Soft Computing (2020) 24:11407–11423
https://doi.org/10.1007/s00500-019-04604-0

METHODOLOGIES AND APPLICATION

A novel topic model for documents by incorporating semantic relations between words

Jihong Chen[1] · Kai Zhang[1] · Yuan Zhou[2] · Zheng Chen[1] · Yufei Liu[2,3] · Zhuo Tang[4] · Li Yin[1]

Published online: 23 December 2019

Abstract
Topic models have been widely used to infer latent topics in text documents. However, the unsupervised topic models often result in incoherent topics, which always confused users in applications. Incorporating prior domain knowledge into topic models is an effective strategy to extract coherent and meaningful topics. In this paper, we go one step further to explore how different forms of prior semantic relations of words can be encoded into models to improve the performance of topic modeling process. We develop a novel topic model—called Mixed Word Correlation Knowledge-based Latent Dirichlet Allocation—to infer latent topics from text corpus. Specifically, the proposed model mines two forms of lexical semantic knowledge based on recent progress in word embedding, which can represent semantic information of words in a continuous vector space. To incorporate generated prior knowledge, a Mixed Markov Random Field is constructed over the latent topic layer to regularize the topic assignment of each word during the topic sampling process. Experimental results on two public benchmark datasets illustrate the superior performance of the proposed approach over several state-of-the-art baseline models.

Keywords Topic model · Must-links · Cannot-links · Word embeddings · Gibbs sampling

1 Introduction

Topic models, such as Latent Dirichlet Allocation (LDA), and its extensions provide a powerful method for inferring latent topics in document corpus (Blei et al. 2003; Gao et al. 2017). The standard unsupervised topic models extract the hidden structures from corpus based on the basic assumptions that each topic is represented as a multinomial distribution over a given vocabulary and each document is represented as a multinomial distribution over these topics (Li et al. 2018). However, many researchers have found that the unsupervised models often result in incoherent topics, which constitutes the primary obstacle to acceptance of these models in applications (Chang et al. 2009; Mimno et al. 2011).

The key reason to generate incoherent topics for the above standard topic models lies in that the objective functions of these models do not always correlate well with human judgments (Chang et al. 2009). Specifically, the above unsupervised models assume the topic-word distribution follows a Dirichlet prior, which will result in words under each document to be uncorrelated and generated independently (Ahmed et al. 2017). As a result, the topic modeling process will ignore lexical semantic information between words to learn meaningful and coherent topics, which are accordance with human cognition (Blei and Lafferty 2005). Knowledge-based topic models have been demonstrated to be an effective strategy to deal with the above problem (Petterson et al. 2010).

Several knowledge-based topic models have been proposed by researchers in recent years (Fu et al. 2018; Xu

Communicated by V. Loia.

✉ Yuan Zhou
zhou_yuan@mail.tsinghua.edu.cn

[1] National Numerical Control Systems Engineering Research Center, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2] School of Public Policy and Management, Tsinghua University, Beijing 100084, China

[3] The CAE Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

[4] College of Information Science and Engineering, Hunan University, and National Supercomputing Center in Changsha, Hunan 410082, China

et al. 2018). Early knowledge-based model asks users to provide prior domain knowledge to extract more coherent topics (Andrzejewski et al. 2009). In these models, users will be requested to provide two kinds of prior lexical semantic knowledge in the form of must-links and cannot-links (Jagarlamudi et al. 2012). A must-link means that two words should belong to the same topic and a cannot-link means that two words should not be in the same topic. Compared to the popular unsupervised topic models, combining the above two forms of word–word correlation knowledge generated by users can improve the performance of modeling to some extent (Hu et al. 2014). However, the key weakness exists in the above early knowledge-based models is to require users to provide prior domain knowledge. The quantity and quality of prior knowledge generated artificially will be limited as a user may have no idea what knowledge to provide and the knowledge generation process will be non-automatic or inefficient. In addition, early knowledge-based models lack the judgment mechanism of knowledge applicability and aim to incorporate word–word correlation in a hard and topic-independent way, which ignores the fact that whether two words are correlated depends on which topic they appear in (Xie et al. 2015).

To solve the problem of knowledge acquisition and knowledge judgment in early studies, researchers explore to mine knowledge of word–word correlation automatically and measure incorrect knowledge in modeling process based on statistical information of word co-occurrence in corpus (Chen and Liu 2014a, b). These models solve the problems in early knowledge-based topic models and improve the coherence of topic modeling results further (Shams and Baraani-Dastjerdi 2017). However, the limitations are obvious because these models need a large amount of related domain datasets to mine valuable prior knowledge, which is not applicable in practical applications. Given the fact that semantically related words may not co-occur frequently in text corpus, the knowledge mining method based on word co-occurrence information will generate limited knowledge, which will further increase dependence on massive amount of datasets. In addition, the evaluation mechanism of knowledge correctness in these models depends only on co-occurrence information of words and cannot capture the lexical semantic relations explicitly.

With the development of neural language models, word embeddings make it feasible and practical to represent words in a semantic vector space with the purpose of fully retaining the semantical and syntactical relations between words (Mikolov et al. 2013a). The characteristics of word embeddings provide an effective method to mine lexical semantic knowledge by simply calculating the similarity between words in continuous embedding space. Recently, more attention has been paid on how to incorporate word embeddings into topic models to produce more meaningful topics (Qiang et al. 2017; Xie et al. 2015). However, these existing models mainly explore to incorporate must-links based on word embeddings, but ignore the incorporation of cannot-links in topic modeling (Chen and Liu 2014b; Xie et al. 2015; Yao et al. 2017). The word–word correlation of cannot-link is completely different from must-links and imposes a completely different effect on topic assignment when topic modeling (Yang et al. 2015c). Incorporating and coordinating two different forms of word–word correlation derived from word embeddings in topic models simultaneously are critical to extract more coherent topics, but little attention has been paid on this direction.

In this paper, we propose a novel topic model—called Mixed Word Correlation Knowledge-based Latent Dirichlet Allocation (MWCK-LDA)—to combine both must-links and cannot-links based on word embeddings. The developed MWCK-LDA can not only incorporate both must-links and cannot-links in a topic-dependent way but also balance the effect of two different knowledge forms on topic sampling effectively. Compared to standard topic models that produce knowledge of word–word correlation based on users' experience or word co-occurrence information in corpus, our model can mine more abundant knowledge automatically based on word embeddings, which is superior in both effectiveness and efficiency for knowledge generation. To incorporate and balance two completely different forms of lexical semantic knowledge, gaining insights from (Xie et al. 2015; Zhu and Xing 2010), a Mixed Markov Random Field is constructed over the latent topic layer to regularize the topic assignment of each word during topic modeling process, which will give must-links a better chance to be put into the same topic and cannot-links a better chance to be not. Compared to those models that only take must-links based on word vectors into consideration, our model can also incorporate more abundant cannot-links in modeling process simultaneously, which will combine more comprehensive lexical semantic information and further improve topics quality. In addition to the improved knowledge incorporation capabilities, our model also provides a soft mechanism to determine the applicability of must-links and cannot-links during topic modeling process, which is different from the hard and topic-independent judgment mechanism of knowledge applicability in standard knowledge-based models. Specifically, the mechanism in our model does not specify which topic a must-link or cannot-link should belong to or should not belong to directly, but leaves this to be assessed by text documents automatically. What is more, our model can balance the effect between two forms of prior knowledge and word co-occurrence information in corpus on

topic assignment over each word in document. The main contributions of this paper include:

1. Proposing a novel knowledge-based topic model coordinating and balancing two different lexical semantic knowledge of must-links and cannot-links derived from word embeddings in a soft and balanced manner.
2. Providing a collapsed Gibbs sampling method to infer posterior distribution and estimate parameters of the proposed model, which could incorporate both must-links and cannot-links adaptively.
3. Extensive experiments demonstrate that the proposed model outperforms the state-of-the-art baseline models on two public benchmark datasets in terms of topic coherence in both qualitative and quantitative metrics.

The remainder of this paper is organized as follows. Section 2 shows related researches. Section 3 presents our proposed topic model. Section 4 discusses the experiments, and finally, Sect. 5 concludes.

2 Related work

The most related research to our work is the knowledge-based topic models aiming at incorporating external domain knowledge into modeling process to improve topic quality. In this section, we focus on existing works in knowledge-based models and give a brief summarization.

Several knowledge-based topic models have been proposed in recent years. For example, Andrzejewski et al. (2009) proposed a topic model using a Dirichlet forest to replace the Dirichlet as the prior over the topic-word multinomial distribution, which enables the developed model to encode the set of must-links and cannot-links. Chen et al. (2013b) proposed a topic model that could exploit the knowledge form of s-sets derived from multiple past domains to extract more coherent and meaningful topics in a new domain. The form of s-set is an extended knowledge form of must-links, which is composed of multiple words that should belong to the same topic. In practice, the knowledge applied by the above models is produced by users. However, asking users to provide prior domain knowledge to guide generative process of topic models can be problematic as users may not know what to provide (Lee et al. 2017). It also makes the topic modeling process non-automatic. In addition, these models incorporate the external domain knowledge into modeling process in a primitive way and lack the judgment mechanism of knowledge applicability, which limits the performance of topic modeling.

In order to deal with the above problems, based on word co-occurrence information in corpus some other researchers explore to mine prior knowledge automatically and eliminate incorrect knowledge. Chen et al. proposed LTM to extract must-links by frequent itemset mining (FIM) from prior topics which already inferred in past domains and use pointwise mutual information (PMI) to assess the applicability of mined knowledge in current topic modeling process automatically (Chen and Liu 2014b). As an extension of LTM, Chen et al. developed a new model which expanded the form of prior knowledge, incorporating both must-links and cannot-links into generative process using generalized PÓlya urn (GPU) model (Chen and Liu 2014a). The generation and correctness measurement of knowledge applied in these models are all based on statistical word co-occurrence information in documents, which makes these models have a superior performance than early knowledge-based topic models (Xu et al. 2018). However, the external knowledge incorporation strategy realized by GPU is essentially a hard and topic-independent way, which simply assumes correlated words should be similar in each topic probabilistically. Although researchers introduce word co-occurrence information to assess the correctness of knowledge to current modeling process, the evaluation mechanism of knowledge cannot capture the semantic relations between words effectively given that semantically related words may not co-occur frequently in document corpus (Yao et al. 2016). In addition, the statistical knowledge mining method will generate limited knowledge, making these models need a large amount of past datasets to provide abundant information for knowledge generation, which is always not applicable in majority topic modeling tasks. Yang et al. (2015c) presented Sparse Constrained LDA (SC-LDA) to incorporate prior knowledge of word–word correlation into LDA. In SC-LDA, must-links stem from synsets in WordNet 3.0, which restricts the scope of knowledge generation. For example, “deep” and “learning” is not synonymous, but should constitute a pair of must-link when topic modeling. In addition, SC-LDA is unable to balance the effects of must-links and cannot-links during the generative process of topics. SC-LDA also ignores the fact that whether two words are correlated depends on which topic they appear in and lacks the judgment mechanism of knowledge applicability.

The development of word embeddings provides an effective strategy to represent words in continuous vector space, which can retain abundant syntactic and semantic information between words (Yang et al. 2015a). The continuous vector representations of words make it feasible to measure the lexical semantic similarity by the distance between words in the vector space (Mikolov et al. 2013a, b; Pennington et al. 2014). Recently, more attention has been paid on how to combine word embeddings into topic models to produce high-quality topics (Fang et al. 2016).

Xie et al. (2015) proposed MRF-LDA to use must-links generated by word embeddings to improve the performance of topic modeling. Yao et al. (2017) developed WE-LDA to incorporate must-links by word vectors. However, existing works exploit only one knowledge form of must-links derived from word embeddings in models and ignore more extensive lexical semantic knowledge form of cannot-links. Both forms of the above word–word correlation are all essential to understand document contents. Given that cannot-link is a completely different knowledge form from must-link and plays a completely different role when topic sampling, it is critical to explore the strategy or mechanism to coordinate different kinds of knowledge effectively in topic models. In addition, some other researchers also explored to apply word embeddings in topic modeling tasks for short texts to alleviate the content sparsity problem of short texts (Xun et al. 2016; Yang et al. 2015b; Yao et al. 2016). These short text topic models only explored to use must-link knowledge. However, in this paper we focus on improving topic modeling performance for long documents.

3 The proposed MWCK-LDA model

The proposed model in this paper is based on LDA. In this section, we start with the introduction of LDA and then describe MWCK-LDA in detail along with its inference method. The notations and their corresponding meanings used in this paper are summarized in Table 1. To infer posterior probability and estimate parameters of our model, we derive the Gibbs sampler and give Gibbs sampling algorithm in Table 2.

3.1 Brief review of LDA

In this section, we will briefly discuss the LDA proposed by Blei et al. LDA is a generative probabilistic model that aims at inferring latent topics from documents corpus. LDA follows two basic assumptions: (a) A document is assumed to be a multinomial distribution of topics and (b) a topic is assumed to be a multinomial distribution of words in the given vocabulary. The graph model of LDA is shown in Fig. 1a.

Given a documents corpus containing M documents, the vocabulary derived from this corpus consists of V different words. We assume the number of latent topics contained in text documents is K. LDA is a probabilistic generative model. When generating the mth ($m \in [1, M]$) document in corpus, LDA samples a document-topic multinomial distribution $\vec{\theta}_m$ from a prior Dirichlet distribution with hyper-parameters $\vec{\alpha}$: $p\left(\vec{\theta}_m|\vec{\alpha}\right) = Dir(\vec{\theta}_m|\vec{\alpha})$. $\vec{\alpha}$ and $\vec{\theta}_m$ are both K-dimensional vectors, and the elements of $\vec{\theta}_m$ are satisfied with: $\sum_k \vec{\theta}_{m,k} = 1$, $k = 1, \ldots, K$. Then LDA assigns a latent topic $z_{m,n}$ for each word $w_{m,n}$ in document m based on the topic multinomial distribution $\vec{\theta}_m$ of the mth document. n states the position of the word in document and satisfies collection $n \in [1, N_m]$ where N_m is the number of words in document m. LDA assumes all words is independent from each other in a document. For document m, the joint probability of topic assignments for all words is shown in formula (1) where $\vec{z}_m$ characterizes topic assignments for all words in document.

$$p\left(\vec{z}_m|\vec{\theta}_m\right) = \prod_{n=1}^{N_m} p\left(z_{m,n}|\vec{\theta}_m\right) \tag{1}$$

As discussed above, the kth topic is assumed to be a multinomial distribution over V words in the vocabulary. According to LDA, each topic-word multinomial distribution $\vec{\varphi}_k$ follows a prior Dirichlet distribution with hyper-parameters $\vec{\beta}$: $p\left(\emptyset|\vec{\beta}\right) = Dir(\vec{\varphi}_k|\vec{\beta})$ where both $\vec{\varphi}_k$ and $\vec{\beta}$ are V-dimensional vectors. $\emptyset$ denotes a matrix with $K \times V$ dimension containing all topics' multinomial distributions over words. $\vec{\varphi}_{k,w}$ denotes the probability of generating word w given topic k and it satisfies $\sum_w \vec{\varphi}_{k,w} = 1$, where $w = 1, \ldots, V$. Then each word in document m can be generated by sampling from the assigned multinomial distribution of latent topic: $p\left(w_{m,n}|z = z_{m,n}\right) = \vec{\varphi}_{z_{m,n}}$. The generative process of LDA is listed as follows:

1. For kth topic, where $k = 1, 2, \ldots, K$
   Draw $\vec{\varphi}_k \sim Dir\left(\vec{\beta}\right)$
2. For mth document, where $m = 1, 2, \ldots, M$
   (a) Draw $\vec{\theta}_m \sim Dir(\vec{\alpha})$
   (b) For nth word in mth document, where $n = 1, 2, \ldots, N_m$
   Draw $z_{m,n} \sim Multi\left(\vec{\theta}_m\right)$
   Draw $w_{m,n} \sim Multi\left(\vec{\varphi}_{z_{m,n}}\right)$

Given a documents corpus, $w_{m,n}$ is observable variable. α and β are prior hyper-parameters. $z_{m,n}$, $\vec{\theta}_m$ and $\vec{\varphi}_{z_{m,n}}$ are hidden variables which can be estimated by the observed words in corpus. The joint distribution of all variables is as follows:

Table 1 Notations used in this paper

Notations	Meaning
M	Number of documents in the corpus
$\vec{d}$	Documents in the corpus
V	Number of words in the vocabulary
K	Number of predefined latent topics
N_m	Number of words in document m
m, k, n	Index for document, topic, word in a document
$\vec{\theta}_m$	Multinomial distribution over topics for document m
$\vec{z}_m$	Topic labels of document m
$\vec{\varphi}_k$	Multinomial distribution over words for topic k
$n_m^{(k)}$	Number of tokens assigned to topic k in document m
$\neg i$	Word w_i is excluded from the counting
$n_k^{(w_i)}$	Number of times of word w_i assigned to topic k
I	Number of iterations
n_m	Number of topics in document m
n_k	Number of words in topic k
α, β	Parameter of Dirichlet distribution for $\vec{\theta}_m$ and $\vec{\varphi}_k$
G_{P_m}	The undirected graph corresponding to must-link in document m
G_{N_m}	The undirected graph corresponding to cannot-link in document m
P_m	The set of undirected edges in G_{P_m}
N_m	The set of undirected edges in G_{N_m}
$\|P_m\|$	The number of edges in P_m
$\|N_m\|$	The number of edges in N_m
λ_1	User-defined hyper-parameter balancing the effect of must-links
λ_2	User-defined hyper-parameter balancing the effect of cannot-links

$$p\left(\vec{w}_m, \vec{z}_m, \vec{\theta}_m, \emptyset | \vec{\alpha}, \vec{\beta}\right) = \prod_{n=1}^{N_m} p\left(w_{m,n} | \vec{\varphi}_{z_{m,n}}\right) \cdot p\left(z_{m,n} | \vec{\theta}_m\right) \cdot p\left(\vec{\theta}_m | \vec{\alpha}\right) \cdot p\left(\emptyset | \vec{\beta}\right) \tag{2}$$

The hidden variables in the generative process can be approximated by applying the inference strategy of Markov chain Monte Carlo (MCMC). Specifically, it is implemented using the Gibbs sampling method. The parameter inference method conducted by Gibbs sampling in LDA was first proposed by Griffiths et al. and has been widely used in various parameters estimating tasks in the field of topic models (Griffiths and Steyvers 2004). In order to simplify the inference, we always assume the prior Dirichlet distributions in topic model are symmetrical Dirichlet distributions. Based on the above assumptions, we can derive the conditional distribution to sample a topic z for each word in corpus as follows:

$$\mathrm{p}(z_i = k | \vec{z}_{\neg i}, \vec{w}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)} + \alpha\right)} \cdot \frac{n_{k,\neg i}^{(w_i)} + \beta}{\sum_{i=1}^{V}\left(n_{k,\neg i}^{(w_i)} + \beta\right)} \tag{3}$$

where $n_{m,\neg i}^{(k)}$ is the number of topic k assigned to document m and $n_{k,\neg i}^{(w_i)}$ is the number of word w_i assigned to topic k. Symbol $\neg i$ denotes that the ith word is excluded from the counting. α and β are hyper-parameters of two symmetric prior Dirichlet distributions, respectively.

3.2 MWCK-LDA

In this section, we will discuss how to incorporate lexical semantic correlation into learning process of topic model and present our proposed model, MWCK-LDA. As discussed above, standard LDA model assumes that topic-word multinomial distribution follows a prior Dirichlet distribution. Under this assumption, the words in the corpus are regarded as a bag-of-words model where any word is independent from each other, which ignores the effects of semantic correlation between words on topic assignment

Table 2 Algorithm of Gibbs sampling for MWCK-LDA

Algorithm 1: Gibbs sampling for MWCK-LDA

Input: Topic number $\boldsymbol{K}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{I}$, $\boldsymbol{\lambda_1}$, $\boldsymbol{\lambda_2}$, $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, Documents in the corpus $\vec{\boldsymbol{d}}$
Result: topic assignments of all words in the corpus
begin
  initialize $\boldsymbol{n_m^{(k)}}$, $\boldsymbol{n_k^{(w_i)}}$, $\boldsymbol{n_k}$ as zero for each topic k
  for each document $\boldsymbol{m} \in [1,M]$ do
    for each word $\boldsymbol{n} \in [\boldsymbol{1}, \boldsymbol{N_m}]$ do
      mine must-links set $\boldsymbol{ML_{m,n}}$ and cannot-links set $\boldsymbol{CL_{m,n}}$ for $\boldsymbol{w_{m,n}}$
      count the number of words in above two sets $\left|\boldsymbol{ML_{m,n}}\right|$ and $\left|\boldsymbol{CL_{m,n}}\right|$
      sample a topic $\boldsymbol{z_{m,n}}$ for $\boldsymbol{w_{m,n}}$: $\boldsymbol{z_{m,n}} = \boldsymbol{k} \sim \boldsymbol{Multinomial(1/K)}$
      $\boldsymbol{n_m^{(k)} \leftarrow n_m^{(k)} + 1}$; $\boldsymbol{n_k^{(w_{m,n})} \leftarrow n_k^{(w_{m,n})} + 1}$; $\boldsymbol{n_k \leftarrow n_k + 1}$
    end for
  end for
  for i = 1 to I do
    for each document $\boldsymbol{m} \in [1,M]$ do
      for each word $\boldsymbol{n} \in [\boldsymbol{1}, \boldsymbol{N_m}]$ do
        $\boldsymbol{n_m^{(k)} \leftarrow n_m^{(k)} - 1}$; $\boldsymbol{n_k^{(w_{m,n})} \leftarrow n_k^{(w_{m,n})} - 1}$; $\boldsymbol{n_k \leftarrow n_k - 1}$
        sample a topic $\boldsymbol{z_{m,n}}$ for $\boldsymbol{w_{m,n}}$:$\boldsymbol{z_{m,n}} = \boldsymbol{k} \sim \boldsymbol{p(z_i = k|\vec{z}_{\neg i}, \vec{w})}$
        $\boldsymbol{n_m^{(k)} \leftarrow n_m^{(k)} + 1}$; $\boldsymbol{n_k^{(w_{m,n})} \leftarrow n_k^{(w_{m,n})} + 1}$; $\boldsymbol{n_k \leftarrow n_k + 1}$
      end for
    end for
  end for

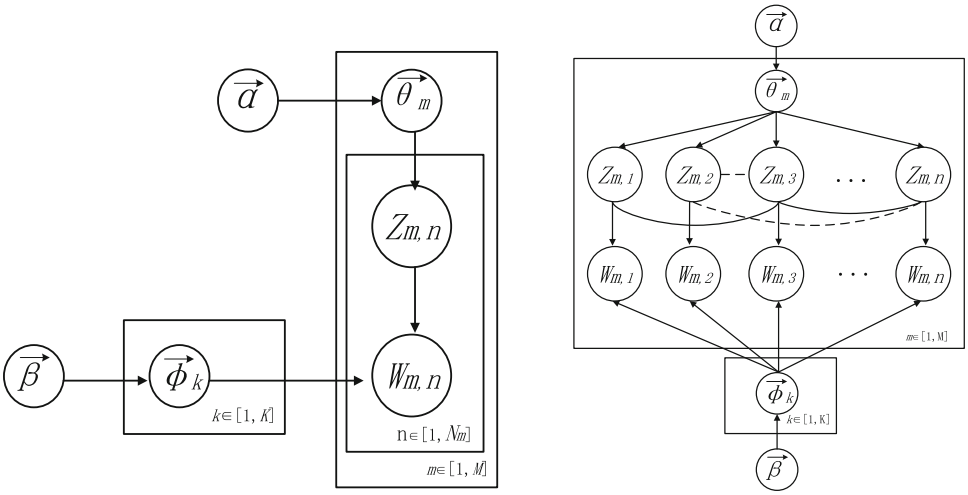


Fig. 1 Graphical model of LDA (**a**) and MWCK-LDA (**b**)

(a): Graphical model of LDA **(b)**: Graphical model of MWCK-LDA

for each word in a document. Although this simplified assumption can improve the modeling efficiency of LDA, it makes the model generate incoherent topics in practice. The incorporation of prior lexical semantic correlation in topic model is of great significance for improving the quality of learned topics.

There are two forms of lexical semantic correlations, which can be divided into must-links and cannot-links. Must-links refer to positive semantic correlation between words, and they should belong to the same topic when topic modeling. Cannot-links refer to negative semantic correlation between words, and they should belong to different topics. For example, "game" and "team" are more likely to belong to the same topic, while "team" and "apple" should belong to different topics. It is critical to use and balance both must-links and cannot-links when topic inferring. Word embeddings aim at embedding the syntax information of words into a continuous vector space and representing a discrete word token using a vector, where words that are semantically related are close to each other

in the distribution of the space. The vector representation of a word in word embeddings makes it possible to model the semantic relationship between words simply by measuring the similarity between word vectors. In this paper, we apply cosine similarity to calculate the correlation between words based on two thresholds μ_1 and μ_2. If the cosine distance between two words' vectors is higher than μ_1, then the two words can be considered to construct a must-link pair indicating a positive semantic correlation exists between these two words. Otherwise, if the cosine similarity between two words' vectors is lower than μ_2, the two words can be considered to construct a cannot-link pair indicating a negative semantic correlation exists between them.

The key idea of the MWCK-LDA in this paper is that if two words in a document construct a must-link they are more likely to belong to the same topic, and otherwise, if two words construct a cannot-link they are more likely to belong to different topics. The graph model of the proposed MWCK-LDA is depicted in Fig. 1b. In MWCK-LDA, we define a mechanism to coordinate must-links and cannot-links during topic modeling process by imposing the Mixed Markov Random Field over the latent topic layer. The Mixed Markov Random Field consists of two kinds of Markov Random Field (MRF) to incorporate different forms of lexical correlation knowledge, respectively, including semantically positively correlated MRF and semantically negatively correlated MRF. By adding this mechanism, the objective function of our proposed model will become more consistent with human judgments than existing models, which can be able to produce more coherent topics.

Given a document m containing N_m words: $\{w_{m,n}\}_{n=1}^{N_m}$, traversing all words and calculating the cosine distances between any word pair based on their corresponding representations of word embeddings. If the distance of a word pair $(w_{m,i}, w_{m,j})$ is higher than threshold μ_1, it indicates that the two words of the word pair can constitute a must-link and semantically positively correlated MRF will define a positive semantic undirected edge between their topic assignments $(z_{m,i}, z_{m,j})$, which is represented by solid undirected line in Fig. 1b. On the contrary, if the distance of a word pair $(w_{m,i}, w_{m,j})$ is lower than threshold μ_2, it indicates that the two words of the pair can constitute a cannot-link and semantically negatively correlated MRF will define a negative semantic undirected edge between their topic assignments $(z_{m,i}, z_{m,j})$. In Fig. 1b, the negative semantic undirected edge is represented by dotted undirected line. After having traversed all word pairs contained in the mth document, the Mixed Markov Random Field creates two undirected graphs, G_{P_m} and G_{N_m}. G_{P_m} consists of the connects between topic assignments corresponding to word pairs with must-link relationship, where latent topic assignments denote nodes and the connects between them denote edges. The set of all edges in the undirected graph G_{P_m} is represented by the notations of P_m. Compared to G_{P_m}, G_{N_m} consists of the connects between topic assignments corresponding to words pairs with cannot-link correlation, where nodes denote latent topic assignments and the connects between them denote edges represented by the notations of N_m. In Fig. 1b, $P_m = \{(z_{m,1}, z_{m,3}), (z_{m,3}, z_{m,n}), \ldots\}$ and $N_m = \{(z_{m,2}, z_{m,3}), (z_{m,2}, z_{m,n}), \ldots\}$.

To encode semantic correlation between words, for each edge in P_m and N_m, the Mixed Markov Random Field defines a binary potential to make words in a must-link tend to be assigned to the same topic, while words in a cannot-link tend not to be. Specifically, to encode must-link correlation, for each edge $(z_{m,i}, z_{m,j})$ in P_m our strategy defines a binary potential as $\exp\{I(z_{m,i} = z_{m,j})\}$, where $I(\cdot)$ is the indicator function. Under this circumstance, if the latent topic assignments are the same, the binary potential function will generate a large value, and otherwise, the binary potential function will generate a small value if two topics are different. As a result, the potential function will increase the probability to assign the words in a must-link to the same topic. Then to encode cannot-link correlation, for each edge $(z_{m,f}, z_{m,g})$ in N_m, the proposed method defines another binary potential as $\exp\{I(z_{m,f} \neq z_{m,g})\}$. If two topics are different, the binary potential function will yield a large value which leads to two words in a cannot-link that can be encouraged to be assigned to different topics. In our model, the joint probability of all topic assignments $\{z_{m,n}\}_{n=1}^{N_m}$ in document m can be written as

$$p(\vec{z}_m|\vec{\theta}_m, \lambda_1, \lambda_2) = \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\theta}_m) \exp\left\{\lambda_1 \frac{\sum_{(z_{m,i}, z_{m,j}) \in P_m} I(z_{m,i} = z_{m,j})}{|P_m|}\right\} \exp\left\{\lambda_2 \frac{\sum_{(z_{m,f}, z_{m,g}) \in N_m} I(z_{m,f} \neq z_{m,g})}{|N_m|}\right\} \quad (4)$$

where $|P_m|$ is the number of all edges in P_m and $|N_m|$ is the number of all edges in N_m. λ_1 and λ_2 are two hyper-parameters to be set by users to balance the effect of external knowledge and statistical information of word co-occurrence contained in corpus. From the joint probability of all topic assignments in document m, the sampling of latent topic for each word is the joint effect of multinomial topic distribution $\vec{\theta}_m$ and word correlation knowledge in current document. The generative process of MWCK-LDA is described as follows.

1. For kth topic, where $k = 1, 2, \ldots, K$

   Draw $\vec{\varphi}_k \sim Dir\left(\vec{\beta}\right)$

2. For mth document, where $m = 1, 2, \ldots, M$

   (a) Draw $\vec{\theta}_m \sim Dir(\vec{\alpha})$

   (b) For nth token in mth document, where $n = 1, 2, \ldots, N_m$

   Draw $z_{m,n} \sim p\left(\vec{z}_m | \vec{\theta}_m, \lambda_1, \lambda_2\right)$

   Draw $w_{m,n} \sim Multi\left(\vec{\varphi}_{z_{m,n}}\right)$

In this paper, collapsed Gibbs sampling inference method is used to estimate parameters of the proposed model. Collapsed Gibbs sampling has been widely used in many topic models, and the derivation process of final Gibbs sampler for MWCK-LDA will be discussed in detail in the next section.

The derived approximate Gibbs sampler has the following conditional distribution.

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)} + \alpha\right)} \cdot \frac{n_{k,\neg i}^{(w_i)} + \beta}{\sum_{i=1}^{V}\left(n_{k,\neg i}^{(w_i)} + \beta\right)} \cdot \exp\left(\lambda_1 \frac{\sum_{j \in ML_{m,i}}\left(z_j = k\right)}{\left|ML_{m,i}\right|}\right) \cdot \exp\left(\lambda_2 \frac{\sum_{j \in CL_{m,i}}\left(z_j \neq k\right)}{\left|CL_{m,i}\right|}\right) \tag{5}$$

where $n_{m,\neg i}^{(k)}$ is the number of words assigned to topic k in document m, $n_{k,\neg i}^{(w_i)}$ is the number of times of word w_i assigned to topic k and $\neg i$ denotes word w_i is excluded from counting in document m. $ML_{m,i}$ denotes the set of words which are labeled to construct must-links with w_i in mth document and $\left|ML_{m,i}\right|$ is the number of all words in $ML_{m,i}$. $CL_{m,i}$ denotes the set of words which are labeled to construct cannot-links with w_i in mth document and $\left|CL_{m,i}\right|$ is the number of all words in $CL_{m,i}$. α and β are predefined Dirichlet hyper-parameters. K is the number of predefined latent topics. The details of Gibbs sampling process of MWCK-LDA are described in Algorithm 1. At the beginning, we randomly assign an initial topic for each word in the corpus and count $n_m^{(k)}$, $n_k^{\left(w_{m,n}\right)}$ and n_k. Given the iterations number of Gibbs sampling I, a new latent topic is reassigned for each word in the corpus according to the derived approximate Gibbs sampler and then update the above parameters in each iteration.

As discussed in Sect. 3.1, standard LDA model follows the Dirichlet distributions to draw discrete document-topic distributions and topic-word distributions. The above assumption strategy is mostly motivated from the perspective of learning efficiency due to the conjugacy between the Dirichlet distribution and the multinomial distribution. However, using Dirichlet distributions as prior will ignore the correlations between tokens in the documents, and in fact, the tokens of a sample are independent. In addition, modeling the word co-occurrence patterns from texts corpus is hard due to the heavy tail nature of the vocabulary. Equation (3) is the sampler to infer hidden variables in standard topic model. From the sampler, we can observe the standard topic model mostly focuses on modeling irrelevant but common word co-occurrence patterns in the corpus, which cannot guide the model toward the intended state for the latent variables. On the contrary, if the distribution of the latent variables can capture lexical semantic correlation, it will be more accordance with human judgments. Based on the above analysis, it is critical to retain desired constraints and bias the latent distribution toward more intended state. Furthermore, there exists abundant lexical semantic correlation in documents corpus like must-links and cannot-links, which are accordance with human cognition. In our MWCK-LDA, to encode prior lexical correlation of must-links and cannot-links, the Mixed Markov Random Field is constructed over the latent topic layer in each document. Equation (4) is the joint probability for each document m of our method. From the likelihood of all topic assignments in document m, we can observe that the sampling of latent topic for each word is not independent, but retaining the desired lexical correlation knowledge in current document. The sampler of our model is listed in Eq. (5). Unlike standard topic model, we use the word–word correlation knowledge to influence the posterior distribution to bias the model to allocate similar words to the same topic and dissimilar words to the different topics. By imposing regularization on the posterior distribution of latent variables during topic modeling, we ensure desired constraints information retained in the learned model. As a result, this knowledge incorporation mechanism in our method can bias our model toward more accordance with human judgments.

3.3 Gibbs sampling method for MWCK-LDA

In this section, we will discuss how to derive the conditional distribution $p(z_i = k|\vec{z}_{\neg i}, \vec{w})$ of approximate Gibbs sampler used in our MWCK-LDA model. According to the definition of conditional probability, $p(z_i = k|\vec{z}_{\neg i}, \vec{w})$ can be obtained as follows

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = \frac{p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)}{p(\vec{w}, \vec{z}_{\neg i}|\alpha, \beta, \lambda_1, \lambda_2)} \propto \frac{p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)}{p(\vec{w}_{\neg i}, \vec{z}_{\neg i}|\alpha, \beta, \lambda_1, \lambda_2)}. \tag{6}$$

From Eq. (6), to derive the conditional distribution we need first to explore how to derive the distribution $p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)$. The graphical model of MWCK-LDA in Fig. 1b shows $p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)$ as follows:

$$p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2) = p(\vec{w}|\vec{z}, \beta)p(\vec{z}|\alpha, \lambda_1, \lambda_2). \tag{7}$$

As given in Eq. (7), deriving the distribution $p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)$ can be translated to derive $p(\vec{w}|\vec{z}, \beta)$ and $p(\vec{z}|\alpha, \lambda_1, \lambda_2)$. For $p(\vec{w}|\vec{z}, \beta)$ in MWCK-LDA, the derivation process is the same as that in LDA model. We can get $p(\vec{w}|\vec{z}, \beta)$ by integrating with respect to $\emptyset$, $p(\vec{w}|\vec{z}, \beta) = \int p(\vec{w}|\vec{z}, \emptyset)p\left(\emptyset|\vec{\beta}\right)\mathrm{d}\emptyset$. According to assumptions of topic model, $p\left(\emptyset|\vec{\beta}\right)$ is a Dirichlet distribution and $p(\vec{w}|\vec{z}, \emptyset)$ is a multinomial distribution. Because of the conjugacy between Dirichlet distribution and multinomial distribution, $p(\vec{w}|\vec{z}, \beta)$ can be finally obtained as in Eq. (8).

$$p(\vec{w}|\vec{z}, \beta) = \int p(\vec{w}|\vec{z}, \emptyset)p\left(\emptyset|\vec{\beta}\right)\mathrm{d}\emptyset = \prod_{k=1}^{K}\frac{\Delta\left(\vec{n}_k + \vec{\beta}\right)}{\Delta\left(\vec{\beta}\right)} \tag{8}$$

where $\vec{n}_k = \left\{n_k^{(w_i)}\right\}_{i=1}^{V}$, $n_k^{(w_i)}$ is the number of times that word w_i assigned to topic k. Here, we use the Δ function defined by Heinrich (2005). The expression of Δ function is described in Eq. (9).

$$\Delta\left(\vec{\beta}\right) = \frac{\prod_{t=1}^{V}\Gamma(\beta)}{\Gamma\left(\sum_{t=1}^{V}\beta\right)} \tag{9}$$

Based on the definition of Δ function, for $\Delta\left(\vec{n}_k + \vec{\beta}\right)$ we have

$$\Delta\left(\vec{n}_k + \vec{\beta}\right) = \frac{\prod_{t=1}^{V}\Gamma\left(n_k^{(w_i)} + \beta\right)}{\Gamma\left(\sum_{t=1}^{V}\left(n_k^{(w_i)} + \beta\right)\right)} = \frac{\prod_{t=1}^{V}\Gamma\left(n_k^{(w_i)} + \beta\right)}{\Gamma(n_k + V\beta)} \tag{10}$$

where n_k denotes the total number of words occur in topic k, $n_k = \sum_{i=1}^{V}\left(n_k^{(w_i)}\right)$.

Then we investigate how to derive $p(\vec{z}|\alpha, \lambda_1, \lambda_2)$. Similar to $p(\vec{w}|\vec{z}, \beta)$, we can get $p(\vec{z}|\alpha, \lambda_1, \lambda_2)$ by integrating over θ, $p(\vec{z}|\alpha, \lambda_1, \lambda_2) = \int p(\vec{z}|\theta, \lambda_1, \lambda_2)p(\theta|\alpha)\mathrm{d}\theta$. From Eq. (4), we can get

$$p(\vec{z}|\alpha, \lambda_1, \lambda_2) = \int p(\vec{z}|\theta)\exp\left\{\lambda_1\frac{\sum_{(z_{m,i}, z_{m,j})\in P_m} I\left(z_{m,i} = z_{m,j}\right)}{|P_m|}\right\} \exp\left\{\lambda_2\frac{\sum_{(z_{m,f}, z_{m,g})\in N_m} I\left(z_{m,f} \neq z_{m,g}\right)}{|N_m|}\right\}p(\theta|\alpha)\mathrm{d}\theta \tag{11}$$

where $p(\vec{z}|\theta)$ is a multinomial distribution and $p(\theta|\alpha)$ follows Dirichlet distribution. Two binary potentials, $\exp\left\{\lambda_1\frac{\sum_{(z_{m,i}, z_{m,j})\in P_m} I\left(z_{m,i} = z_{m,j}\right)}{|P_m|}\right\}$ and $\exp\left\{\lambda_2\frac{\sum_{(z_{m,f}, z_{m,g})\in N_m} I\left(z_{m,f} \neq z_{m,g}\right)}{|N_m|}\right\}$, are not related to variable θ. Similar to $p(\vec{w}|\vec{z}, \beta)$, taking conjugacy between $p(\vec{z}|\theta)$ and $p(\theta|\alpha)$ into consideration we can obtain

$$p(\vec{z}|\alpha, \lambda_1, \lambda_2) = \prod_{m=1}^{M}\frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \exp\left\{\lambda_1\frac{\sum_{(z_{m,i}, z_{m,j})\in P_m} I\left(z_{m,i} = z_{m,j}\right)}{|P_m|}\right\} \exp\left\{\lambda_2\frac{\sum_{(z_{m,f}, z_{m,g})\in N_m} I\left(z_{m,f} \neq z_{m,g}\right)}{|N_m|}\right\}. \tag{12}$$

After having obtained $p(\vec{w}|\vec{z}, \beta)$ and $p(\vec{z}|\alpha, \lambda_1, \lambda_2)$, the conditional distribution $p(z_i = k|\vec{z}_{\neg i}, \vec{w})$ of Gibbs sampling can be derived combining Eqs. (6) and (7) as follows:

$$\begin{aligned} p(z_i = k|\vec{z}_{\neg i}, \vec{w}) &= \frac{p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)}{p(\vec{w}, \vec{z}_{\neg i}|\alpha, \beta, \lambda_1, \lambda_2)} \\ &\propto \frac{p(\vec{w}, \vec{z}|\alpha, \beta, \lambda_1, \lambda_2)}{p(\vec{w}_{\neg i}, \vec{z}_{\neg i}|\alpha, \beta, \lambda_1, \lambda_2)} \\ &\propto \frac{\Delta\left(\vec{n}_k + \vec{\beta}\right)}{\Delta\left(\vec{n}_{k,\neg i} + \vec{\beta}\right)} \cdot \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})} \\ &\quad \cdot \exp\left(\lambda_1\frac{\sum_{j\in ML_{m,i}}\left(z_j = k\right)}{\left|ML_{m,i}\right|}\right) \cdot \exp\left(\lambda_2\frac{\sum_{j\in CL_{m,i}}\left(z_j \neq k\right)}{\left|CL_{m,i}\right|}\right) \\ &\propto \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)} + \alpha\right)} \cdot \frac{n_{k,\neg i}^{(w_i)} + \beta}{\sum_{i=1}^{V}\left(n_{k,\neg i}^{(w_i)} + \beta\right)} \\ &\quad \cdot \exp\left(\lambda_1\frac{\sum_{j\in ML_{m,i}}\left(z_j = k\right)}{\left|ML_{m,i}\right|}\right) \cdot \exp\left(\lambda_2\frac{\sum_{j\in CL_{m,i}}\left(z_j \neq k\right)}{\left|CL_{m,i}\right|}\right). \end{aligned} \tag{13}$$

We can get latent topic assignments over each word in the corpus after finishing Gibbs sampling and then estimate the document-topic distribution $\vec{\theta}_m$, topic-word distribution $\vec{\varphi}_k$ as follows:

Table 3 Topics inferred from NIPS dataset

Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)	Topic 1 (Vision)	Topic 2 (Neural Net)	Topic 3 (Speech)	Topic 4 (Circuits)
LDA				*GK-LDA*			
image	network	hmm	circuit	image	network	speech	circuit
images	neural	system	analog	images	neural	word	analog
pixel	**information**	context	chip	pixel	**result**	recognition	current
scale	feedforward	recognition	current	region	architecture	speaker	chip
filter	**paper**	speech	voltage	vision	**paper**	phoneme	voltage
wavelet	architecture	hidden	vlsi	local	feedforward	acoustic	vlsi
resolution	research	training	**output**	texture	feed	system	transistor
scene	**san**	markov	**input**	**disparity**	architectures	letter	**output**
visual	processing	states	transistor	scene	research	vowel	pulse
vision	propagation	probabilities	pulse	surface	**application**	tdnn	silicon
natural	artificial	continuous	**synapse**	**result**	artificial	frame	implementation
texture	**university**	**mlp**	implementation	depth	**introduction**	phonetic	**mead**
region	**pages**	hybrid	**mead**	edge	forward	**waibel**	cmos
result	architectures	word	silicon	edges	**vol**	signal	signal
level	**application**	**dependent**	cmos	contour	**form**	experiment	**winner**
SC-LDA				*WE-LDA*			
image	layer	speech	vlsi	image	network	speech	circuit
images	net	word	chip	images	neural	recognition	voltage
pixel	activation	recognition	pulse	pixel	net	system	current
visual	connection	speaker	silicon	object	**result**	speaker	analog
vision	architecture	acoustic	intensity	vision	**problem**	word	intensity
filter	sigmoid	letter	**retina**	visual	**paper**	hmm	chip
sensor	nodes	segmentation	**mead**	scale	architecture	training	**vibration**
intensity	backpropagation	phoneme	low	recognition	**university**	context	**viscosity**
location	hinton	tdnn	signal	filter	research	acoustic	lightness
features	item	vowel	photoreceptor	view	feedforward	phoneme	vlsi
edges	multilayer	segment	circuit	system	**application**	continuous	transistor
statistic	**spiral**	phonetic	measured	scene	approach	phonetic	brightness
geman	backprop	recognizer	**background**	**presentation**	artificial	**mlp**	pulse
contour	**required**	frames	analog	**information**	general	frame	velocity
optical	**development**	**waibel**	**control**	resolution	**system**	experiment	silicon
MRF-LDA				*MWCK-LDA*			
image	network	speech	chip	image	network	speech	analog
images	unit	recognition	implementation	object	neural	speaker	chip
pixel	input	hmm	processor	images	architecture	acoustic	implementation
vision	hidden	speaker	parallel	pixel	propagation	vowel	vlsi
region	output	acoustic	bit	vision	research	phoneme	hardware
scene	layer	phoneme	hardware	contour	forward	signal	digital
scale	neural	signal	digital	shape	connected	frame	bit
left	**pattern**	**mlp**	**synapse**	scene	artificial	phonetic	template
camera	net	hybrid	block	edge	**desired**	continuous	implemented
color	activation	tdnn	operation	texture	feed	**independent**	device
texture	propagation	continuous	processing	frame	feedforward	consonant	speed
intensity	connection	phonetic	**instruction**	depth	architectures	tdnn	gate
natural	connected	vowel	design	color	fully	spoken	storage
pyramid	forward	frame	speed	grouping	multilayer	utterances	parallel
ieee	**spiral**	**dependent**	computation	perception	backpropagation	phonemes	vol

Table 4 Topics inferred from 20-Newsgroups dataset

Topic 1 (Insurance)	Topic 2 (Sports)	Topic 3 (Health)	Topic 4 (Sex)	Topic 1 (Insurance)	Topic 2 (Sports)	Topic 3 (Health)	Topic 4 (Sex)
LDA				*GK-LDA*			
money	game	health	sex	insurance	game	medical	men
pay	games	medical	men	money	team	health	sex
insurance	baseball	cancer	homosexuality	care	hockey	cancer	gay
cost	team	aids	gay	pay	games	disease	people
care	win	research	homosexual	tax	play	aids	**cramer**
tax	runs	disease	**cramer**	health	season	**number**	**optilink**
private	**year**	**number**	sexual	private	nhl	hiv	sexual
year	**morris**	hiv	people	**canada**	**year**	research	**study**
health	**last**	**use**	**optilink**	government	players	**use**	**virginia**
costs	run	**april**	homosexuals	taxes	teams	**april**	**article**
market	**two**	risk	male	**make**	fans	information	**clayton**
business	**jack**	**page**	**study**	system	cup	**page**	**writes**
buy	series	**newsletter**	**clayton**	**people**	espn	patients	women
make	pitcher	volume	marriage	free	league	volume	homosexual
taxes	pitching	study	**article**	**gary**	win	**national**	homosexuals
SC-LDA				*WE-LDA*			
insurance	game	cancer	homosexuality	cost	playoffs	care	sexual
health	games	health	homosexual	pay	game	healthcare	homosexuals
private	espn	medical	men	pension	seasons	nutrition	gay
care	baseball	drug	sex	paid	matchups	medicare	heterosexual
canada	hockey	hiv	gay	expenditure	homestand	childcare	promiscuity
money	**gant**	disease	**cramer**	tax	games	**education**	sex
pay	play	aids	people	insurance	**doubleheader**	welfare	adultery
canadian	fans	**number**	sexual	subsidy	matchup	**educational**	lesbianism
companies	**time**	patients	**paul**	costing	**streak**	counseling	homophobia
geico	**hirschbeck**	**use**	homosexuals	excise	playoff	counseling	bestiality
doctors	**pens**	**april**	**optilink**	outlay	**shutout**	prevention	sexuality
costs	coverage	drugs	**clayton**	revenues	lineup	tutoring	prostitution
coverage	night	research	male	mortgage	tournaments	sanitation	pedophilia
sgi	caps	**page**	women	fees	victory	health	**abuse**
gld	braves	volume	**study**	totaling	match	**schooling**	homosexuality
MRF-LDA				*MWCK-LDA*			
money	game	medical	sex	money	game	medical	men
cost	team	health	men	cost	team	cancer	sex
buy	**year**	cancer	homosexuality	pay	games	patients	women
price	hockey	disease	homosexual	buy	play	treatment	homosexuality
year	play	aids	gay	insurance	hockey	doctor	homosexual
pay	goal	hiv	**cramer**	care	season	medicine	gay
market	cup	**number**	sexual	market	win	disease	sexual
costs	win	patients	people	tax	fans	study	male
business	player	research	**optilink**	business	teams	studies	marriage
years	series	**april**	male	company	nhl	doctors	love
per	**last**	information	homosexuals	costs	players	candida	homosexuals
sell	**wings**	volume	**study**	private	player	patient	behavior
tax	games	**newsletter**	**clayton**	companies	series	yeast	man
billion	**leafs**	study	**rutgers**	**per**	league	infection	**issues**
company	played	**page**	child	price	division	symptoms	married

$$\theta_m^{(k)} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^{K}\left(n_m^{(k)} + \alpha\right)} \tag{14}$$

$$\varphi_k^{(w_i)} = \frac{n_k^{(w_i)} + \beta}{\sum_{i=1}^{V}\left(n_k^{(w_i)} + \beta\right)}. \tag{15}$$

4 Experiment

This section evaluates the proposed model in this paper and compares it with five state-of-the-art baseline topic models on two benchmark datasets.

LDA: LDA is a classic unsupervised topic model that has been widely used in many tasks (Blei et al. 2003). Many extensions of knowledge-based topic model are all based on LDA.

GK-LDA: GK-LDA is a knowledge-based topic model which can exploit the lexical correlation knowledge in dictionary (Chen et al. 2013a). In addition, GK-LDA has the mechanism of dealing with wrong knowledge based on the ratio between word probabilities under each topic. However, GK-LDA cannot mine any prior knowledge automatically. Thus, we feed GK-LDA, the knowledge produced using the proposed knowledge mining method in our paper, which allows us to assess the knowledge fusion capabilities of each model.

MRF-LDA: A knowledge-based topic model that can use external knowledge of word correlation generated by word embeddings (Xie et al. 2015). However, MRF-LDA can only incorporate must-links, but ignores other forms of word–word correlation.

SC-LDA: A knowledge-based topic model that can incorporate prior word correlation knowledge based on Word-Net 3.0 and word embeddings (Yang et al. 2015c). Given that extracting word correlation knowledge from synsets in Word-Net 3.0 may restrict the quantity of knowledge generation, we feed the knowledge generated in MWCK-LDA into SC-LDA to compare the knowledge incorporation mechanism between them.

WE-LDA: WE-LDA is a knowledge-based topic model which can use must-links to improve topic modeling performance (Yao et al. 2017). Similar to MRF-LDA, WE-LDA cannot be able to incorporate other forms of knowledge.

4.1 Experimental settings

- *Datasets* We use two public benchmark datasets for our experiments, which have been used to validate the performance of many topic models. The first dataset is 20-Newsgroups.[1] The 20-Newsgroups dataset has been one of the most commonly used benchmarks in natural language processing domain including topic modeling, text classification and text clustering. It collects about 20,000 newsgroup documents divided into 20 different categories. The second dataset we choose to evaluate the performance of topic models is NIPS.[2] The NIPS dataset is collected from Annual Conference on Neural Information Processing Systems (NIPS) published in the time range from 1987 to 1999 and it contains about 1500 documents. For 20-Newsgroups, we conduct the following preprocessing: (1) convert letters into lowercase and (2) remove stop words using stop words list of NLTK. For NIPS, we use the original documents for topic modeling.
- *Word embeddings* In our paper, Web Eigenwords[3] is selected to be the source to mine word–word correlation of must-links and cannot-links. In Web Eigenwords, a word is represented by a real-valued vector. To generate word correlation knowledge, we should calculate the cosine distance between word vectors corresponding to two words.
- *Parameter settings* In experiment for all models, posterior estimates of latent variables are obtained with 2000 iterations of Gibbs sampling. For two benchmark datasets, we set the number of latent topics as 100. For other parameters of baseline models, we set them according to their original papers. For LDA, α and β are set as $50/K$ and 0.01. For GK-LDA, we set $\alpha = 1$ and $\beta = 0.1$. For MRF-LDA, we set α, β as $50/K$ and 0.01. In MRF-LDA, to incorporate must-links from Web Eigenwords we set threshold μ_1 to be 0.99 and λ_1 to be 1. For SC-LDA, α and β are set as 1.0 and 0.01. For WE-LDA, the parameters are set as suggested in the original paper. For MWCK-LDA model, we set α and β to be $50/K$ and 0.01. λ_1 and λ_2 are both set as 1 to balance the effect between must-links and cannot-links. To generate knowledge of word–word correlation in this paper, two thresholds μ_1 and μ_2 are set as 0.99 and 0.1. Word pairs with similarity higher than μ_1 are labeled as must-links and word pairs with similarity lower than μ_2 are labeled as cannot-links.

4.2 Experimental results

In this section, we evaluate the performance of the developed model in both qualitative and quantitative evaluations. In Sect. 4.2.1, we present the qualitative evaluation

[1] http://qwone.com/jason/20Newsgroups/.

[2] http://archive.ics.uci.edu/ml/datasets/Bag+of+Wor.

[3] http://www.cis.upenn.edu/ungar/eigenwords/.

of our model. In Sect. 4.2.2, we illustrate the performance of the proposed model from the perspective of the quantitative evaluation. In addition, we discuss the effects of the thresholds μ_1 and μ_2 to the modeling performance of MWCK-LDA in Sect. 4.2.3. Then finally, we discuss the effects of λ_1 and λ_2 to the MWCK-LDA.

4.2.1 Qualitative evaluation

In this section, we will evaluate our MWCK-LDA with the baseline models in qualitative. Table 3 shows some topic modeling results extracted from the NIPS dataset by six models. Each topic is represented by the top fifteen words, and we highlight the words that lack representativeness in bold. From the meanings of words contained in each topic, we can know that these four topics are about "Vision," "Neural Net," "Speech" and "Circuits," respectively. Table 3 shows that MWCK-LDA proposed in this paper can mine more coherent topic results with fewer noisy and meaningless words than all other baseline models. From the topics extracted by LDA, we can see that some noisy words which cannot effectively characterize a topic appear at the top positions of each topic due to their high frequency, such as word "natural" in topic 1 and word "paper" in topic 2. This is because LDA is a totally unsupervised topic model which generates the words independently based only on word co-occurrence information and lacks the mechanism to incorporate external word correlation knowledge. As to GK-LDA and MRF-LDA, both models can be capable of applying word correlation knowledge of must-links during topic modeling process. Compared to standard LDA, they can take similarity relationships among words into consideration and mine topics that are more meaningful. We can see that the number of noisy words in topics inferred by GK-LDA and MRF-LDA is all lower than that in LDA. The difference between GK-LDA and MRF-LDA lies in the mechanism of incorporating and assessing word correlation knowledge. Compared to GK-LDA, MRF-LDA has a superior mechanism that can generate topics with better quality. The quality of topics generated by GK-LDA is unsatisfactory. For example, GK-LDA cannot solve the confusion problem in "Neural Net" topic which confuses a "paper" topic and a "neural network" topic, even though GK-LDA can extract higher-quality topics overall. In comparison, MRF-LDA can solve the above problem and generate more coherent topic. Although GK-LDA and MRF-LDA can be able to incorporate word correlation knowledge of must-links into topic modeling process, there are still some noise words existing in each topic. For example, in MRF-LDA, noise words such as "left" and "ieee" appear in topic "Vision," and "pattern" and "spiral" appears in topic "Neural Net." In GK-LDA, noise words such as "result" and "paper" appear in topic "Neural Net," and "waibel" appears in topic "Speech." These noise words in each topic are all semantic uncorrelated with other words. Our model MWCK-LDA defines the Mixed Markov Random Field at the latent topic layer to improve possibilities for correlated words to be assigned into the same topic and uncorrelated words to be assigned to different topics. As a result, Table 3 shows that the topics mined by our model are more coherent than those mined by baseline models. The extracted results by MWCK-LDA are more coherent and contain fewer meaningless words.

Table 4 shows the topics mined from 20-Newsgroups dataset. The four topics are about "Insurance," "Sports," "Health" and "Sex," respectively. Form the table, the modeling results mined by our MWCK-LDA are far better than those mined by baseline models. It demonstrates that our model is more effective than other baseline topic models.

4.2.2 Quantitative evaluation

For quantitative evaluation of topic modeling results, we choose Coherence Measure (CM) as quantitative metric to compare topic quality to other baseline models. Coherence Measure has been used as quantitative metric of topic modeling performance in many works of topic model (Qiang et al. 2017; Xie and Xing 2013; Xie et al. 2015). In our experiment, we select the top 15 words of each topic to assess whether words are relevant to current topic by judgments of human annotators. During the assessment process, we ask annotators to evaluate whether the meaning of current topic is obvious or not at the beginning. If not, the fifteen candidate words in current topic are both marked as to be irrelevant. Otherwise, annotators need to label words that are relevant to current topic. Finally, the metric of CM can be calculated by the percentage of the number of relevant words over total number of candidate words.

In our experiment, there are four graduate students participating in annotation experiment. For each dataset, ten topics were randomly selected for evaluating. The Coherence Measure results on NIPS dataset and 20-Newsgroups dataset are shown in Tables 5 and 6, respectively. We can observe that our model performs better than other baseline models on two benchmark datasets. From Tables 5 and 6, our model achieves 83.5% and 70.75% at the average Coherence Measure metric on the NIPS dataset and 20-Newsgroups dataset, respectively, which are higher than other baseline models. In addition, from the experimental results we can also assess the consistency performance by different annotators. We can observe that performance by different annotators shows good consistency between each other. Experimental results

Table 5 CM (%) on NIPS dataset

Method	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Mean	Standard deviation
LDA	56	62	58	54	57.5	2.96
GK-LDA	66	68	62	63	64.75	2.38
SC-LDA	50	55	57	53	53.75	2.59
WE-LDA	59	61	59	56	58.75	1.79
MRF-LDA	59	64	60	60	60.75	1.92
MWCK-LDA	86	82	82	84	83.5	1.66

Table 6 CM (%) on 20-Newsgroups dataset

Method	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Mean	Standard deviation
LDA	42	46	50	42	45	3.32
GK-LDA	38	46	40	42	41.5	2.96
SC-LDA	48	51	51	55	51.25	2.49
WE-LDA	61	62	65	60	62	1.88
MRF-LDA	44	50	46	49	47.25	2.38
MWCK-LDA	72	68	72	71	70.75	1.64

in quantitative evaluation demonstrate the effective of our model.

4.2.3 Influence of μ_1 and μ_2

In this section, we discuss the influence of the thresholds μ_1 and μ_2 in MWCK-LDA. In this experiment, we choose the metric of topic coherence to measure the quality of topic modeling results. The metric of topic coherence is proposed to assess topic quality. Given a topic z and its top T most probable words, metric of topic coherence is obtained as follows and a higher metric value implies a better performance of topic modeling:

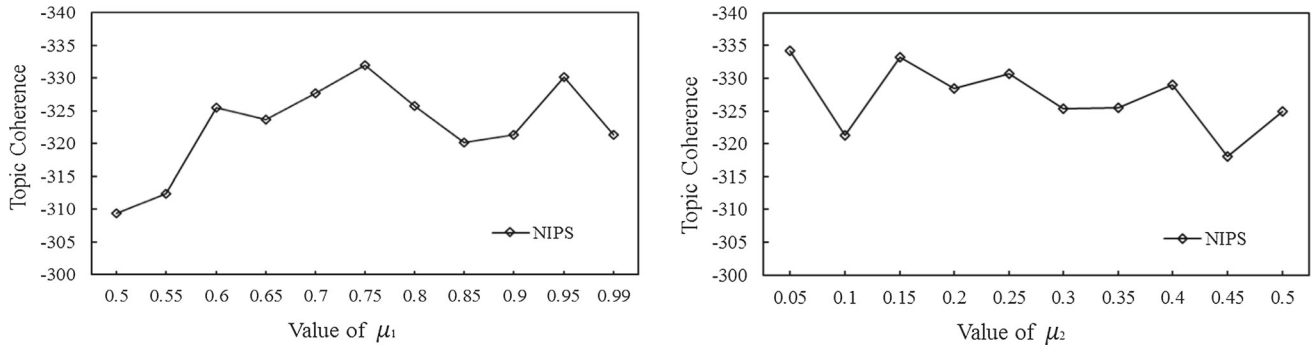
$$\mathrm{C} = \frac{1}{K}\sum_{z=1}^{K}\sum_{t=2}^{T}\sum_{l=1}^{t-1}\log\frac{D\left(w_t^z, w_l^z\right)+1}{D\left(w_l^z\right)} \tag{16}$$

where $\left(w_1^z, \cdots\cdots, w_T^z\right)$ is a list of the T most probable words in topic z. $D(w)$ denotes the document frequency of word w. $D(w, w')$ denotes the co-document frequency of w and w'. Figure 2 shows the metric of topic coherence over different μ_1 and μ_2 on NIPS dataset with the setting of K = 100 and T = 20. In addition, Fig. 3 presents the metric of topic coherence over different μ_1 and μ_2 on 20-Newsgroups dataset. Other parameters are all fixed as the settings in the experiment. As discussed in Sect. 3.2, thresholds μ_1 and μ_2 determine the quality of the generated word correlation knowledge.

Figure 2a shows that MWCK-LDA achieves the highest topic coherence value when $\mu_1 = 0.5$ on NIPS dataset. This indicates that setting the value of μ_1 equals 0.5 can generate higher quality of must-link correlation knowledge. When the value of μ_1 is less than the threshold of 0.5, the generated must-link correlation knowledge contains much noise and this will hinder our model inferring high-quality topics. In contrast, when the value of μ_1 is set to more than 0.5, the generated must-link correlation knowledge will be limited because of the high threshold. The limited must-link correlation knowledge will also hinder our model working well. Figure 2b shows that MWCK-LDA achieves the highest topic coherence value when $\mu_2 = 0.45$ on NIPS dataset. This indicates that setting the value of μ_2 equal to 0.45 can generate higher quality of cannot-link correlation knowledge. Different from lexical correlation of must-links, when the value of μ_2 is set to more than 0.45, the generated cannot-link correlation knowledge will contain much noise knowledge and hinder MWCK-LDA mining coherent topics. Furthermore, when the value of μ_2 is set to less than 0.45, the generated cannot-link knowledge will be limited. Similar phenomenon can also be observed in 20-Newsgroups dataset from Fig. 3a, b. Figure 3a, b shows that when μ_1 is set as 0.55 or μ_2 is set as 0.3 our model performs best on 20-Newsgroups dataset. The optimal values of μ_1 or μ_2 depend on the specific dataset. In addition, we chose Web Eigenwords as the source to mine lexical correlation knowledge. The approximate thresholds of μ_1 and μ_2 also depend on the selected word vectors.

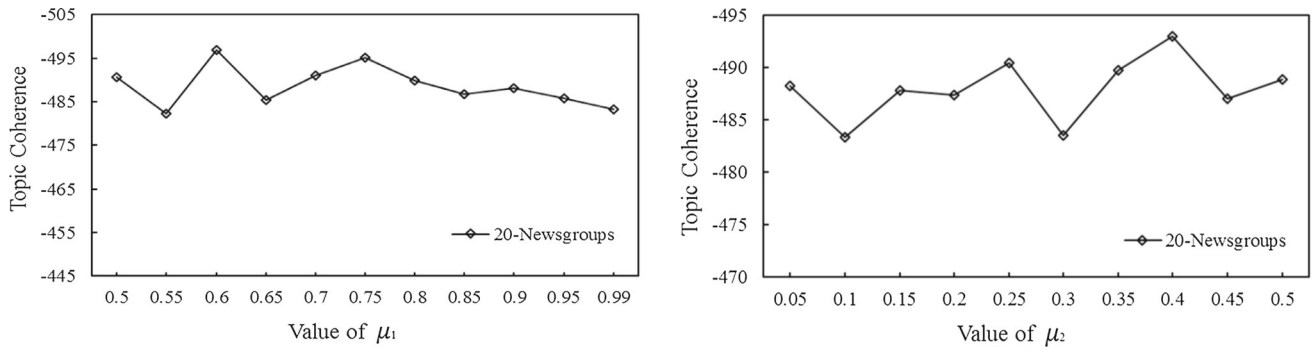
4.2.4 Influence of λ_1 and λ_2

In this section, we investigate the effect of parameters λ_1 and λ_2 in MWCK-LDA. Figure 4 shows the topic coherence over different λ_1 and λ_2 on NIPS dataset with the setting of K = 100 and T = 20. Figure 5 shows the topic coherence over different λ_1 and λ_2 on 20-Newsgroups dataset with the setting of K = 100 and T = 20. When changing parameters λ_1 or λ_2, other parameters are all



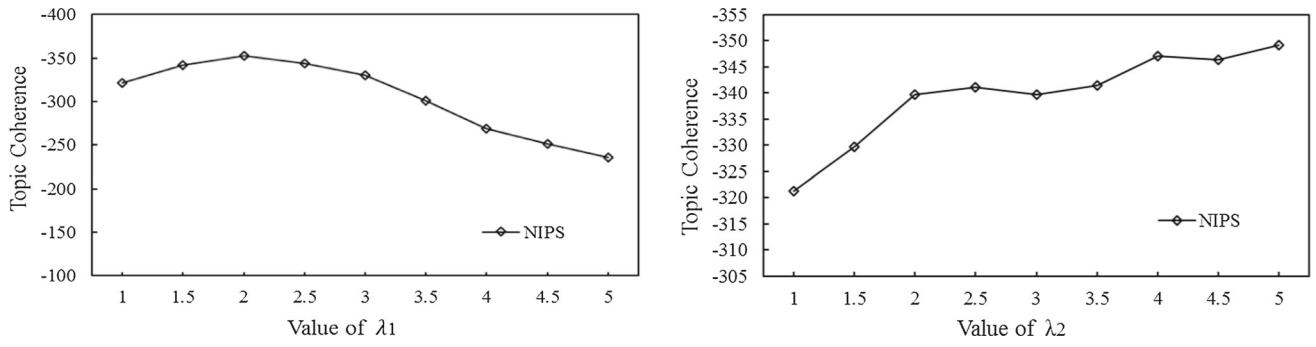
(a) Effect of μ_1 on NIPS **(b)** Effect of μ_2 on NIPS

Fig. 2 Effect of μ_1 and μ_2 on NIPS dataset under $K = 100$ and $T = 20$



(a) Effect of μ_1 on 20-Newsgroups **(b)** Effect of μ_2 on 20-Newsgroups

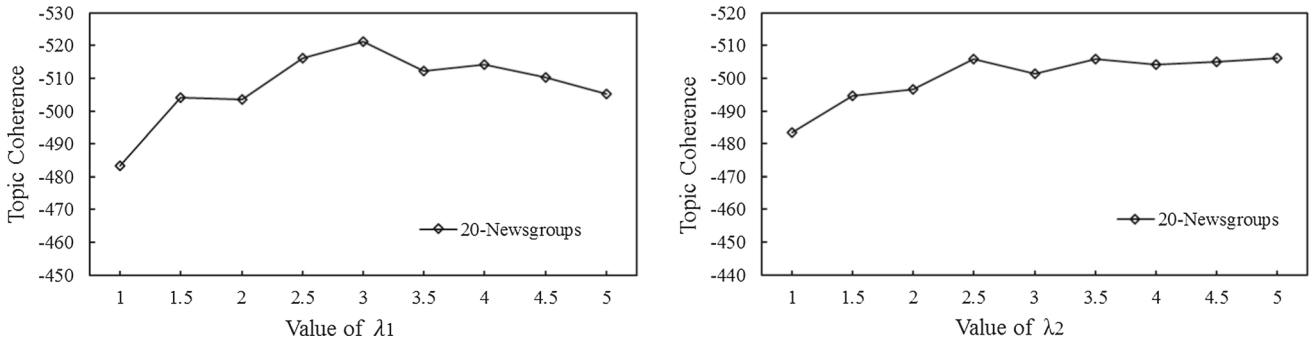
Fig. 3 Effect of μ_1 and μ_2 on 20-Newsgroups dataset under $K = 100$ and $T = 20$



(a) Effect of λ_1 on NIPS **(b)** Effect of λ_2 on NIPS

Fig. 4 Effect of λ_1 and λ_2 on NIPS dataset under $K = 100$ and $T = 20$

fixed, following the experimental settings. As discussed in Sect. 3.2, in our model parameter λ_1 balances the effect between external correlation knowledge of must-links and other information patterns when inferring latent topics. Figure 4a shows that MWCK-LDA has the highest topic coherence value when $\lambda_1 = 5$ on NIPS dataset. In 20-Newsgroups dataset, when the value of λ_1 equals 1, our model performs the best. This indicates that we need to set specific parameter value of λ_1 for different datasets to balance the effect of prior correlation knowledge of must-links and other information sources. For a specific dataset, we can find an optimal value of λ_1. When the parameter of λ_1 is less than the suitable threshold, the effect of must-link correlation knowledge on topic extraction will be weakened. In contrast, when the parameter of λ_1 is more than the threshold, the effect of other information patterns like co-occurring word pair et al. on topic extraction will be weakened. It is important to balance the effects of different



(a) Effect of λ_1 on 20-Newsgroups

(b) Effect of λ_2 on 20-Newsgroups

Fig. 5 Effect of λ_1 and λ_2 on 20-Newsgroups dataset under $K = 100$ and $T = 20$

information sources on topic sampling in MWCK-LDA to mine coherent topic results. The situations discussed above will hinder our model working well. As for parameter of λ_2, in MWCK-LDA λ_2 balances the effect between prior cannot-link correlation knowledge and other knowledge patterns when topic modeling. Figures 4b and 5b show that when the value of λ_2 equals 1, our model performs the best on both 20-Newsgroups dataset and NIPS dataset. This indicates that we should set λ_2 equals 1 for different datasets to incorporate cannot-link correlation knowledge well.

5 Conclusions

In this paper, we propose a Mixed Word Correlation Knowledge-based Latent Dirichlet Allocation topic model (abbr. to MWCK-LDA) to infer latent topics from documents corpus. We find that MWCK-LDA can incorporate both must-links and cannot-links generated by word embeddings in a soft manner. To incorporate the above knowledge, the Mixed Markov Random Field is constructed over the latent topic layer to regularize the topic assignment of each word during the topic modeling process, which will give must-links a better chance to be put into the same topic and cannot-links a better chance to be not. In addition, the developed knowledge incorporation mechanism enable a good balance between two forms of external knowledge and word co-occurrence information contained in documents when topic sampling. Experimental results show that MWCK-LDA can achieve significantly better performance than baseline topic models. As future research work, we will focus on how to incorporate more abundant knowledge forms into topic models to improve the coherence of modeling results further.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Nos. 91646102, L1824039, L1724034, L1724026, L1524015, L1624045), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (16JDGC011), the Construction Project of China Knowledge Center for Engineering Sciences and Technology (No. CKCEST-2019-2-13), the UK–China Industry Academia Partnership Program (UK-CIAPP/260), the Tsinghua University Project of Volvo-supported Green Economy and Sustainable Development (20153000181) and the Tsinghua Initiative Research Project (2016THZW).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Ahmed A, Long J, Silva D, Wang Y (2017) A practical algorithm for solving the incoherence problem of topic models in industrial applications. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1713–1721

Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In: Proceedings of the 26th annual international conference on machine learning, pp 25–32

Blei DM, Lafferty JD (2005) Correlated topic models. In: Proceedings of the 18th international conference on neural information processing systems, pp 147–154

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res Arch 3:993–1022

Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Proceedings of the 22nd international conference on neural information processing systems, pp 288–296

Chen Z, Liu B (2014a) Mining topics in documents: standing on the shoulders of big data. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1116–1125

Chen Z, Liu B (2014b) Topic modeling using topics from many domains, lifelong learning and big data. In: Proceedings of the 31st international conference on international conference on machine learning, pp II-703–II-711

Chen Z, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013a) Discovering coherent topics using general knowledge. In: Proceedings of the 22nd ACM international conference on information & knowledge management, pp 209–218

Chen Z, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013b) Leveraging multi-domain prior knowledge in topic models. In: Proceedings of the twenty-third international joint conference on artificial Intelligence, pp 2071–2077

Fang A, Macdonald C, Ounis I, Habel P (2016) Using word embedding to evaluate the coherence of topics from Twitter data. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, pp 1057–1060

Fu X, Sun X, Wu H, Cui L, Huang JZ (2018) Weakly supervised topic sentiment joint model with word embeddings. Knowl-Based Syst 147:43–54

Gao S, Li X, Yu Z, Qin Y, Zhang Y (2017) Combining paper cooperative network and topic model for expert topic analysis and extraction. Neurocomputing 257:136–143

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci USA 101(Suppl 1):5228–5235

Heinrich, G (2005) Parameter estimation for text analysis. Technical report

Hu Y, Boyd-Graber J, Satinoff B, Smith A (2014) Interactive topic modeling. Mach Learn 95:423–469

Jagarlamudi J, Daumé H III, Udupa R (2012) Incorporating lexical priors into topic models. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics, pp 204–213

Lee TY, Alison S, Seppi K, Elmqvist N, Boyd-Graber J, Findlater L (2017) The human touch: how non-expert users perceive, interpret, and fix topic models. Int J Hum Comput Stud 105:28–42

Li X, Ma Z, Peng P, Guo X, Huang F, Wang X, Guo J (2018) Supervised latent Dirichlet allocation with a mixture of sparse softmax. Neurocomputing 312:324–335

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations (ICLR), pp 1–12

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013b) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems, pp 3111–3119

Mimno D, Wallach HM, Talley E, Leenders M, Mccallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing, pp 262–272

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

Petterson J, Smola AJ, Caetano TS, Buntine WL, Narayanamurthy S (2010) Word features for latent Dirichlet allocation. In: Proceedings of the 23rd international conference on neural information processing systems, pp 1921–1929

Qiang J, Chen P, Wang T, Wu X (2017) Topic modeling over short texts by incorporating word embeddings. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp 363–374

Shams M, Baraani-Dastjerdi A (2017) Enriched LDA (ELDA): combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. Expert Syst Appl 80:136–146

Xie P, Xing EP (2013) Integrating document clustering and topic modeling. In: Proceedings of the 20th conference on uncertainty in artificial intelligence, pp 694–703

Xie P, Yang D, Xing E (2015) Incorporating word correlation knowledge into topic modeling. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp 725–734

Xu Y, Yin J, Huang J, Yin Y (2018) Hierarchical topic modeling with automatic knowledge mining. Expert Syst Appl 103:106–117

Xun G, Gopalakrishnan V, Ma F, Li Y, Gao J, Zhang A (2016) Topic discovery for short texts using word embeddings. In: 2016 IEEE 16th international conference on data mining (ICDM), pp 1299–1304

Yang L, Liu Z, Chua TS, Sun M (2015a) Topical word embeddings. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, pp 2418–2424

Yang S, Lu W, Yang D, Yao L, Wei B (2015b) Short text understanding by leveraging knowledge into topic model. In: The 2015 annual conference of the North American Chapter of the ACL, pp 1232–1237

Yang Y, Downey D, Boyd-Graber J (2015c) Efficient methods for incorporating knowledge into topic models. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 308–317

Yao L, Zhang Y, Wei B, Li L, Wu F, Zhang P, Bian Y (2016) Concept over time: the combination of probabilistic topic model with wikipedia knowledge. Expert Syst Appl 60:27–38

Yao L, Zhang Y, Chen Q, Qian H, Wei B, Hu Z (2017) Mining coherent topics in documents using word embeddings and large-scale text data. Eng Appl Artif Intell 64:432–439

Zhu J, Xing EP (2010) Conditional topic random fields. In: Proceedings of the 27th international conference on international conference on machine learning, pp 1239–1246