ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201806.00109

Labeled Bilingual Topic Models for Cross-lingual Text Classification and Label Recommendation: A Postprint

Authors: Tian Mingjie, Cui Rongyi

Date: 2018-06-19T00:00:00+00:00

Abstract

To address the increasingly abundant cross-lingual textual information resources and multi-label data present in news reports and scientific literature, and to mine the correlations between cross-lingual information as well as the associations among data attributes, we propose a labeled bilingual topic model for cross-lingual text classification and tag recommendation. First, assuming that keywords in scientific literature exhibit content relevance with their abstract sections, we extract and label these keywords, subsequently mapping the labels to topics within the topic model to instantiate the "latent" topics. Second, the abstract sections are trained iteratively using the labeled bilingual topic model. Finally, cross-lingual text classification and tag recommendation are performed on newly added documents. Experimental results demonstrate that the Micro-F1 reaches 94.81% on the cross-lingual text classification task, and the recommended tags effectively exhibit semantic relevance.

Full Text

Preamble

Title: Research on Labeled Bilingual Topic Model for Cross-Lingual Text Classification and Label Recommendation

Authors: Tian Mingjie, Cui Rongyi†

Affiliation: Intelligent Information Processing Laboratory, Department of Computer Science & Technology, Yanbian University, Yanji, Jilin 133002, China

Abstract: With the increasing abundance of multilingual information resources and multi-label data in news reports and scientific literature, this paper proposes a labeled bilingual topic model to mine correlations between languages

and associations among data attributes for cross-lingual text classification and label recommendation. First, assuming that keywords in scientific literature are content-related to their abstracts, we extract and label these keywords, aligning them with topics in the topic model to instantiate "latent" topics. Second, we train the abstract sections iteratively using the proposed labeled bilingual topic model. Finally, we perform cross-lingual text classification and label recommendation for newly added documents. Experimental results demonstrate that the Micro-F1 measure reaches 94.81% in cross-lingual text classification tasks, and the recommended labels effectively reflect semantic relevance.

Keywords: topic model; label; cross-lingual text classification; label recommendation; latent topic

0 Introduction

The proliferation of the Internet has ushered society into an era of information explosion, making effective management and mining of massive information resources critically important. Today' s information resources are not only growing rapidly in scale but also becoming increasingly diverse in type and language. While this linguistic diversity enriches information resources, language differences inevitably hinder users' ability to utilize them effectively. In this context, cross-lingual text classification technology is needed to organize multilingual information resources systematically and address the problem of information chaos. Cross-lingual text classification leverages labeled training documents in one language to train a classifier for categorizing unlabeled documents in another language. Compared to traditional text classification, cross-lingual text classification is a relatively new research area that started later.

In 2003, Bel et al. [1] formally introduced the concept of cross-lingual text categorization, defining it as the extension of existing text classification systems from monolingual to two or more languages without human intervention. Researchers have since proposed various approaches based on bilingual dictionaries, machine translation, and latent topic models. Bel et al. [1] constructed category features from the top n words of source language documents for each category, then used a bilingual dictionary to translate unclassified documents into the target language for similarity-based classification. Olsson et al. [2] employed probabilistic bilingual dictionaries to translate English training documents into Czech. Machine translation-based methods translate all texts into another language before classification, either translating the source language training set into the target language or vice versa. Rigutini et al. [3] combined machine translation with EM algorithms for cross-lingual text classification between English and Italian. Wei et al. [4] used machine translation to translate pivot features in structural correspondence learning for cross-lingual sentiment classification. Mimno et al. [5] proposed the PLTM topic model for modeling parallel and comparable corpora, applying it to machine translation and cross-lingual

topic tracking. Ni et al. [6] mined multilingual topics from Chinese-English comparable Wikipedia corpora using LDA, projecting multilingual texts into a latent topic space for cross-lingual text classification.

Dictionary-based methods suffer from word ambiguity and difficulties in selecting the top n feature words for translation, and they cannot handle out-of-vocabulary domain terms. While machine translation systems provide richer semantic information than bilingual dictionaries, their accuracy significantly impacts classification quality. Latent topic model-based approaches suffer from poor interpretability, as each latent topic lacks explicit definition. This paper leverages multi-label information from scientific literature and news reports, combined with LDA topic models for cross-lingual text processing, to propose a Labeled Bilingual Topic Model (LBTM). Our approach addresses the ambiguity of latent topics in conventional LDA by giving topics explicit semantics and better interpretability. By mining shared "topics" across multiple documents, we discover inter-document relationships and correlations, enabling label recommendation for new documents.

LDA Topic Model

The LDA topic model [7], proposed by Blei in 2003, is a document topic generation model and a three-level Bayesian probabilistic model comprising words, topics, and documents. Blei posited that each word in a document is generated by first selecting a topic with a certain probability, then selecting a word from that topic. Documents follow a multinomial distribution over topics, and topics follow a multinomial distribution over words. This assumption facilitates dimensionality reduction in large-scale data processing by projecting documents into a topic space. LDA introduces Dirichlet prior parameters [8] on the document-topic and topic-word distributions, addressing overfitting issues in large corpora.

LDA is a typical directed probabilistic graphical model [9]. Its graphical model is shown in [Figure 1: see original paper]. The variable $w_{m,n}$ represents the n-th word in the m-th document; α and β are hyperparameters of the Dirichlet distribution set empirically; $z_{m,n}$ is the topic assigned to word $w_{m,n}$; θ_m is the topic distribution for document m; and ϕ_k is the word distribution for topic k. Given a document collection D consisting of M documents, where the m-th document contains N_m words, and assuming K topics in D, LDA generates documents as follows:

1. Draw document length $N_m \sim \text{Poisson}(\xi)$
2. Draw topic distribution $\theta_m \sim \text{Dirichlet}(\alpha)$
3. For each word position $n = 1$ to N_m: 3.1. Choose a topic $z_{m,n} \sim \text{Multinomial}(\theta_m)$ 3.2. Choose a word $w_{m,n} \sim \text{Multinomial}(\phi_{z_{m,n}})$

Here, Poisson$(\cdot)$, Dirichlet$(\cdot)$, and Multinomial$(\cdot)$ denote Poisson, Dirichlet, and multinomial distributions, respectively.

Model parameter estimation is required during LDA construction. Common

methods include variational Bayesian inference [10,11], EM algorithms [12], and Collapsed Gibbs sampling [13]. Gibbs sampling is intuitive and simple to implement, effectively sampling topics from large document collections. Its parameter estimation can be viewed as the inverse of document generation: estimating parameters given the document collection. From the graphical model, the probability of a document is:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

By integrating out the parameters to be estimated, we can sample topics for each word instead. The conditional probability of the topic sequence given the word sequence is:

$$p(\mathbf{z}|\mathbf{w},\alpha,\beta) = \frac{p(\mathbf{w},\mathbf{z}|\alpha,\beta)}{p(\mathbf{w}|\alpha,\beta)}$$

The collapsed Gibbs sampling formula is:

$$p(z_i = k|\mathbf{z}_{-i},\mathbf{w}) \propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{v=1}^{V} n_{k,-i}^{(v)} + \beta_v} \cdot \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_{m,-i}^{(j)} + \alpha_j}$$

where z_i is the topic for the i-th word; $\mathbf{z}_{-i}$ denotes all topic assignments except the i-th; $n_{k,-i}^{(t)}$ is the count of word t assigned to topic k excluding the current assignment; β_t is the Dirichlet prior for word t; $n_{m,-i}^{(k)}$ is the count of topic k in document m excluding the current assignment; and α_k is the Dirichlet prior for topic k.

Once topic assignments for all word tokens are obtained, parameters can be computed as:

$$\phi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{v=1}^{V} n_k^{(v)} + \beta_v}$$

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_m^{(j)} + \alpha_j}$$

where $\phi_{k,t}$ represents the probability of word t in topic k, and $\theta_{m,k}$ represents the probability of topic k in document m.

2 Labeled Bilingual Topic Model

LDA represents high-dimensional word information through low-dimensional "latent" topics, capturing document semantics. However, each "latent" topic lacks explicit meaning and interpretability. This paper improves LDA by proposing the Labeled Bilingual Topic Model (LBTM), which utilizes multi-label information (e.g., keywords in papers) from scientific literature and news reports. LBTM treats labels as topics, giving topics explicit semantics and interpretability by instantiating "latent" topics with concrete 内涵. Documents modeled with LBTM are represented by "explicit" topics, providing more specific descriptions, while these explicit topics effectively represent the document collection' s word set. Through LBTM modeling, each document in the collection obtains a probability distribution over "explicit" topics, represented as vectors in vector space models to enable cross-lingual text classification and label recommendation.

2.1 Basic Idea

Assume a document collection consists of M documents, each described in two languages L_1 and L_2 with identical content. LBTM uses a set of language-independent "universal" topics to model both language representations, where each universal topic has two language-specific representations. The probabilistic graphical model of LBTM is shown in [Figure 2: see original paper].

In [Figure 2: see original paper], $w_{m,n}$ represents the n-th word in the m-th document; α is the hyperparameter of the Dirichlet distribution; β_{L_j} is the Dirichlet hyperparameter for topics in language L_j $(j = 1, 2)$; γ is the hyperparameter for the Bernoulli distribution constraining document-topic relationships; $\Lambda_{m,k}$ represents the relationship constraint between document m and topic k, unique for each document-topic pair; $z_{m,n}$ is the topic assigned to word $w_{m,n}$; θ_m is the topic distribution for document m; and ϕ_{k,L_j} is the word distribution for universal topic k in language L_j.

Given document collection D with M documents, each containing representations in both languages, where the L_j portion of document m contains N_{m,L_j} words, and assuming K topics in D, LBTM generates documents as follows:

1. For each universal topic z $(z = 1, 2, ..., K)$: 1.1. For each language L_j $(j = 1, 2)$: 1.1.1. Choose word distribution $\phi_{z,L_j} \sim \text{Dirichlet}(\beta_{L_j})$
2. For each document m in the collection: 2.1. For each universal topic z $(z = 1, 2, ..., K)$: 2.1.1. Choose $\Lambda_{m,k} \in \{0, 1\} \sim \text{Bernoulli}(\gamma)$ 2.2. Choose topic distribution $\theta_m \sim \text{Dirichlet}(\alpha, \Lambda_m)$ 2.3. For each word position n: 2.3.1. Choose topic $z_{m,n} \sim \text{Multinomial}(\theta_m)$ 2.3.2. Choose word $w_{m,n} \sim \text{Multinomial}(\phi_{z_{m,n},L_j})$

2.2 Parameter Estimation

In the parameter estimation phase, we modify Gibbs sampling to accommodate labels and bilingual characteristics. The conditional probability of topic sequences given word sequences is:

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(t)}_{k,-i} + \beta_t}{\sum_{v=1}^{V} n^{(v)}_{k,-i} + \beta_v} \cdot \frac{n^{(k)}_{m,-i} + \alpha_k}{\sum_{j=1}^{K} n^{(j)}_{m,-i} + \alpha_j}$$

For LBTM, this extends to bilingual settings. Let $\mathbf{w}_{L_j}$ denote the vector of all word terms in language L_j, and $\mathbf{z}_{L_j}$ their topic assignments. The conditional probability formula becomes:

$$p(z_{L_j,i} = k|\mathbf{z}_{L_j,-i}, \mathbf{w}_{L_j}) \propto \frac{n^{(t)}_{k,-i} + \beta_{L_j,t}}{\sum_{v=1}^{V_{L_j}} n^{(v)}_{k,-i} + \beta_{L_j,v}} \cdot \frac{n^{(k)}_{m,-i} + \alpha_k}{\sum_{j=1}^{K} n^{(j)}_{m,-i} + \alpha_j}$$

where $n^{(t)}_{k,-i}$ represents the count of word t assigned to topic k in language L_j, excluding the current assignment; V_{L_j} is the vocabulary size of language L_j; and $n^{(k)}_{m,-i}$ represents the count of topic k in document m, excluding the current assignment.

After obtaining topic assignments for each word in each language, the document-topic representation for LBTM is:

$$\theta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{l=1}^{K} n^{(l)}_m + \alpha_l}$$

2.3 Estimating Topic Distribution of New Documents

For a new document, its topic distribution can be predicted using trained model parameters, projecting the document onto the "topic" dimension. We compute the conditional probability $p(z_{L_j,i} = k|\mathbf{z}_{L_j,-i}, \mathbf{w}_{L_j})$, where $\mathbf{w}_d$ is the document's Bag-of-Words vector. The topic assignment for the current word term t depends on other words' current topic assignments and all word topic assignments:

$$p(z_{L_j,i} = k|\mathbf{z}_{L_j,-i}, \mathbf{w}_{L_j}) \propto \frac{n^{(t)}_{k,-i} + \beta_{L_j,t}}{\sum_{v=1}^{V_{L_j}} n^{(v)}_{k,-i} + \beta_{L_j,v}} \cdot \frac{n^{(k)}_{d,-i} + \alpha_k}{\sum_{l=1}^{K} n^{(l)}_{d,-i} + \alpha_l}$$

where $n^{(k)}_{d,-i}$ is the count of topic k assigned to other word positions in document d, excluding the current position. The final topic distribution θ_d for the new document is computed as:

$$\theta_{d,k} = \frac{n_d^{(k)} + \alpha_k}{\sum_{l=1}^{K} n_d^{(l)} + \alpha_l}$$

2.4 Differences from "Latent" Topic Models

Compared to LDA and other "latent" topic models, LBTM utilizes multi-label data linked to documents to instantiate "latent" topics, making topic meanings "explicit" rather than "implicit." Key differences include:

a) **Topic number K**: In LDA, determining K is challenging and requires experimental selection within a range. In LBTM, K is fixed as the number of unique labels after deduplication in the corpus.

b) **Document vector representation**: In LDA, each document' s words can be assigned to any of K topics, so all topic components in the document vector may be non-zero. In LBTM, each document has constraint relationships with its labels, so only topic components corresponding to constrained topics are non-zero.

c) **Document generation process**: LDA selects topic distribution θ_m from a Dirichlet prior α. LBTM uses a Bernoulli prior γ to determine document-topic constraints, as each document either has or lacks constraints with each topic (label).

d) **Sampling range**: LDA computes conditional probabilities between each word and all K topics. LBTM restricts sampling to only those topics (labels) that have constraint relationships with the document.

e) **Computational complexity**: Due to the reduced sampling range, LBTM offers computational advantages. During each iteration, it only computes conditional probabilities for topics associated with the document, unlike LDA which computes probabilities for all topics.

f) **Inference for new documents**: LDA samples from all K topics during inference. LBTM only samples from topics assigned to the current word during training, further reducing computational complexity.

3 Experiments

3.1 Cross-Lingual Text Classification

To validate LBTM' s effectiveness and feasibility for cross-lingual text classification, we trained classifiers on Chinese and Korean scientific literature training sets and applied them to classify test documents in the other language. The corpus is a parallel corpus, so Chinese and Korean documents have identical content (semantic alignment). We compared our approach with the traditional "latent" LDA topic model from [6] that does not use label information.

3.1.1 Dataset The bilingual corpus consists of Chinese-Korean parallel scientific literature from the Science and Technology Department of Yanbian Korean Autonomous Prefecture. It comprises 9,000 papers with sentence-level aligned keywords and abstracts in both languages, as shown in [Figure 3: see original paper]. The collection includes 6,000 ecology papers and 3,000 aerospace papers, categorized by journal type. The training-to-test split ratio is 9:1 for each category. Chinese abstracts were segmented using the ICTCLAS system, and Korean abstracts using the Hannanum system.

3.1.2 Evaluation Metrics We evaluate classification performance using Macro-F1 and Micro-F1 scores. Macro-F1 computes metrics for each class separately and averages them arithmetically, while Micro-F1 aggregates all test instances into a global confusion matrix regardless of class. Definitions are as follows:

$$\text{Macro-F1} = \frac{2 \times \text{Macro-precision} \times \text{Macro-recall}}{\text{Macro-precision} + \text{Macro-recall}}$$

$$\text{Micro-F1} = \frac{2 \times \text{Micro-precision} \times \text{Micro-recall}}{\text{Micro-precision} + \text{Micro-recall}}$$

where precision is the ratio of correctly classified documents to all documents assigned to a category, and recall is the ratio of correctly classified documents to all documents actually belonging to that category. Macro-F1 is sensitive to minority classes, while Micro-F1 is influenced by majority classes [14,15].

3.1.3 Parameter Settings Topic models require preset parameters including topic number K, Dirichlet priors α and β, and training/test iteration counts. Parameter settings for LBTM and "latent" LDA are shown in .

Table 1: Parameter Settings for Comparative Experiments

Parameter	Labeled Bilingual Topic Model	"Latent" LDA Topic Model
Topic number K	18,170 (unique labels after deduplication)	400
Prior α	50/K	50/K
Prior β	0.01	0.01
Training iterations	1,000	1,000
Test iterations	100	100

In LBTM, K is fixed at 18,170—the number of unique labels in the training set. For LDA, K was set to 400 for comparison. Training and test iterations were set to 1,000 and 100 respectively to ensure convergence.

During classification, we first model the training set. In each training iteration, we compute the probability of assigning each word to each topic using Equation (5) and update assignments via Gibbs sampling. After training, we obtain model parameters (document-topic distributions) using Equation (6). For test documents, we initialize topic distributions using trained parameters and known word terms, then iteratively compute assignment probabilities using Equation (7) and update via Gibbs sampling, finally obtaining the test document' s topic distribution using Equation (8).

3.1.4 Experimental Results and Analysis We compared LBTM with the traditional "latent" LDA model [6] on the Chinese-Korean scientific literature corpus. Both training and test sets were represented as topic distributions. We used a Naive Bayes classifier for cross-lingual classification, performing two tasks: (1) training on Korean documents to classify Chinese test documents (KOR->CHN), and (2) training on Chinese documents to classify Korean test documents (CHN->KOR). Training and test sets were randomly split 9:1. We built label indices and dictionaries from training set labels and content, estimated LBTM parameters, inferred topic distributions for test documents, trained the Naive Bayes classifier on training data, and finally classified test documents.

Table 2: Comparative Experimental Results

Model	Micro-F1	Macro-F1	Training Time
Labeled Bilingual Topic Model	94.81%	92.41%	8 hours
"Latent" LDA Topic Model [6]	94.04%	92.76%	48 hours

LBTM achieved a maximum Micro-F1 of 94.81% and Macro-F1 of 92.41%, demonstrating practical applicability for automatic cross-lingual classification of scientific literature. Compared to LDA, LBTM' s Micro-F1 is higher while Macro-F1 is slightly lower. In the 6,000 ecology documents, LBTM outperformed LDA, while in the 3,000 aerospace documents, LDA performed better —confirming that Micro-F1 is influenced by majority classes and Macro-F1 by minority classes. Comprehensive evaluation requires both metrics.

The total time consumption was 8 hours for LBTM versus 48 hours for LDA (a 1:6 ratio). During training, LBTM' s per-word topic sampling range is limited to a document' s keywords (typically 5-6), whereas LDA computes probabilities for all topics. During inference, LBTM only computes probabilities for topics assigned during training, while LDA computes probabilities for all topics. Thus, LBTM achieves higher classification accuracy with significantly reduced computational cost.

3.2 Label Recommendation

By instantiating and clarifying "latent" topics, LBTM can be applied to label recommendation. For new documents without labels, we infer their topic

distributions using the trained model. When representing documents as topic-dimensional vectors, each component value indicates the proportion of words belonging to that topic—higher values indicate stronger relevance.

We used the same document collection as the classification task. After inferring test documents' topic distributions, we selected the top two topics with highest component values as recommended labels and compared them with original keywords. Tables 3 and 4 show sample results.

Table 3: Label Recommendation Results for Chinese Scientific Literature

Paper Title	Original Keywords	Recommended Labels
Subsurface Drip Irrigation Technology: Eco-Economic Sustainability Analysis—A Case Study of Cotton in the Manas River Basin, Xinjiang	Subsurface drip irrigation; Sustainability analysis; Cotton; Numerical simulation; Bossel theory	Sustainability analysis; Bossel theory; Numerical simulation; Fertilization; Soil N2O flux dynamics; Chestnut forest; Mach 4 hydrogen autoignition-assisted ethylene ignition; Direct-connect pulse combustion wind tunnel; Ignition test; Sub-combustion mode; Soil organic carbon; Water-soluble organic carbon

Table 4: Label Recommendation Results for Korean Scientific Literature

Paper Title	Original Keywords	Recommended Labels
LFM 광대역압축센싱 (Blind Compressive Sensing Model for LFM Broadband Radar Signals)	압축센싱 (Compressive sensing); 선형주파수변조신호 (Linear frequency modulated signal); 분수계푸리에변환 (Fractional Fourier transform)	압축센싱 (Compressive sensing); 선형주파수변조신호 (Linear frequency modulated signal); 분수계푸리에변환 (Fractional Fourier transform); 신호분석 (Signal analysis)

The recommended labels show semantic relevance to original keywords. For instance, the Chinese paper about cotton irrigation includes "棉花" (cotton) in

both original keywords and recommendations. The Korean paper on compressive sensing includes "압축센싱" (compressive sensing) in both. Since authors manually add keywords with subjective perspectives, exact alignment is unrealistic, but semantic associations enable auxiliary label recommendation.

4 Conclusion

This paper proposes a Labeled Bilingual Topic Model (LBTM) that leverages multi-label information from scientific literature and news reports. LBTM offers several advantages:

a) Compared to traditional "latent" LDA, LBTM "instantiates" topics with explicit semantics and better interpretability.
b) With constrained sampling ranges, LBTM achieves faster parameter estimation and document inference than LDA.
c) On Chinese-Korean parallel scientific literature, LBTM achieves 92.41% Macro-F1 and 94.81% Micro-F1, making it suitable for practical applications.
d) Its explicit topic meanings enable auxiliary label recommendation.

Future research directions include:

a) Extracting domain-specific feature words for each labeled category to improve cross-lingual classification accuracy.
b) Extending the model to multiple languages given parallel corpora with multi-label annotations across more languages.

References

[1] Bel N, Koster C H A, Villegas M. Cross-lingual text categorization [C]// Proc of the 7th European Conference on Research and Advanced Technology for Digital Libraries. Berlin: Springer, 2003: 126-139.

[2] Olsson J S, Oard D W, Hajic J. Cross-language text classification [C]// Proc of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005: 577-584.

[3] Rigutini L, Maggini M, Liu Bing. An EM based training algorithm for cross-language text categorization [C]// Proc of IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC: IEEE Computer Society, 2005: 529-535.

[4] Wei Bin, Pal C. Cross lingual adaptation: an experiment on sentiment classifications [C]// Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 258-262.

[5] Mimno D, Wallach H M, Naradowsky J, et al. Polylingual topic models [C]// Proc of Conference on Empirical Methods in Natural Language Processing.

Stroudsburg, PA: ACL, 2009: 880-889.

[6] Ni Xiaochuan, Sun Jiantao, Hu Jian, et al. Mining multilingual topics from Wikipedia [C]// Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 1155-1156.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[8] 徐谦, 周俊生, 陈家骏. Dirichlet 过程及其在自然语言处理中的应用 [J]. 中文信息学报, 2009, 23(5): 25-32, 46. (Xu Qian, Zhou Junsheng, Chen Jiajun. Dirichlet process and its applications in natural language processing[J]. Journal of Chinese Information Processing, 2009, 23(5): 25-32, 46.)

[9] 徐戈, 王厚峰. 自然语言处理中主题模型的发展 [J]. 计算机学报, 2011, 34(8): 1423-1436. (Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.)

[10] Beal M J. Variational algorithms for approximate Bayesian inference[D]. London: University of London, 2003.

[11] Fang Anjie, Macdonald C, Ounis I, et al. Exploring time-sensitive variational Bayesian inference LDA for social media data[C]// Proc of the 39th European Conference on Information Retrieval. Berlin: Springer, 2017: 252-265.

[12] 王爱平, 张功营, 刘方. EM 算法研究与应用 [J]. 计算机技术与发展, 2009, 19(9): 108-110. (Wang Aiping, Zhang Gongying, Liu Fang. Research and application of EM algorithm[J]. Computer Technology and Development, 2009, 19(9): 108-110.)

[13] Yerebakan H Z, Dundar M. Partially collapsed parallel Gibbs sampler for Dirichlet process mixture models[J]. Pattern Recognition Letters, 2017, 90: 22-27.

[14] 张启蕊, 张凌, 董守斌, 等. 训练集类别分布对文本分类的影响 [J]. 清华大学学报: 自然科学版, 2005, 45(S1): 1802-1805. (Zhang Qirui, Zhang Ling, Dong Shoubin, et al. Effects of category distribution in a training set on text categorization[J]. Journal of Tsinghua University: Science and Technology, 2005, 45(S1): 1802-1805.)

[15] Luo Le, Li Li. Defining and evaluating classification algorithm for high-dimensional data based on latent topics[J/OL]. PLoS One, 2014, 9(1): e82119. (2014-01-09)[2018-05-05]. https://doi.org/10.1371/journal.pone.0082119.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.