Topic Model Postprint Based on Semantic Distribution Similarity

Authors: Ju Yaya, Yang Lu, Yan Jianfeng

Date: 2018-10-11T00:00:00+00:00

Abstract

Latent Dirichlet Allocation (LDA) is a popular three-layer Bayesian probabilistic model that achieves clustering of documents and words within documents at the topic level. LDA is based on the Bag-of-Words (BOW) model, which simplifies modeling complexity but results in poor semantic coherence of topics and weak document representation capability. To address this issue, a topic model based on semantic distribution similarity is proposed. This model, within the EM (Expectation Maximization) algorithm framework, employs the GPU (generalized Pólya urn) model to incorporate word-word and document-topic semantic distribution similarities to guide topic modeling, thereby weakening the influence of the bag-of-words assumption on topic generation from the semantic association level. Experiments on four public datasets demonstrate that the topic model based on semantic distribution similarity exhibits superior performance in terms of topic semantic coherence and text classification accuracy compared to currently popular topic modeling algorithms, while also improving convergence speed and model precision.

Full Text

Preamble

Semantic Distribution Similarity Based Topic Model

Ju Yaya, Yang Lu, Yan Jianfeng
(School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Latent Dirichlet Allocation (LDA) is a popular three-layer Bayesian probability model that implements clustering of words and texts at the topic level. LDA is based on the Bag-of-Words (BOW) model, which simplifies modeling complexity but results in poor semantic coherence of topics and weak document representation capabilities. To address this problem, this paper proposes a

semantic distribution similarity based topic model. Under the EM (Expectation Maximization) algorithm framework, this model employs the GPU (generalized Pólya urn) model to incorporate word-word and document-topic semantic distribution similarities to guide topic modeling, thereby weakening the impact of the bag-of-words assumption on topic generation from the semantic association level. Experiments on four public datasets demonstrate that the semantic distribution similarity based topic model exhibits superior performance compared to current popular topic modeling algorithms in terms of topic semantic coherence and text classification accuracy, while also improving convergence speed and model precision.

Keywords: latent Dirichlet allocation; semantic distribution similarity; topic model; GPU model

0 Introduction

With the rapid development of Internet technology, network data has exploded, primarily including microblogs, news, web pages, images, and audio, among which textual information occupies a dominant position. How to obtain needed knowledge from massive text information is a major challenge currently facing people, and topic models are effective tools for solving this problem. Topic models are statistical models that use unsupervised machine learning algorithms to extract latent topic information hidden in documents and words. Among them, Latent Dirichlet Allocation (LDA) [**?**] is a commonly used probabilistic topic model that explicitly extracts semantic information from text by using topics as an intermediate layer feature representation between documents and words, and is often used for text classification [**?**, **?**], summarization extraction [**?**], topic detection and tracking [**?**], and other tasks.

Currently, popular inference algorithms for the LDA topic model mainly include Variational Bayesian (VB) [**?**], Gibbs Sampling (GS) [**?**], and Expectation Maximization (EM) [**?**, **?**]. Variants based on these three inference algorithms have been developed for specific application scenarios, such as Stochastic Variational Bayesian [**?**], Author-Topic Model [**?**], and Adaptive Expectation Maximization [**?**]. Although these algorithms can achieve certain modeling effects, they still face a series of challenges. First, current topic model variants typically incorporate external prior knowledge to guide modeling for functional or semantic enhancement. For example, Chen et al. [**?**] proposed the GK-LDA model (General Knowledge LDA), which utilizes domain-independent general knowledge to obtain semantic relationships between words and integrates them into the topic modeling process to improve topic coherence. However, this approach mainly targets improvements for short text tasks in specific domains and lacks universality, while the acquired prior knowledge may contain errors. Second, current topic modeling does not effectively combine relevant semantic enhancement mechanisms. For instance, Bekoulis et al. [**?**] proposed a graph-based weighting method that assumes the more frequently two words co-occur in a document, the larger their corresponding weight, enabling topic models to obtain more dis-

criminative topics from long documents. However, this weighting method does not consider semantic relationships between words during modeling, and therefore cannot obtain topics with optimal semantic coherence and interpretability [**?**]. Additionally, most current topic modeling uses Gibbs Sampling (GS) to estimate parameters, which often prevents the model from converging to an ideal state, resulting in weak semantic representation capabilities of documents.

To address the above problems, this paper constructs a text topic model based on semantic distribution similarity, aiming to enhance topic semantic coherence, improve text classification accuracy, and increase convergence speed and precision. This paper studies the semantic enhancement problem of probabilistic topic models and proposes the Semantic Distribution Similarity based Topic Model (SDS_TM). Under the EM algorithm framework, the generalized Pólya urn (GPU) model [**?**] is used to perform semantic enhancement from both word-word and document-topic perspectives, and to implement parameter estimation for SDS_TM. First, for word-word semantic enhancement, similarity between words is obtained through semantic distribution representation of words. Second, for document-topic semantic enhancement, representative words of document semantics are obtained by calculating the similarity between the semantic distribution representation of documents and the semantic distribution representation of words in documents, and the probability of corresponding topics in the document is increased through their quantitative enhancement. This paper compares the semantic distribution similarity based topic model with currently popular inference algorithms: Variational Bayesian (VB), Gibbs Sampling (GS), and Expectation Maximization (EM). Experiments show that the semantic distribution similarity based topic model exhibits superior performance in topic semantic coherence and text classification accuracy, while effectively improving convergence speed and precision.

1.1 LDA Topic Model

The LDA model is an unsupervised three-layer Bayesian probabilistic graphical model consisting of three layers: document, topic, and word. The LDA graphical model is shown in [Figure 1: see original paper], where non-shaded circles represent parameters or hidden variables to be estimated; shaded circles represent observable variables; arrows indicate dependency relationships between variables; boxes indicate repeated processes; and subscripts in boxes indicate the number of repetitions. The LDA model assumes that the entire text collection has K topics, each document d can be represented as a topic distribution of length K, and each topic k is represented as a word distribution of length equal to the vocabulary size W. The generative process for a document is as follows:

The LDA modeling process is the reverse of the generation process: given a text collection, the model first obtains the distribution for each topic k from a Dirichlet distribution with prior parameter β, and for a document d, obtains its topic probability distribution from a Dirichlet distribution with prior parameter α. Then, for each word t in document d, a topic z_{di} is sampled from θ_d, and the

word is obtained from the topic-word distribution $\phi_{z_{di}}$. This process is repeated until all documents are generated. lists some parameters used in this paper.

The inference goal of LDA is to maximize a specific posterior probability from the joint probability distribution $p(x, z, \theta, \phi|\alpha, \beta)$. Different LDA algorithms have different interpretations of the posterior probability. Currently, mainstream inference algorithms mainly include Variational Bayesian (VB), Gibbs Sampling (GS), and Expectation Maximization (EM). Since these inference algorithms optimize different lower bounds of the posterior probability, their modeling results differ.

1.2 Variational Inference Based Algorithm

When Blei proposed the LDA model, a parameter estimation method based on variational inference (VB) was presented. The core of this algorithm is to use variational inference methods to replace the unsolvable posterior probability distribution with a solvable approximate distribution, and to solve for variational parameters through this approximate distribution. The optimization objective is defined as:

Variational inference utilizes mean field approximation theory, which endows the approximate distribution with a fully factorizable property. This approximate distribution is defined as:

where γ and δ are document-level free parameters. The simplified LDA probabilistic graphical model is shown in [Figure 2: see original paper].

By minimizing the Kullback-Leibler (KL) distance between the approximate distribution and the true distribution to derive parameter values, the update formula for the approximate distribution can be obtained as:

1.3 Gibbs Sampling Based Algorithm

Gibbs Sampling (GS) is a solution for approximate inference problems in LDA models, performing approximate sampling of the joint posterior probability of hidden variables that are difficult to solve. The optimization objective of GS is:

GS is a special case of Markov Chain Monte Carlo (MCMC) [**?**] algorithms, using MCMC to sample from the target distribution. First, the topic label z_{di} of the current word i is removed. Then, based on the topic label distribution of all words except word i, the probability of assigning the current word to each topic is estimated. Finally, a topic label is randomly sampled and assigned to the current word. This process iterates until convergence. The update formula is:

1.4 Expectation Maximization Based Algorithm

The Expectation Maximization algorithm is similar to the LDA posterior probability maximization algorithm (Maximum A Posteriori, MAP) [**?**], with the

optimization objective:

Maximizing this posterior probability can be understood as finding the optimal $\{\theta, \phi\}$ that fits x. Expanding the likelihood probability $p(x|\theta, \phi)$ and maximizing it using Jensen's inequality yields the EM framework for the EM algorithm. The E-step updates the probability that word w in document d belongs to topic k:

The M-step updates the sufficient statistics:

1.5 Analysis and Comparison of Current Algorithms

The above are the three mainstream inference algorithms for LDA. Since these algorithms optimize different combinations of hidden variables and indirectly solve LDA' s $p(x, z, \theta, \phi|\alpha, \beta)$, there are many differences between them. Both Variational Bayesian (VB) and Gibbs Sampling (GS) use approximate inference methods to implement topic modeling. Additionally, the VB algorithm introduces the digamma function when calculating topic distributions, resulting in lower algorithm precision and slower convergence speed. However, Expectation Maximization (EM) uses exact inference to obtain the exact lower bound of the posterior probability when solving parameters $\{\theta, \phi\}$, so this algorithm is superior to VB and GS algorithms in both convergence speed and precision [**?**]. Nevertheless, these three inference algorithms are all based on the bag-of-words (BOW) model assumption, which neither considers the relationship between documents and words in documents nor the relationship between words and words. While this assumption simplifies modeling complexity, it leads to unsatisfactory topic modeling effects.

2 Semantic Distribution Similarity Based Topic Modeling

Currently popular LDA model inference algorithms are all based on the bag-of-words (BOW) model assumption, which represents documents as word frequency vectors. This ignores semantic associations between documents and words, and between words and words, during modeling, losing syntactic and grammatical information of documents. Therefore, many studies have made extensions to topic models, but these extensions mainly target specific tasks or introduce external prior knowledge to guide topic modeling [**?**], representing expansions or improvements to traditional topic model applications without substantial differences.

This paper proposes a semantic distribution similarity based topic model. This model performs semantic enhancement from both word-word and document-topic perspectives under the EM algorithm framework. The main idea is to consider semantic associations between words: words with strong semantic associations with the sampled word have a higher probability of belonging to the same topic. It also considers semantic associations between documents and words in documents: words with close semantic relationships to the document have an increased probability of being selected by the document's corresponding

topic, thereby implementing document-topic semantic enhancement. Through bidirectional semantic enhancement to improve the topic modeling process, the semantic coherence of topics and document representation capabilities are effectively enhanced.

2.1 Semantic Enhancement Based on GPU

The generalized Pólya urn (GPU) model is commonly used in the sampling process of topic models. In contextual topic models, a word is regarded as a ball of a certain color, a topic as an urn, and the distribution of topics is reflected by the number of balls of different colors in the urn. The reason LDA models follow the generalized Pólya urn model is that when a ball of a specific color is taken from the urn, balls of the same color are returned to the urn together with the ball. Over time, the change in the number of balls in the urn is a self-reinforcing phenomenon, i.e., "the rich get richer," which is consistent with the word topic sampling process in topic models. This paper uses the GPU model to perform semantic enhancement for topic modeling from both word-word and document-topic perspectives.

2.1.1 Word-Word Semantic Enhancement Most previous studies mainly obtained semantic relationships between words through external prior knowledge, but such semantic knowledge may not conform to the modeling corpus. Therefore, without introducing external prior knowledge, this paper considers semantic associations between words in the corpus from two perspectives: local contextual syntactic information and global document-level semantic information.

Word-word GPU semantic enhancement is achieved by calculating the cosine similarity between word semantic distribution representations. The local semantic distribution of words is obtained through the word2vec model [**?**]. The word2vec model represents words as a distributed vector form, examining word semantics only from the surrounding document information of the word' s location, ignoring topic information of the word in global documents. A fixed-size sliding window is used to perform contextual statistics for each word in the corpus, obtaining the contextual semantic distribution representation v_w^t of word w, with dimension K.

The global semantic distribution representation of words is the topic-word distribution generated by the LDA model, which produces semantic information within the global document scope during modeling on the corpus. The topic distribution of word w is represented as a K-dimensional vector ϕ_w. Reference [**?**] studied the topic distribution vector of words. ϕ_w is a sparse matrix. When K is sufficiently large, $\sum_{k=1}^{K} \phi_{wk} \to 0$, and due to the influence of the bag-of-words (BOW) model, high-frequency words in documents have lower sparsity, while keywords or low-frequency words have higher sparsity. In traditional topic modeling, high-frequency words almost occupy all topics. Therefore, this paper

introduces the L2 norm in word semantic distribution representation to suppress the influence of high-frequency words on modeling. The L2 norm is used to measure the sparsity of vectors. Equation (15) is the calculation formula for the sparsity of word w' s topic vector, where K represents the number of topics.

Therefore, by linearly weighting and summing the local semantic distribution v_w^t and global semantic distribution ϕ_w of word w, the semantic distribution representation v_w of word w can be obtained:

where the weight λ adjusts the position of words in the vector space, making words under the same topic closer in the vector space. For a sampled word w, words with cosine similarity greater than threshold ρ constitute the similar word set A_w of that word. Assuming the word similarity matrix is A, when word w is sampled, all words in set A_w will have their probability values on the sampling topic increased by the corresponding cosine similarity. The enhancement for the current word w itself remains unchanged at 1, and no enhancement is performed for other cases. The specific enhancement method is shown in Equation (16):

2.1.2 Document-Topic Semantic Enhancement Most previous research on topic model enhancement has only focused on semantic associations between semantically similar words, without considering semantic associations between words and texts. This paper starts from the semantic distribution representation of documents, considering semantic associations between the document-topic distribution generated by modeling and the responsibility values μ_{wdk} of words in the document to obtain representative words of document semantics. Since μ_{wdk} is a sparse matrix, its L2 norm is used to constrain the influence of the bag-of-words (BOW) model on semantic enhancement. The semantic association between a word and its document is reflected in the GPU model enhancement process: when word w is sampled by topic k, if the word has a close semantic association with document d, the probability value of topic k in document d will be enhanced. Document-topic GPU semantic enhancement is achieved by calculating the semantic similarity between the semantic distribution of words in the corpus and the semantic distribution of their documents, as shown in Equation (17), where δ_w represents the sparsity of word w, μ_{wdk} is the probability value of word w in document d belonging to topic k, θ_d is the document-topic distribution, and W_d is the set of all words in document d. If the similarity between them is greater than ρ, it is considered that semantic enhancement is needed between document d and topic k; otherwise, no enhancement is performed. The enhancement matrix is B_{wd}, and the specific enhancement method is shown in Equation (18):

2.2 SDS_TM Model Structure

This paper proposes the Semantic Distribution Similarity based Topic Model (SDS_TM). SDS_TM builds upon the LDA model and uses the GPU model to fuse word-word and document-topic semantic distribution similarities to implement semantic enhancement during topic modeling.

The graphical model of SDS_TM is shown in [Figure 3: see original paper]. The diagonal shaded parts in the figure represent the GPU semantic enhancement for document-topic and word-word components. The former depends on the document-topic distribution and topic-word distribution generated during topic modeling, while the latter depends not only on the topic-word distribution but also on Skip-Gram word embeddings, i.e., the local semantic distribution of words obtained using the Skip-Gram model in word2vec.

2.3 Model Parameter Inference

Currently, mainstream topic model inference algorithms include Variational Bayesian (VB), Gibbs Sampling (GS), and Expectation Maximization (EM) algorithms. Among them, the EM algorithm directly optimizes the exact lower bound of the LDA model' s posterior probability, demonstrating superior performance in generalization and precision compared to VB and GS algorithms. Therefore, this paper infers the parameters of the SDS_TM model based on the EM algorithm framework. According to the update formulas of the EM algorithm, the update formula for word w in document d on topic k is shown in Equation (19), using the GPU model to fuse word-word and document-topic semantic distribution similarities. The update formulas for sufficient statistics are shown in Equations (20) and (21):

Combining the graphical model and update formulas of SDS_TM, its training process is as follows: when the model initially converges (iteration number greater than lower bound), the results obtained from the LDA model are combined with the local semantic distribution of words from word2vec, and the GPU model is used for semantic enhancement during topic modeling. Since the calculation of similarity between words takes a long time, matrix A is updated at intervals after the model initially converges.

3.1 Experimental Environment and Datasets

The experiments in this paper were conducted on a single-machine multi-core server consisting of 2 Intel(R) Xeon(R) CPUs @ 2.10GHz, with 8 cores per CPU (16 cores total) and 140GB of memory.

The experiments were performed on four public datasets: Cora, WebKB, Reuters R8 (R8), and 20 Newsgroups (20 News), which are introduced in reference [**?**]. briefly summarizes these four datasets, where D is the number of documents in the corpus, W is the vocabulary size, NNZ is the number of non-zero elements, and Category is the number of text categories in the dataset. Before the experiments, some preprocessing work was performed on the datasets, mainly including removing standard stop words, removing words that appear fewer than 3 times, and stemming words.

In topic model research and applications, the selection of prior parameters has a certain impact on topic modeling [**?**]. However, parameter research is not the focus of this paper. To ensure fairness and simplicity in comparative experiments,

referencing the parameter settings in [**?**], all algorithms' prior parameters were set to $\alpha = 50/K$ and $\beta = 0.01$, where K is the number of topics. The total number of iterations in the experiments was set to T=1000. This paper sets corresponding similarity thresholds based on semantic distribution similarity, taking the top 2, and the sliding window size of the word2vec model was set to 4.

3.2 Evaluation Metrics

This paper evaluates the modeling capability of topic models using commonly used performance evaluation metrics in the general domain of topic models: Pointwise Mutual Information (PMI) [**?**, **?**], classification accuracy (Accuracy) [**?**], and Perplexity [**?**, **?**, **?**].

PMI is a commonly used evaluation metric for measuring topic semantic coherence. Its main idea is that the top N words with the highest probability values in the topic-word distribution are more likely to appear in the same document in the corpus. The PMI evaluation metric is usually consistent with manual evaluation results, using the correlation between the top N words in a topic as the PMI value. Higher PMI indicates stronger topic semantic coherence. The PMI calculation formula for topic k is:

where $Q(w_i)$ represents the number of documents containing word w_i in the corpus, $Q(w_i, w_j)$ represents the number of documents containing both words w_i and w_j, $\{w_{k1}, ..., w_{kN}\}$ is the list of the N words with the highest probability in topic k, and ϵ is a small positive integer used to avoid logarithm of zero. This paper sets $N = 10$ and $\epsilon = 1$.

Classification accuracy is a commonly used metric for measuring document semantic representation capability. Topics are used as document features to implement text classification. This paper divides datasets into training and testing sets at a 6:4 ratio and uses a Support Vector Machine (SVM) classifier for classification tasks. The average of ten experiments is taken as the accuracy. Without loss of generality, experimental verification shows that classification results with other classifiers are consistent. The classification accuracy calculation formula is:

where $|C|$ represents the number of text categories, D_i represents the number of texts in category c_i, and T_i represents the number of texts correctly classified in category c_i.

Perplexity is a commonly used evaluation metric for assessing the quality of LDA model modeling. It can be understood as the inverse of the geometric mean of the likelihood values of all words in the corpus. Lower perplexity indicates better generalization performance. Its calculation formula is:

3.3.1 Semantic Coherence Analysis

This paper compares the currently popular LDA inference algorithms—Variational Bayesian (VB), Gibbs Sampling (GS), and Expectation Maximization (EM)—with the proposed Semantic Distribution Similarity based Topic Model (SDS_TM). [Figure 4: see original paper] shows the PMI value comparisons of the four algorithms on the Cora, WebKB, R8, and 20 News datasets under different topic numbers K. It can be observed that the PMI values of SDS_TM proposed in this paper are generally higher, indicating that the topics it extracts have higher semantic coherence.

The VB algorithm optimizes an approximate lower bound of the posterior probability, while the GS algorithm obtains word topic labels through simple sampling methods. Both are approximate inference methods. The EM algorithm precisely optimizes the posterior probability representation, so it can obtain topics with stronger semantic relevance than VB and GS algorithms. However, these three inference algorithms are all based on the bag-of-words (BOW) model, ignoring semantic relationships in topic models. In contrast, SDS_TM can effectively integrate semantic associations between word-word and document-topic into topic modeling, thus obtaining topics with higher semantic coherence.

3.3.2 Text Classification Effect Analysis

This paper applies the SDS_TM model to text classification tasks to verify the overall effectiveness of the model. Higher text classification accuracy indicates stronger feature expression capability of topics. shows the classification accuracy of the four algorithms on the R8 and 20 News datasets as the number of topics K changes. It can be seen that the SDS_TM model can achieve higher accuracy on both datasets. The exact inference algorithm EM achieves higher accuracy than the approximate inference algorithms VB and GS. The classification accuracy on the R8 dataset is higher than on the 20 News dataset, possibly due to the influence of document size on topic modeling. R8 has a shorter vocabulary list than 20 News, resulting in less sparsity in texts on the R8 dataset and enabling the acquisition of more similar semantic information, which more effectively guides topic modeling.

3.3.3 Algorithm Convergence and Model Precision Analysis

Convergence is a commonly used metric for evaluating model training speed. [Figure 5: see original paper] and [Figure 6: see original paper] show the perplexity changes with iteration numbers for the four LDA algorithms on the R8 and 20 News datasets. Since the SDS_TM model effectively integrates semantic distribution similarity information during modeling, it has the fastest convergence speed. The VB and EM algorithms retain all topic information for each word, so their convergence speed is faster than the GS algorithm. The GS algorithm only samples one topic for each word, and the sampling process is relatively slow, so its convergence speed is the slowest. Additionally, the

SDS_TM model has advantages in final perplexity compared to the other three algorithms. Lower perplexity indicates higher model precision and stronger generalization capability on unknown datasets. The VB algorithm has the highest perplexity after convergence because it simplifies model complexity, resulting in precision loss. When the iteration number exceeds 30, the topic model tends to roughly converge. By introducing word-word and document-topic semantic distribution similarities to guide topic modeling, the perplexity reduction amplitude of the SDS_TM model increases and quickly approaches the convergence state. Therefore, SDS_TM exhibits superior performance in both convergence speed and model precision compared to other algorithms.

4 Conclusion

This paper addresses the shortcomings of current topic model inference algorithms, such as poor semantic coherence and weak document representation capability, by proposing the Semantic Distribution Similarity based Topic Model (SDS_TM). This model effectively uses the GPU model under the EM algorithm framework to integrate semantic associations between word-word and document-topic into the topic modeling process, thereby inferring topic model parameters. Experiments demonstrate that SDS_TM exhibits excellent performance in topic semantic coherence, text classification accuracy, convergence speed, and model precision. Future research on SDS_TM will mainly focus on improving the speed of calculating semantic distribution similarity, its application on big data streams, and parallel acceleration, aiming to accelerate model training while improving model precision.

References

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

[2] Hoffman M D, Blei D M, Bach F R. Online Learning for Latent Dirichlet Allocation [J]. Advances in Neural Information Processing Systems, 2010, 23: 856-864.

[3] Dunlavy D M, O' Leary D P, Conroy J M, et al. QCS: A system for querying, clustering and summarizing documents [J]. Information Processing & Management, 2007, 43 (6): 1588-1605.

[4] Niebles Juan Carlos, Wang Hongcheng, Li Feifei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words [J]. International Journal of Computer Vision, 2008, 79 (3): 299-318.

[5] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National academy of Sciences, 2004, 101 (Suppl 1): 5228-5235.

[6] Liu Xiaosheng, Zeng Jia, Yang Xi, et al. Scalable Parallel EM Algorithms for Latent Dirichlet Allocation in Multi-Core Systems [C]// International Con-

ference on World Wide Web. New York: ACM Press, 2015: 669-679.

[7] Zhang JianWei, Zeng Jia, Yuan Mingxuan, et al. LDA Revisited: Entropy, Prior and Convergence [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2016: 1763-1772.

[8] Foulds J, Boyles L, Dubois C, et al. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 446-454.

[9] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents [C]// Proceedings of the 20th Conference on Uncertainty in Artifical Intelligence. 2012: 487-494.

[10] Chen Zhiyuan, Mukherjee Arjun, Liu Bing, et al. Discovering coherent topics using general knowledge [C]// Proceedings of the 22th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2013: 209-218.

[11] Bekoulis G, Rousseau F. Graph-Based Term Weighting Scheme for Topic Modeling [C]// International Conference on Data Mining Workshops. New York: IEEE, 2016: 1039-1044.

[12] Mimno D, Wallach H M, Talley E, et al. Optimizing Semantic Coherence in Topic Models [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011: 262-272.

[13] Kotz S, Mahmoud H, Robert P. On generalized Polya urn models [J]. Statistics Probability Letters, 2000, 49 (2): 163-173.

[14] Gilks W R. Markov Chain Monte Carlo [M]. Numerical Analysis for Statisticians. New York: Springer, 1999: 238-245.

[15] Asuncion A, Welling M, Smyth P, et al. On smoothing and inference for topic models [C]// Conference on Uncertainty in Artificial Intelligence. 2009: 25-32.

[16] Andrzejewski David, Zhu Xiaojin, Craven Mark. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors [C]// International Conference on Machine Learning. New York: ACM Press, 2009: 25-32.

[17] Mikolov Tomas, Sutskever Ilya, Chen Kai, et al. Distributed Representations of Words and Phrases and their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[18] Wu Xiaona, Zeng Jia, Yan Jianfeng, et al. Finding Better Topics: Features, Priors and Constraints [C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York: Springer, 2014: 296-310.

[19] Wallach H M, Mimno D M, Mccallum A. Rethinking LDA: Why priors matter [C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009: 1973-1981.

[20] Newman D, Lau J H, Grieser K, et al. Automatic evaluation of topic coherence [C]// Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010: 100-108.

[21] 常东亚, 严建峰, 杨璐. 基于中心词的上下文主题模型 [J]. 计算机应用研究, 2018, 35 (4): 1005-1009. (Chang Dongya, Yan Jianfeng, Yang Lu. Centroid-word based context topic model [J]. Application Research of Computers, 2018, 35 (4): 1005-1009.)

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