Topic Model-Based Entity Alignment for Encyclopedic Knowledge Bases: Postprint

Authors: Zhenpeng Liu, He Mengjie, Zhang Bin, Dong Jing, Xu Jianmin

Date: 2018-08-13T00:00:00+00:00

Abstract

To address the problem that traditional entity alignment methods fail to represent latent semantic information, we optimize them to achieve more significant entity alignment performance. The LDA model is used to model unstructured web encyclopedia data, and an improved BP algorithm is adopted to solve for the hidden parameters in the LDA model, thereby generating entity feature vectors for similarity computation and determining alignment feasibility based on the results. Experimental results show that, compared with three traditional algorithms, the proposed algorithm improves on all three evaluation metrics: accuracy, recall, and the comprehensive F-measure. For web encyclopedia entities with descriptive information, this algorithm can effectively enhance entity alignment performance.

Full Text

Preamble

Entity Alignment for Encyclopedia Knowledge Base Based on Topic Model

Liu Zhenpenga,b, He Mengjiea, Zhang Binb, Dong Jinga, Xu Jianminc

aSchool of Electronic Information Engineering; bInformation Technology Center; cSchool of Cyber Security & Computer, Hebei University, Baoding, Hebei 071002, China

Abstract: Traditional entity alignment methods fail to capture latent semantic information. This paper optimizes these methods to achieve more significant alignment performance. The proposed approach employs the LDA model to model unstructured data from web encyclopedias, uses an improved BP algorithm to solve for hidden parameters in the LDA model, and generates entity

feature vectors for similarity calculation to determine alignment. Experimental results demonstrate that compared with three traditional algorithms, the proposed algorithm improves performance across three evaluation metrics: precision, recall, and F-score. For network encyclopedia entities with descriptive information, this algorithm can effectively enhance entity alignment effectiveness.

Keywords: entity alignment; LDA model; BP algorithm; knowledge fusion

0 Introduction

Entity alignment, also known as entity linking [9], aims to determine whether two entities from different data sources [10] refer to the same real-world object. In recent years, the internet has produced increasingly large-scale knowledge bases, including representative foreign knowledge bases such as FreeBase [1], DBpedia [2], YAGO [3], and Omega [4]; in China, prominent knowledge bases include Baidu Zhixin, Sogou Zhili, and Tsinghua University' s bilingual knowledge base XLore [5]. Knowledge bases play a crucial role in natural language processing and artificial intelligence domains such as knowledge graphs [6], information fusion, and intelligent semantic question answering [7]. In Chinese knowledge base construction, complete and reliable data resources are scarce. To obtain comprehensive knowledge, data from different knowledge bases must be integrated, consolidated, and reused. As an important method for knowledge fusion, entity alignment significantly impacts knowledge base construction and expansion.

Current research on entity alignment methods primarily includes three approaches: Web Ontology Language (OWL) based semantic analysis [11], rule-based analysis, and similarity-based determination. For Chinese web encyclopedias, which lack complete ontology information, alignment through OWL semantics is difficult. Moreover, encyclopedias contain entities from numerous domains, making rule-based alignment impractical as different domains require different rules, limiting generalizability. The most widely used approach is similarity-based determination, which typically involves assigning weights to attribute values [12] and calculating similarity between the same attributes of different entities. In recent years, with the popularity of topic models, methods applying topic models to entity descriptive text for modeling and subsequent similarity-based alignment have emerged. References [13,14] utilized RDFS vocabulary to normalize attributes and combined attribute similarity with topic feature similarity of descriptive text to achieve entity alignment. Reference [15] proposed a semi-supervised co-training approach for entity alignment, incorporating entity names, attributes, descriptive text, and key information such as time and numerical values. Reference [16] introduced an ontology-independent entity alignment method based on attribute semantic features, still relying on entity attribute information. However, such methods

are unsuitable for entities with scarce attribute information. For Chinese web encyclopedias, identical attributes often have inconsistent naming and information across different platforms. For example, the attribute "English name" appears as "Foreign name" in Baidu Baike but as "English name" in Hudong Baike. For the public figure "Zhang Jie" (a singer), the "alias" attribute is "Jie Ge" in Baidu Baike but "Zhang Xiaojie" in Hudong Baike. This phenomenon increases the difficulty of attribute-based entity alignment, as it first requires unifying attribute names. Without guaranteed accuracy in attribute alignment, final results are significantly impacted. Research shows that for Chinese web encyclopedias, improper handling of attribute information can produce adverse effects and increase workload. Therefore, the extensive entity summary and descriptive text information in encyclopedia knowledge bases can be effectively utilized. The challenge addressed in this paper is how to construct effective entity features using only unstructured text for entity alignment.

To effectively leverage entity unstructured text, this paper proposes an entity alignment algorithm for encyclopedia knowledge bases based on topic models. The algorithm uses the LDA model for topic modeling of web encyclopedia entity text information and employs an improved BP algorithm to solve for hidden parameters, thereby completing the entity alignment task. Experiments demonstrate that the proposed method effectively improves entity alignment accuracy and exhibits good generalizability for entities with descriptive text.

The main contributions of this work are: (a) effectively utilizing unstructured data from encyclopedia entities, using the LDA model to extract latent semantic information from text, and proposing a broadly applicable entity alignment algorithm for encyclopedia entities with descriptive information; (b) proposing an improved BP algorithm for estimating model hidden parameters during LDA model inference; and (c) conducting experimental validation using data from Baidu Baike and Chinese Wikipedia, comparing with similar algorithms and analyzing the effectiveness of the proposed approach.

1.1 LDA Model

Latent Dirichlet Allocation (LDA) [17] is a three-layer Bayesian probability model proposed by Blei et al. in 2003, comprising word, topic, and document layers. [Figure 1: see original paper] illustrates the LDA graphical model.

In [Figure 1: see original paper], white circles represent hidden variables, gray circles represent observable variables, and arrows between circles indicate probabilistic relationships between variables. Boxes represent replication, with subscripts indicating the number of replications. α and β denote prior parameters for distributions θ and ϕ, respectively. In this paper and experiments, both α and β are set to 0.1. w represents a word in a document, z represents the topic of a word in a document, K represents the total number of topics, N represents the number of words in a document, and D represents the number of documents.

The LDA graphical model in [Figure 1: see original paper] introduces the model from a document generation perspective, showing the process of selecting word w.

This model simplifies text generation into a probability sampling process, representing documents as a probabilistic mixture of multiple topics (the "document-topic" probability matrix θ), while topics consist of different words (the "topic-word" probability matrix ϕ). To generate a document, topics are first sampled to obtain a word collection under each topic, and multiple words are iteratively extracted to form a complete document.

For the algorithm proposed in this paper, parameter estimation for θ and ϕ appearing in the model is required to conduct entity alignment experiments. Currently, three mainstream parameter estimation methods exist: Variational Bayesian (VB), Gibbs Sampling (GS), and Belief Propagation (BP). While VB and GS have made considerable progress in approximate inference, the BP algorithm demonstrates strong competitiveness in learning speed and accuracy. This paper adopts the classic neural network Belief Propagation (BP) algorithm with optimization.

1.2 Belief Propagation Algorithm

The BP algorithm, proposed by Pearl [18], is a message-passing algorithm for inferring parameters in graphical models and an effective method for solving conditional marginal probabilities. Zeng et al. [19] applied this algorithm in 2011 to solve hidden variables in the LDA model, i.e., to estimate θ and ϕ values, achieving significant progress. [Figure 2: see original paper] shows the LDA factor graph based on the BP algorithm proposed by Zeng Jia.

[Figure 2: see original paper] represents the same model as the LDA graphical model in [Figure 1: see original paper], but while [Figure 1: see original paper] focuses on document generation, [Figure 2: see original paper] emphasizes relationships between topic labels and highlights the mathematical relationships for solving topic labels. In [Figure 2: see original paper], gray boxes represent hidden variables θ and ϕ to be solved, while α and β remain prior parameters for θ and ϕ. Other elements represent topic labels. $z_{d,w}$ connects θ_d and ϕ_w, representing the topic label of word w in document d; $z^{-}_{d,w}$ represents topic labels of other words in document d besides word w, meaning $z^{-}_{d,w}$ connects all word topic labels in the same document d. Meanwhile, $z^{-}_{d,w}$ connects to θ_d and ϕ_w, where $z^{-}_{d,w}$ refers to topic labels of word w in all documents except the current document d, thus $z^{-}_{d,w}$ connects topic labels of word w across all documents. Arrows in [Figure 2: see original paper] indicate information propagation directions, carrying topic information.

When using the BP algorithm for LDA parameter estimation, factors influencing $z_{d,w}$ include all connected topic labels and parameters. The topic update formula is:

$$\mu_k^{z_{d,w}} \propto \left(\alpha_k + \sum_{i \neq w} \mu_k^{z_{d,i}}\right) \times \left(\beta_k + \sum_{j \neq d} \mu_k^{z_{j,w}}\right)$$

where $\mu_k^{z_{d,w}}$ represents the probability distribution of topic k for word w in document d; $\sum_{i \neq w} \mu_k^{z_{d,i}}$ represents the topic probability distribution of all words in document d except word w; $\sum_{j \neq d} \mu_k^{z_{j,w}}$ represents the topic probability distribution of word w in all documents except document d.

Here, x represents observed values. In the formula, $x_{\bar{d},w}$ represents observed values of all words in document d except word w, and $x_{\bar{d},w}$ represents observed values of word w in all documents except document d.

Information updates are locally normalized:

$$\sum_{k=1}^{K} \mu_k^{z_{d,w}} = 1$$

Parameters θ and ϕ are estimated as:

$$\theta_{d,k} = \frac{\alpha_k + \sum_w \mu_k^{z_{d,w}}}{\sum_{k'} (\alpha_{k'} + \sum_w \mu_{k'}^{z_{d,w}})}$$

$$\phi_{k,w} = \frac{\beta_k + \sum_d \mu_k^{z_{d,w}}}{\sum_{w'} (\beta_{w'} + \sum_d \mu_k^{z_{d,w'}})}$$

where $\sum_w \mu_k^{z_{d,w}}$ represents the topic probability distribution of all words in document d, and $\sum_d \mu_k^{z_{d,w}}$ represents the topic probability distribution of word w across all documents.

2.1 Algorithm Overview

The core task of this paper is to calculate the latent semantic similarity between encyclopedia entities with identical entry names and perform entity alignment. The specific algorithm process is illustrated in [Figure 3: see original paper].

As shown in [Figure 3: see original paper], the algorithm comprises four modules. The first module is data acquisition and preprocessing. This paper obtains corpus from Chinese Wikipedia and partial Baidu Baike, including encyclopedia entity entry names and related descriptive information. After data acquisition, word segmentation and stop-word removal are performed. The second module applies LDA modeling to the processed text, followed by parameter estimation using the improved BP algorithm—this core step is detailed in Section 2.2. The third module involves feature generation and similarity calculation. Feature

generation processes the obtained "document-topic" matrix θ to produce entity feature vectors, while similarity calculation uses cosine similarity, detailed in Section 2.3. The final module evaluates similarity calculation results; when the similarity between two entities exceeds threshold ω, they are determined to be alignable; otherwise, the entity to be aligned is saved as a new entity in the entity repository and added to the candidate entity' s sense entry.

2.2 Improved BP Algorithm

Traditional LDA models are based on the bag-of-words model, which ignores word order. While this simplifies the model, it provides opportunities for improvement [20]. The BP algorithm offers advantages of high precision and fast speed when inferring LDA model parameters. However, due to inherent limitations of the LDA model and the fact that this algorithm targets Chinese web encyclopedias—where word meaning is largely context-dependent—this paper proposes an improved BP algorithm. [Figure 4: see original paper] shows the factor graph of the improved BP algorithm.

In the improved BP algorithm, a new component is added: the context of word w (denoted by c). When calculating the topic distribution of a word, the word serves as the center, and several words before and after it are expanded to form a word set window. The topic distribution of each word is calculated based on this short text window. After iteration, the topic distribution of each word converges. In [Figure 4: see original paper], z_{d,w_i} represents the topic label of the i-th word w in document d; z^{-}_{d,w_c} represents topic labels of other words in the context window besides word w; $z^{-}_{d,w}$ represents topic labels of other words in document d outside the context window. Additionally, α, β, θ_d, and ϕ_w represent the same content as in the BP algorithm.

The optimization of the BP algorithm in this work first introduces the concept of context. This addresses the LDA model' s limitation of ignoring word order and the BP algorithm' s limitation of assigning identical semantic information to the same word within a document. After incorporating context, word order within a document is preserved, and considering Chinese language characteristics where understanding a word requires its context, identical words within the same document are assigned different topics, making word semantics more aligned with their actual context. Second, this improvement modifies the $\sum_{j\neq d}\mu_k^{z_{j,w}}$ term to $\sum_{i\neq c}\mu_k^{z_{d,w_i}}$. Since this paper aims to align entities by comparing descriptive information of encyclopedia entities with identical names, incorporating topic information of identical words from other documents would confuse the topic information of the current document. Therefore, this paper adopts $\sum_{i\neq c}\mu_k^{z_{d,w_i}}$, effectively using only information from identical words within the same document, making the document topic more explicit.

From the algorithm' s factor graph, the topic of the i-th word w in document d is determined by two components: (1) the influence of different word topics in the context window, and (2) the influence of identical word topics in the

same document but outside the context window. This yields the topic update formula:

$$\mu_k^{z_{d,w_i}} \propto \left(\alpha_k + \sum_{i \neq c} \mu_k^{z_{d,w_i}}\right) \times \left(\beta_k + \sum_{c \neq w} \mu_k^{z_{d,w_c}^{-}}\right)$$

where $\sum_{c \neq w} \mu_k^{z_{d,w_c}^{-}}$ represents topic information of other words in the context window besides word w, and $\sum_{i \neq c} \mu_k^{z_{d,w_i}}$ represents topic information of word w in document d outside the context window.

The final model parameters are:

$$\theta_{d,k} = \frac{\alpha_k + \sum_w \mu_k^{z_{d,w}}}{\sum_{k'}(\alpha_{k'} + \sum_w \mu_{k'}^{z_{d,w}})}$$

$$\phi_{k,w} = \frac{\beta_k + \sum_d \mu_k^{z_{d,w}}}{\sum_{w'}(\beta_{w'} + \sum_d \mu_k^{z_{d,w'}})}$$

The training process for estimating LDA model hidden parameters using the improved BP algorithm is:

a) Randomly initialize a topic for each word;
b) Iterate through the entire corpus, updating each word's topic using update formula (6);
c) Repeat the above process until convergence;
d) Use formulas (7) and (8) to calculate parameters.

2.3 Feature Vector Generation and Similarity Calculation

When using LDA for topic modeling, text topics are hidden variables, meaning θ and ϕ values are unknown. This paper uses the improved BP algorithm to estimate the model' s unknown parameters.

1) Calculate the "document-topic" probability matrix θ: By performing topic modeling on entity descriptive information and using the improved BP algorithm to estimate hidden variables, the "document-topic" probability matrix θ is obtained:

$$\theta = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nK} \end{bmatrix}$$

where p_{ij} represents the probability of topic j belonging to document i; n represents the number of documents in the document set, and K represents the number of topics generated during LDA modeling.

2) Split matrix θ by rows to generate "document-topic" vectors: The input document set is $D = (d_0, d_1, d_2, \dots, d_n)$, where d_0 represents the text to be aligned, and other texts represent entries with identical names in the entity repository.

$$\begin{aligned}
\theta_{d_0} &= (p_{11}, p_{12}, \dots, p_{1K}) \\
\theta_{d_1} &= (p_{21}, p_{22}, \dots, p_{2K}) \\
\theta_{d_2} &= (p_{31}, p_{32}, \dots, p_{3K}) \\
&\vdots \\
\theta_{d_n} &= (p_{n1}, p_{n2}, \dots, p_{nK})
\end{aligned}$$

3) Similarity calculation: Calculate cosine similarity between θ_{d_0} and other "document-topic" vectors to determine document similarity between two articles with identical names. For example, the similarity between entity e_a represented by θ_{d_0} and another entity e_b is:

$$sim(e_a, e_b) = \frac{\theta_{d_0} \cdot \theta_{d_i}}{|\theta_{d_0}| \times |\theta_{d_i}|}$$

where θ_{d_i} represents the "document-topic" vector of entity e_b.

3.1 Experimental Data

To verify the effectiveness of the proposed algorithm, this paper conducts experiments using high-quality Chinese corpora from Chinese Wikipedia and Baidu Baike. Wikipedia regularly updates and releases its corpus packages; this paper downloads the latest Wikipedia corpus for experiments, including entry names and corresponding descriptive information. Since Wikipedia corpus is comprehensive and contains extensive information, it serves as the entity repository in this experiment. Baidu Baike corpus is crawled from the website, with 200 entries each from person, society, science, and art categories (800 total Baidu Baike entries), including entry names and corresponding descriptive information as entities to be aligned.

After obtaining experimental data, preprocessing is performed using Python. The Jieba segmentation tool is used for word segmentation, and the "Harbin Institute of Technology Stopword List" is used for stop-word removal. Experimental data statistics are shown in .

Table 1 Statistics of Entity Alignment Data

Baidu Baike Entities	Wikipedia Homonymous Entities	Alignable Pairs
800	[Value]	[Value]

Table 1 summarizes the data volume used in experiments. As shown, 800 Baidu Baike entities were crawled across four categories (person, society, science, art). Entries were extracted by name to obtain homonymous entities from Wikipedia, along with their counts. Through manual comparison, the number of alignable pairs was also determined.

3.3 Parameter Setting

The main parameters in this paper are threefold: (1) topic number K for the LDA model; (2) prior parameters α and β for the improved BP algorithm; and (3) threshold ω for the proposed LDA-based entity alignment algorithm. The threshold directly relates to alignment results and is therefore discussed in detail.

1) **Impact of Topic Number K on Experimental Results**: To avoid threshold influence, threshold ω is set to 0.9 in this experiment. Results are shown in [Figure 5: see original paper].

[Figure 5: see original paper] shows that for person, society, and art categories, entity alignment precision is not high, likely because descriptive information for these entities is insufficiently precise and the entity repository contains many homonymous entities, increasing the number of entities participating in alignment (N_o). Science category entities, with more rigorous and clear descriptions and fewer proper nouns, have higher precision and fewer homonymous entities (N_o), resulting in higher accuracy under the same topic number.

Results indicate that when topic number K is 8 or 9, entity alignment precision (P), recall (R), and F-score (F) are optimal.

2) **Impact of Threshold ω on Experimental Results**: From previous experiments, topic numbers 8 or 9 yield optimal metrics. In subsequent experiments, topic number K is set to 9. Results are shown in [Figure 6: see original paper].

[Figure 6: see original paper] shows that as threshold increases, precision for all four categories continuously increases while recall continuously decreases. This occurs because increasing threshold ω reduces the number of correctly aligned entities (N_r) after algorithm processing. The F-score initially increases with threshold, reaching its maximum at approximately $\omega = 0.96$. Therefore, when threshold is set to 0.96, the algorithm achieves optimal alignment performance.

3.4 Comparison with Other Algorithms

To demonstrate the effectiveness of the proposed algorithm, comparative experiments are conducted using identical text data and experimental parameters. Three comparison algorithms are: (1) TF-IDF replacing topic modeling in the proposed framework, (2) BP algorithm for estimating LDA model hidden parameters, and (3) Gibbs algorithm for inferring LDA model hidden parameters. Results are shown in .

Table 2 Comparison Results with Other Algorithms

Algorithm	Precision	Recall	F-score
TF-IDF	[Value]	[Value]	[Value]
LDA+BP	[Value]	[Value]	[Value]
LDA+Gibbs	[Value]	[Value]	[Value]
Proposed	[Value]	[Value]	[Value]

Table 2 shows that with identical text data and experimental parameters, different algorithms produce varying alignment effects with significant differences. While the proposed algorithm' s results differ from expected outcomes, its precision is indeed higher than LDA+BP, proving the effectiveness of the BP algorithm improvements. TF-IDF accuracy is slightly lower than the proposed algorithm, likely because TF-IDF only considers term frequency information without capturing document latent semantics. LDA+Gibbs [23] shows metrics roughly equivalent to the proposed algorithm, providing a new research direction for further optimization. Overall, the proposed algorithm demonstrates substantial performance improvements over baseline algorithms and shows good effectiveness for encyclopedia knowledge base entity alignment.

4 Conclusion

In recent years, internet growth has led to massive concentration of online knowledge information. As carriers of knowledge, knowledge bases play an important role in learning. However, single knowledge bases have low knowledge coverage, necessitating knowledge fusion to integrate different knowledge bases. The proposed LDA-based encyclopedia knowledge base entity alignment algorithm effectively addresses this problem and can be practically applied to encyclopedia knowledge base entity alignment tasks.

Future work will consider more effective methods for LDA model parameter estimation, such as Gibbs sampling, and explore additional topic models to improve text similarity, thereby further enhancing knowledge base entity alignment effectiveness.

References

[1] Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge [C]// Proc of AAAI Conference on Artificial Intelligence. British Columbia Canada: IEEE Press. 2007: 1962-1963.

[2] Lehmann J. DBpedia: A large-scale, multilingual knowledge base extracted from wikipedia [J]. Semantic Web, 2015, 6 (2): 167-195.

[3] Suchanek F M, Kasneci G, Weikum G. Yago: a large ontology from Wikipedia and WordNet [J]. Web Semantics Science Services & Agents on the World Wide Web, 2008, 6 (3): 203-217.

[4] Philpot A, Hovy E, Patrick P. The omega ontology [C]// Proc of Ontolex Workshop at LJCNLP. USA: Prep Press. 2005: 59-66.

[5] Li Mingyang, Shi Yao, Wang Zhigang, et al. Building a Large-Scale Cross-Lingual Knowledge Base from Heterogeneous Online Wikis [C]// Proc of CCF Conference on Natural Language Processing and Chinese Computing. New York: Springer-Verlag. 2015: 413-420.

[6] Xu Zenglin, Sheng Yongpan, He Lirong, et al. Review on Knowledge Graph Techniques [J]. Journal of University of Electronic Science and Technology of China, 2016, 45 (4): 589-606.

[7] Liu Kang, Zhang Yuanzhe, Ji Guoliang, et al. Representation Learning for Question Answering over Knowledge Base: An Overview [J]. ACTA Automatica Sinica, 2016, 42 (6): 807-818.

[8] Wang Xuepeng, Liu Kang, He Shizhu, et al. Multi-Source Knowledge Base Entity Alignment by Leveraging Semantic Tags [J]. Chinese Journal of Computers, 2017, 40 (3): 701-711.

[9] Gao Yanhong, Li Aiping, Duan Liguo. Entity disambiguation method based on multi-feature fusion graph model for entity linking [J]. Application Research of Computers, 2017, 34 (10): 2909-2914.

[10] Liu Shulin, Liu Kang, He Shizhu, et al. A probabilistic soft logic based approach to exploiting latent and global information in event classification [C]// Proc of the 30th AAAI Conference on Artificial Intelligence. Phoenix USA: AAAI Press. 2016: 2993-2999.

[11] Stoilos G, Venetis T, Stamou G. A fuzzy extension to the OWL 2 RL ontology language [J]. Computer Journal, 2015, 58 (11): 2956-2971.

[12] Zhang Xiaohui, Jiang Haihua, Di Ruihua. Property Weight Based Co-reference Resolution for Linked Data [J]. Computer Science, 2013, 40 (2): 40-43.

[13] Huang Junfu, Li Tianrui, Jia Zhen, et al. Entity alignment of Chinese heterogeneous encyclopedia knowledge base [J]. Journal of Computer Applications, 2016, 36 (7): 1881-1886.

[14] Yang Xiuzhang. Research and Implementation on Entity Alignment and Attribute Alignment [D]. Beijing: Beijing Institute of Technology, 2016.

[15] Zhang Weili, Huang Tinglei, Liang Xiao. Instance alignment algorithm between encyclopedia based on semi-supervised co-training [J]. Computer and Modernization, 2017 (12): 88-93.

[16] Wan Jing, Li Lin, Yan Huanchun, et al. An entity alignment approach based on the VS-Adaboost algorithm [J]. Journal of Beijing University of Chemical Technology: Natural Science Edition, 2018, 45 (1): 72-77.

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

[18] Andersen S K. Probabilistic reasoning in intelligent systems: networks of plausible inference [J]. Artificial Intelligence, 1991, 48 (1): 117-124.

[19] Zeng Jia, Cheung William K, Liu Jiming. Learning topic models by belief propagation. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (5): 1121-1134.

[20] Chang Dongya, Yan Jianfeng, Yang Lu. Centroid-word based context topic model [J]. Application Research of Computers, 2018, 35 (4): 1005-1009.

[21] Zhuang Yan, Li Guoliang, Feng Jianhua. A survey on entity alignment of knowledge base [J]. Journal of Computer Research and Development, 2016, 53 (1): 165-192.

[22] Wallach H M, Mimno D M, Mccallum A. Rethinking LDA: why priors matter [J]. Advances in Neural Information Processing Systems, 2009, 23: 1973-1981.

[23] Zhang Jianwei. Comparison and Improvement Studies of Topic Model LDA Inference Algorithms [D]. Suzhou: Soochow University, 2017.

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