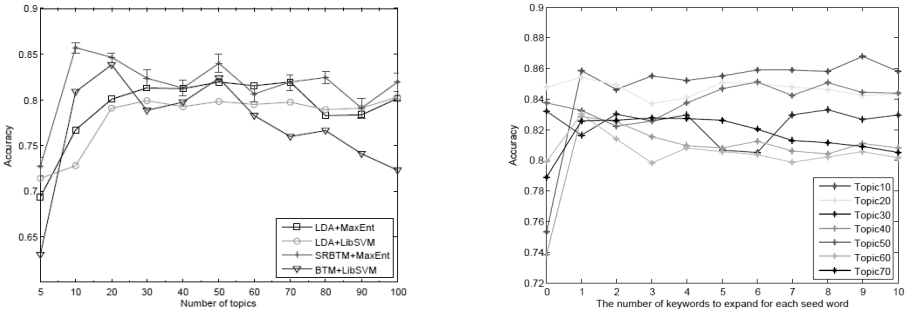


Fig. 5. Time consumed in test stage

When v varies from 0 to 10, the classification results of MaxEnt classifier are demonstrate in Figure 6(b). It is clear that enriching text representation using related keywords can indeed introduce useful information and achieve significant quality enhancement in the classification performance. However, when v is large enough, no more useful information can be added. Especially, when topic number k =10, we can obtain best result.

At last, the time consumed of classifiers over variational topics in prediction stage are compared in Figure 5. It is obviously that the MaxEnt classifier is faster and more stable than LibSVM. The LibSVM classifier will consume more time with the increase of the dimensionality of features. Our method proposed in this paper to enrich short text consume comparative time with that of LDA.



(a) (b)

Fig. 6. Accuracy vary with topic numbers and expanded keyword

4.2 Evaluation on 20Newsgroups

In previous experiments, we have demonstrated the effectiveness of SRBTM on short texts. Although we propose SRBTM for enriching the representation of short text, there is no limitation for our method to be applied on normal text. Therefore, it is also interesting to see whether SRBTM can alleviate the Synonyms and Polysemy problem

in normal text, and further enhance the classification performance. For this purpose, we compared SRBTM with sparse topical coding (STC), proposed by Zhu and Xing [17]. Based on 20Newsgroups, SRBTM is applied to identify topic-focused keywords and expand the original features.

Table 4. Accuracy vary with topic numbers on 20NG

Topic num. / Method	10	20	30	40	50
STC	0.70	0.757	0.789	0.809	0.817
SRBTM+Liblinear	0.8196 (±0.0026)	0.8170 (±0.0018)	0.8126 (±0.0047)	0.8117 (±0.0037)	0.8073 (±0.0066)
Topic num. / Method	60	70	80	90	100
STC	0.788	0.818	0.812	0.784	0.775
SRBTM+Liblinear	0.8090 (±0.0054)	0.8082 (±0.0022)	0.7467 (±0.1895)	0.8054 (±0.0038)	0.7995 (±0.0090)

In our experiments, we employ the output of SRBTM to learn Liblinear classifier, and with STC as baseline. Then, the comparison results are given in Table 4. As the similar settings to the prior experiments, we compute the average and variance of accuracy when v changes from 0 to 10. The results indicate that our approach can introduce discrimination information to some extent on normal text. When k =10, we achieve the highest average accuracy, which is consistent to the validation on short text. To interpret this result, we can find the underlying foundation from Figure 4 that with the increase of k , the number of keywords shared by many topics is increasing, which hurt the quality of likely candidate keywords discovered by SRBTM.

5 Conclusion and Future Work

In this paper, we presented a novel method to enrich short text for classification based on topic-keyword graph and link analysis. The topics drawn from the original short and noisy texts are mapped as cliques to form the topic-keyword graph, which depict the semantic structure of the whole corpus. Then, the link analysis algorithm—SimRank is applied to compute similarities between nodes from the link structure perspective. Finally, K-L divergence is incorporated with the output of SimRank to reliably select candidate words to perform short text enrichment.

The main contributions of our work are: (1) we improve the TFIDF-like score in [10] and use it to re-rank the BTM-derived topic keywords to refine the topic definitions and improve the coherence, interpretability and ultimate usability of learned topics; (2) based on the re-ranked topic keywords, a semantic graph is constructed and a topic weighted SimRank method is proposed to measure the affinity among nodes;

(3) a novel comprehensive similarity measurement is proposed to select the most related keywords for short text expansion without the large-scale external corpus.

In the future, we will study techniques to fully exploit the topic-keywords graph to reduce noise, and to combine the vector representation of words [19] to further improve the short text classification performance.

References

1. Sun, A.: Short text classification using very few words. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1145–1146. ACM (2012)
2. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842. ACM (2010)
3. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume, vol. 3, pp. 1776–1781. AAAI Press (2011)
4. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International Conference on World Wide Web, pp. 91–100. ACM (2008)
5. Zhang, L., Li, C., Liu, J., Wang, H.: Graph-based text similarity measurement by exploiting Wikipedia as background knowledge. World Academy of Science, Engineering and Technology 59, 1548–1553 (2011)
6. Hu, X., Sun, N., Zhang, C., Chua, T.S.: Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 919–928. ACM (2009)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
8. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th International Conference on World Wide Web, pp. 377–386. ACM (2006)
9. Yan, X.H., Guo, J.F., Lan, Y.Y., Cheng, X.Q.: A biterm topic model for short texts. In: Proceedings of the 22nd International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 1445–1456 (2013)
10. Song, Y., Pan, S., Liu, S., et al.: Topic and keyword re-ranking for LDA-based topic modeling. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1757–1760. ACM (2009)
11. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
12. Antonellis, I., Molina, H.G., Chang, C.C.: Simrank++: query rewriting through link analysis of the click graph. Proceedings of the VLDB Endowment 1(1) (August 2008)
13. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. IJCAI 7, 1606–1611 (2007)
14. Evgeniy, G., Markovitch, S.: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In: AAAI, vol. 6, pp. 1301–1306 (2006)

15. Zhu, Y., Li, L., Luo, L.: Learning to classify short text with topic model and external knowledge. In: Wang, M. (ed.) KSEM 2013. LNCS, vol. 8041, pp. 493–503. Springer, Heidelberg (2013)
16. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
17. Zhu, J., Xing, E.P.: Sparse topical coding. arXiv preprint arXiv:1202.3778 (2012)
18. Berger Adam, L., Pietra, V.J.D., DellaPietra, S.A.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (1996)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)