A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling

Mehdi Allahyari
Computer Science Department
Georgia Sothern University
Statesboro, USA.

Seyedamin Pouriyeh
Computer Science Department
University of Georgia
Athens, USA

Krys Kochut
Computer Science Department
University of Georgia
Athens, USA

Hamid Reza Arabnia
Computer Science Department
University of Georgia
Athens, USA

***Abstract*—Probabilistic topic models, which aim to discover latent topics in text corpora define each document as a multinomial distributions over topics and each topic as a multinomial distributions over words. Although, humans can infer a proper label for each topic by looking at top representative words of the topic but, it is not applicable for machines. Automatic Topic Labeling techniques try to address the problem. The ultimate goal of topic labeling techniques are to assign interpretable labels for the learned topics. In this paper, we are taking concepts of ontology into consideration instead of words alone to improve the quality of generated labels for each topic. Our work is different in comparison with the previous efforts in this area, where topics are usually represented with a batch of selected words from topics. We have highlighted some aspects of our approach including: 1) we have incorporated ontology concepts with statistical topic modeling in a unified framework, where each topic is a multinomial probability distribution over the concepts and each concept is represented as a distribution over words; and 2) a topic labeling model according to the meaning of the concepts of the ontology included in the learned topics. The best topic labels are selected with respect to the semantic similarity of the concepts and their ontological categorizations. We demonstrate the effectiveness of considering ontological concepts as richer aspects between topics and words by comprehensive experiments on two different data sets. In another word, representing topics via ontological concepts shows an effective way for generating descriptive and representative labels for the discovered topics.**

***Keywords*—Topic modeling; topic labeling; statistical learning; ontologies; linked open data**

I. Introduction

Recently, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] has been getting considerable attention. A wide variety of text mining approaches, such as sentiment analysis [2], [3], word sense disambiguation [4], [5], information retrieval [6], [7], summarization [8], and others have been successfully utilized LDA in order to uncover latent topics from text documents. In general, Topic models consider that documents are made up of topics, whereas topics are multinomial distributions over the words. It means that the topic proportions of documents can be used as the descriptive themes at the high-level presentations of the semantics of the documents. Additionally, top words in a topic-word distribution illustrate the sense of the topic. Therefore, topic models can be applied as a powerful technique for discovering the latent semantics from unstructured text collections. Table I, for example, explains the role of topic labeling in generating a representative label based on the words with highest probabilities from a topic discovered from a corpus of news articles; a human assessor has labeled the topic "United States Politics".

Although, the top words of every topic are usually related and descriptive themselves but, interpreting the label of the topics based on the distributions of words derived from the text collection is a challenging task for the users and it becomes worse when they do not have a good knowledge of the domain of the documents. Usually, it is not easy to answer questions such as "What is a topic describing?" and "What is a representative label for a topic?"

TABLE I. Example of a Labeling a Topic

Human Label: United States Politics				
republican	house	senate	president	state
republicans	political	campaign	party	democratic

Topic labeling, in general, aims to find one or a few descriptive phrases that can represent the meaning of the topic. Topic labeling becomes more critical when we are dealing with hundreds of topics to generate a proper label for each.

The aim of this research is to *automatically* generate *good* labels for the topics. But, what makes a label good for a topic? We assume that a good label: 1) should be semantically relevant to the topic; 2) should be understandable to the user; and 3) highly cover the meaning of the topic. For instance, "relational databases", "databases" and "database systems" are a few good labels for the example topic illustrated in Table I.

With advent of the Semantic Web, tremendous amount of data resources have been published in the form of ontologies and inter-linked data sets such as Linked Open Data (LOD)[1]. Linked Open Data provides rich knowledge in multiple domains, which is a valuable asset when used in combination with various analyses based on unsupervised topic models, in particular, for topic labeling. For instance, DBpedia [10] (as part of LOD) is one the most prominent knowledge bases that is extracted from Wikipedia in the form of an ontology consisting of a set of concepts and their relationships. DBpedia, which is freely available, makes this extensive quantity of information programmatically obtainable on the Web for human and machine consumption.

The principal objective of the research presented here is to leverage and integrate the semantic knowledge graph of

[1] http://linkeddata.org/

concepts in an ontology, DBpedia in this paper, and their diverse relationships into probabilistic topic models (i.e. LDA). In the proposed model, we define another latent (i.e. hidden) variable called, *concept*, i.e. ontological concept, between topics and words. Thus, each document is a mixture of topics, while each topic is made up of concepts, and finally, each concept is a probability distribution over the vocabulary.

Defining concepts as an extra latent variable (i.e. representing topics over concepts instead of words) are advantageous in several ways including: 1) it describes topics in a more extensive way; 2) it also allows to define more specific topics according to ontological concepts, which can be eventually used to generate labels for topics; 3) it automatically incorporates topics learned from the corpus with knowledge bases. We first presented our Knowledge-based topic model, KB-LDA model, in [11] where we showed that incorporating ontological concepts with topic models improves the quality of topic labeling. In this paper, we elaborate on and extend these results. We also extensively explore the theoretical foundation of our Knowledge-based framework, demonstrating the effectiveness of our proposed model over two datasets.

Our contributions in this work are as follows:

1) In a very high level, we propose a Knowledge-based topic model, namely, KB-LDA, which integrates an ontology as a knowledge base into the statistical topic models in a principled way. Our model integrates the topics to external knowledge bases, which can benefit other research areas such as classification, information retrieval, semantic search and visualization.
2) We define a labeling approach for topics considering the semantics of the concepts that are included in the learned topics in addition to existing ontological relationships between the concepts of the ontology. The proposed model enhances the accuracy of the labels by applying the topic-concept associations. Additionally, it automatically generates labels that are descriptive for explaining and understanding the topics.
3) We demonstrate the usefulness of our approach in two ways. Firstly, we demonstrate how our model connects text documents to concepts of the ontology and their categories. Secondly, we show automatic topic labeling by performing a multiples experiments.

The organization of the paper is as follows. In Section 2, we formally define our model for labeling the topics by integrating the ontological concepts with probabilistic topic models. We present our method for concept-based topic labeling in Section 3. In Section 4, we demonstrate the effectiveness of our method on two different datasets. Finally, we present our conclusions and future work in Section 5.

II. Background

In this section, we formally describe some of the related concepts and notations that will be used throughout this paper.

A. Ontologies

Ontologies are fundamental elements of the Semantic Web and could be thought of knowledge representation methods, which are used to specify the knowledge shared among different systems. An ontology is referred to an "explicit specification of a conceptualization". [12]. In other words, an ontology is a structure consisting of a set of concepts and a set of relationships existing among them.

Ontologies have been widely used as the background knowledge (i.e., knowledge bases) in a variety of text mining and knowledge discovery tasks such as text clustering [13], [14], [15], text classification [16], [17], [18], word sense disambiguation [19], [20], [21], and others. See [22] for a comprehensive review of Semantic Web in data mining and knowledge discovery.

Recently, the topic modeling approach has become a popular method for uncovering the hidden themes from data such as text corpora, images, etc. This model has been widely used for various text mining tasks, such as machine translation, word embedding, automatic topic labeling, and many others. In the topic modeling approach, each document is considered as a mixture of topics, where a topic is a probability distribution over words. When the topic distributions of documents are estimated, they can be considered as the high-level semantic themes of the documents.

B. Probabilistic Topic Models

Probabilistic topic models are a set of algorithms that have become a popular method for uncovering the hidden themes from data such as text corpora, images, etc. This model has been extensively used for various text mining tasks, such as machine translation, word embedding, automatic topic labeling, and many others. The key idea behind the topic modeling is to create a probabilistic model for the collection of text documents. In topic models, documents are probability distributions over topics, where a topic is represented as a multinomial distribution over words. The two primary topic models are Probabilistic Latent Semantic Analysis (pLSA) proposed by Hofmann in 1999 [23] and Latent Dirichlet Allocation (LDA) [1]. Since pLSA model does not give any probabilistic model at the document level, generalizing it to model new unseen documents will be difficult. Blei et al. [1] extended pLSA model by adding a prior from Dirichlet distribution on mixture weights of topics for each document. He then named the model Latent Dirichlet Allocation (LDA). In the following section, we illustrate the LDA model.

The latent Dirichlet allocation (LDA) [1] is a probabilistic generative model for uncovering thematic theme, which is called topic, of a collection of documents. The basic assumption in LDA model is that each document is a mixture of different topics and each topic is a multinomial probability distribution over all words in the corpus.

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$ is the corpus and $\mathcal{V} = \{w_1, w_2, \ldots, w_V\}$ is the vocabulary set of the collection. A topic $z_j, 1 \leq j \leq K$ is described as a multinomial probability distribution over the V words, $p(w_i|z_j), \sum_i^V p(w_i|z_j) = 1$. LDA produces the words in a two-step procedure comprising 1) topics generate words; and 2)documents generate topics. In another word, we can calculate the probability of words given the document as:

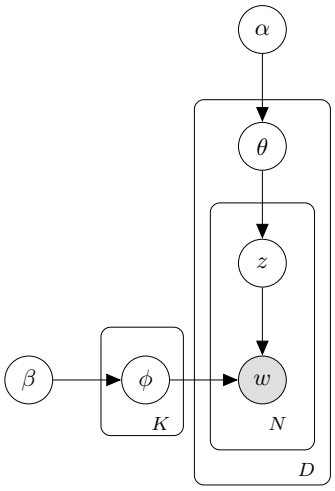


Fig. 1. LDA graphical model.

$$p(w_i|d) = \sum_{j=1}^{K} p(w_i|z_j)p(z_j|d) \quad (1)$$

Fig. 1 shows the graphical model of LDA. The generative process for the document collection $\mathcal{D}$ is as follows:

1) For each topic $k \in \{1, 2, \ldots, K\}$, draw a word distribution $\phi_k \sim \text{Dir}(\beta)$
2) For each document $d \in \{1, 2, \ldots, D\}$,
   (a) draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word w_n, where $n \in \{1, 2, \ldots, N\}$, in document d,
      i. draw a topic $z_i \sim \text{Mult}(\theta_d)$
      ii. draw a word $w_n \sim \text{Mult}(\phi_{z_i})$

The joint distribution of hidden and observed variables in the model is:

$$P(\phi_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{j=1}^{K} P(\phi_j|\beta) \prod_{d=1}^{D} P(\theta_d|\alpha) \left(\prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{1:K}, z_{d,n}) \right) \quad (2)$$

In the LDA model, the word-topic distribution $p(w|z)$ and topic-document distribution $p(z|d)$ are learned entirely in an unsupervised manner, without any prior knowledge about what words are related to the topics and what topics are related to individual documents. One of the most widely-used approximate inference techniques is Gibbs sampling [24]. Gibbs sampling begins with random assignment of words to topics, then the algorithm iterates over all the words in the training documents for a number of iterations (usually on order of 100). In each iteration, it samples a new topic assignment for each word using the conditional distribution of that word given all other current word-topic assignments. After the iterations are finished, the algorithm reaches a steady state, and the word-topic probability distributions can be estimated using word-topic assignments.

III. Motivating Example

Let's presume that we are given a collection of news articles and told to extract the common themes present in this corpus. Manual inspection of the articles is the simplest approach, but it is not practical for large collection of documents. We can make use of topic models to solve this problem by assuming that a collection of text documents comprises of a set of hidden themes, called *topics*. Each topic z is a multinomial distribution $p(w|z)$ over the words w of the vocabulary. Similarly, each document is made up of these topics, which allows multiple topics to be present in the same document. We estimate both the topics and document-topic mixtures from the data simultaneously. After we estimate the distribution of each document over topics, we can use them as the semantic themes of the documents. The top words in each topic-word distribution demonstrates the description of that topic.

For example, Table II shows a sample of four topics with their top-10 words learned from a corpus of news articles. Although the topic-word distributions are usually meaningful,

TABLE II. Example Topics with Top-10 words Learned from a Document Set

Topic 1	Topic 2	Topic 3	Topic 4
company	film	drug	republican
mobile	show	drugs	house
technology	music	cancer	senate
facebook	year	fda	president
google	television	patients	state
apple	singer	reuters	republicans
online	years	disease	political
industry	movie	treatment	campaign
video	band	virus	party
business	actor	health	democratic

it is quite difficult for the users to exactly infer the meanings of the topics just from the top words, particularly when they do not have enough knowledge about the domain of the corpus. Standard LDA model does not *automatically* provide the labels of the topics. Essentially, for each topic it gives a distribution over the entire words of the vocabulary. A *label* is one or a few phrases that adequately describes the meaning of the topic. For instance, As shown in Table II, topics do not have any labels, therefore they must be manually assigned. Topic labeling task can be laborious, specifically when number of topics is substantial. Table III illustrates the same topics that have been labeled (second row in the table) manually by a human.

Automatic topic labeling which aims to to automatically generate interpretable labels for the topics has attracted increasing attention in recent years [25], [26], [27], [28], [29]. Unlike previous works that have essentially concentrated on the topics discovered from LDA topic model and represented the topics by words, we propose an Knowledge-based topic model, KB-LDA, where topics are labeled by ontological concepts.

We believe that the knowledge in the ontology can be integrated with the topic models to automatically generate topic labels that are semantically relevant, understandable for humans and highly cover the discovered topics. In other words, our aim is to use the semantic knowledge graph of concepts in an ontology (e.g., DBpedia) and their diverse relationships

TABLE III. EXAMPLE TOPICS WITH TOP-10 WORDS LEARNED FROM A DOCUMENT SET. THE SECOND ROW PRESENTS THE MANUALLY ASSIGNED LABELS.

Topic 1	Topic 2	Topic 3	Topic 4
"Technology"	"Entertainment"	"Health"	"U.S. Politics"
company	film	drug	republican
mobile	show	drugs	house
technology	music	cancer	senate
facebook	year	fda	president
google	television	patients	state
apple	singer	reuters	republicans
online	years	disease	political
industry	movie	treatment	campaign
video	band	virus	party
business	actor	health	democratic

with unsupervised probabilistic topic models (i.e. LDA), in a principled manner and exploit this information to automatically generate meaningful topic labels.

IV. RELATED WORK

Probabilistic topic modeling has been widely applied to various text mining tasks in virtue of its broad application in applications such as text classification [30], [31], [32], word sense disambiguation [4], [5], sentiment analysis [2], [33], and others. A main challenge in such topic models is to interpret the semantic of each topic in an accurate way.

Early research on topic labeling usually considers the top-n words that are ranked based on their marginal probability $p(w_i|z_j)$ in that topic as the primitive labels [1], [24]. This option is not satisfactory, because it necessitates significant perception to interpret the topic, particularly if the user is not knowledgeable of the topic domain. For example, it would be very hard to infer the meaning of the topic shown in Table I only based on the top terms, if someone is not knowledgeable about the "database" domain. The other conventional approach for topic labeling is to manually generate topic labels [34], [35]. This approach has disadvantages: 1) the labels are prone to subjectivity; and 2) the method can not be scale up, especially when coping with massive number of topics.

Recently, automatic topic labeling has been getting more attention as an area of active research. Wang et al. [25] utilized n-grams to represent topics, so label of the topic was its top n-grams. Mei et al. [26] introduced a method to automatically label the topics by transforming the labeling problem to an optimization problem. First they generate candidate labels by extracting either bigrams or noun chunks from the collection of documents. Then, they rank the candidate labels based on Kullback-Leibler (KL) divergence with a given topic, and choose a candidate label that has the highest mutual information and the lowest KL divergence with the topic to label the corresponding topic. [27] introduced an algorithm for topic labeling based on a given topic hierarchy. Given a topic, they generate label candidate set using Google Directory hierarchy and come with the best matched label according to a set of similarity measures.

Lau et al. [36] introduced a method for topic labeling by selecting the best topic word as its label based on a number of features. They assume that the topic terms are representative enough and appropriate to be considered as labels, which is not always the case. Lau et al. [28] reused the features proposed in [36] and also extended the set of candidate labels exploiting Wikipedia. For each topic they first select the top terms and query the Wikipedia to find top article titles having the these terms according to the features and consider them as extra candidate labels. Then they rank the candidate to find the best label for the topic.

Mao et al. [37] used the sibling and parent-child relations between topics to enhances the topic labeling. They first generate a set of candidate labels by extracting meaningful phrases using Ngram Testing [38] for a topic and adding the top topic terms to the set based on marginal term probabilities. And then rank the candidate labels by exploiting the hierarchical structure between topics and pick the best candidate as the label of the topic.

In a more recent work Hulpus et al. [29] proposed an automatic topic labeling approach by exploiting structured data from DBpedia[2]. Given a topic, they first find the terms with highest marginal probabilities, and then determine a set of DBpedia concepts where each concept represents the identified sense of one of the top terms of the topic. After that, they create a graph out of the concepts and use graph centrality algorithms to identify the most representative concepts for the topic.

The proposed model differs from all prior works as we introduce a topic model that integrates knowledge with data-driven topics within a single general framework. Prior works primarily emphasize on the topics discovered from LDA topic model whereas in our model we introduce another random variable namely *concept* between topics and words. In this case, each document is made up of topics where each topic is defined as a probability distribution over concepts and each concept has a multinomial distribution over vocabulary.

The hierarchical topic models which consider the correlations among topics, are conceptually similar to our KB-LDA model. Mimno et al. [39] proposed the hPAM approach and defined super-topics and sub-topics terms. In their model, a document is considered as a mixture of distributions over super-topics and sub-topics, using a directed acyclic graph to represent a topic hierarchy. Our model, KB-LDA model, is different, because in hPAM, distribution of each super-topic over sub-topics depends on the document, whereas in KB-LDA, distributions of topics over concepts are independent of the corpus and are based on an ontology. The other difference is that sub-topics in the hPAM model are still unigram words, whereas in KB-LDA, ontological concepts are n-grams, which makes them more specific and more representative, a key point in KB-LDA. [40], [41] proposed topic models that integrate concepts with topics. The key idea in their frameworks is that topics of the topic models and ontological concepts both are represented by a set of "focused" words, i.e. distributions over words, and this similarity has been utilized in their models. However, our KB-LDA model is different from these models in that they treat the concepts and topics in the same way, whereas in KB-LDA, topics and concepts make two separate levels in the model.

V. PROBLEM FORMULATION

In this section, we formally describe our model and its learning process. We then explain how to leverage the topic-

[2] http://dbpedia.org

concept distribution to generate meaningful semantic labels for each topic, in Section 4. The notation used in this paper is summarized in Table V.

The intuitive idea behind our model is that using words from the vocabulary of the document corpus to represent topics is not a good way to understand the topics. Words usually demonstrate topics in a broader way in comparison with ontological concepts that can describe the topics in more specific manner. In addition, concepts representations of a topic are closely related and have higher semantic relatedness to each other. For instance, the first column of Table IV shows top words of a topic learned by traditional LDA, whereas the second column represents the same topics through its top ontological concepts learned by the KB-LDA model. We can determine that the topic is about "sports" from the word representation of the topic, but the concept representation of the topic reveals that not only the topic is about "sports", but more precisely about "American sports".

TABLE IV. EXAMPLE OF TOPIC-WORD REPRESENTATION LEARNED BY LDA AND TOPIC-CONCEPT REPRESENTATION LEARNED BY KB-LDA

LDA		**KB-LDA**	
Human Label: Sports		**Human Label:** American Sports	
Topic-word	**Probability**	**Topic-concept**	**Probability**
team	(0.123)	oakland raiders	(0.174)
est	(0.101)	san francisco giants	(0.118)
home	(0.022)	red	(0.087)
league	(0.015)	new jersey devils	(0.074)
games	(0.010)	boston red sox	(0.068)
second	(0.010)	kansas city chiefs	(0.054)

Let $\mathcal{C} = \{c_1, c_2, \ldots, c_C\}$ be the set of concepts from DBpedia, and $\mathcal{D} = \{d_i\}_{i=1}^{D}$ be a text corpus. We describe a document d in the collection $\mathcal{D}$ with a bag of words, i.e., $d = \{w_1, w_2, \ldots, w_V\}$, where V is the size of the vocabulary.

Definition 1. (Concept): A *concept* in a text collection $\mathcal{D}$ is depicted by c and defined as a multinomial probability distribution over the vocabulary $\mathcal{V}$, i.e., $\{p(w|c)\}_{w\in\mathcal{V}}$. Clearly, we have $\sum_{w\in\mathcal{V}} p(w|c) = 1$. We assume that there are $|\mathcal{C}|$ concepts in $\mathcal{D}$ where $\mathcal{C} \subset C$.

Definition 2. (Topic): A *topic* ϕ in a given corpus $\mathcal{D}$ is defined as a multinomial distribution over the *concepts* $\mathcal{C}$, i.e., $\{p(c|\phi)\}_{c\in\mathcal{C}}$. Clearly, we have $\sum_{c\in\mathcal{C}} p(c|\phi) = 1$. We assume that there are K topics in $\mathcal{D}$.

Definition 3. (Topic representation): The *topic representation* of a document d, θ_d, is defined as a probabilistic distribution over K topics, i.e., $\{p(\phi_k|\theta_d)\}_{k\in K}$.

TABLE V. NOTATION USED IN THIS PAPER

Symbol	Description
D	number of documents
K	number of topics
C	number of concepts
V	number of words
N_d	number of words in document d
α_t	asymmetric Dirichlet prior for topic t
β	symmetric Dirichlet prior for topic-concept distribution
γ	symmetric Dirichlet prior for concept-word distribution
z_i	topic assigned to the word at position i in the document d
c_i	concept assigned to the word at position i in the document d
w_i	word at position i in the document d
θ_d	multinomial distribution of topics for document d
ϕ_k	multinomial distribution of concepts for topic k
ζ_c	multinomial distribution of words for concept c

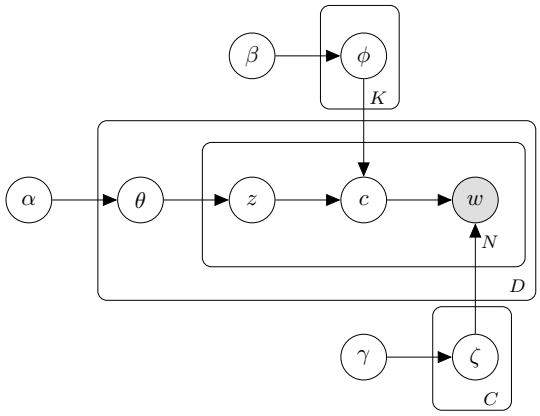


Fig. 2. Graphical representation of KB-LDA model.

Definition 4. (Topic Modeling): Given a collection of text documents, $\mathcal{D}$, the task of *Topic Modeling* aims at discovering and extracting K topics, i.e., $\{\phi_1, \phi_2, \ldots, \phi_K\}$, where the number of topics, K, is specified by the user.

A. The KB-LDA Topic Model

The KB-LDA topic model is based on combining topic models with ontological concepts in a single framework. In this case, topics and concepts are distributions over concepts and words in the corpus, respectively.

The KB-LDA topic model is shown in Fig. 2 and the generative process of the approach is defined as Algorithm 1.

Algorithm 1: KB-LDA Topic Model

```
foreach concept c ∈ {1, 2, ..., C} do
    Sample a word distribution ζ_c ~ Dir(γ)
end
foreach topic k ∈ {1, 2, ..., K} do
    Sample a concept distribution φ_k ~ Dir(β)
end
foreach document d ∈ {1, 2, ..., D} do
    Sample a topic distribution θ_d ~ Dir(α)
    foreach word w of document d do
        Sample a topic z ~ Mult(θ_d)
        Sample a concept c ~ Mult(φ_z)
        Sample a word w from concept c, w ~
          Mult(ζ_c)
    end
end
```

Following this process, the joint probability of generating a corpus $D = \{d_1, d_2, \ldots, d_{|D|}\}$, the topic assignments $\mathbf{z}$ and the concept assignments $\mathbf{c}$ given the hyperparameters α, β and γ is:

$$
\begin{aligned}
&P(\mathbf{w}, \mathbf{c}, \mathbf{z}|\alpha, \beta, \gamma) \\
&= \int_{\zeta} P(\zeta|\gamma) \prod_{d} \sum_{c_d} P(w_d|c_d, \zeta) \\
&\times \int_{\phi} P(\phi|\beta) \int_{\theta} P(\theta|\alpha) P(c_d|\theta, \phi) d\theta d\phi d\zeta
\end{aligned} \tag{3}
$$

B. Inference using Gibbs Sampling

Since the posterior inference of the KB-LDA is intractable, we require an algorithm to estimate the posterior inference of the model. There are different algorithms have been applied to estimate the topic models parameters, such as variational EM [1] and Gibbs sampling [24]. In the current study, we will use collapsed Gibbs sampling procedure for KB-LDA topic model. Collapsed Gibbs sampling [24] is based on Markov Chain Monte Carlo (MCMC) [42] algorithm which builds a Markov chain over the latent variables in the model and converges to the posterior distribution after a number of iterations. In this paper, our goal is to construct a Markov chain that converges to the posterior distribution over $\mathbf{z}$ and $\mathbf{c}$ conditioned on observed words $\mathbf{w}$ and hyperparameters α, β and γ. We use a blocked Gibbs sampling to jointly sample $\mathbf{z}$ and $\mathbf{c}$, although we can alternatively perform hierarchical sampling, i.e., first sample $\mathbf{z}$ and then sample $\mathbf{c}$. Nonetheless, Rosen-Zvi [43] argue that in cases where latent variables are greatly related, blocked sampling boosts convergence of the Markov chain and decreases auto-correlation, as well.

The posterior inference is derived from (3) as follows:

$$
\begin{aligned}
P(\mathbf{z}, \mathbf{c}|\mathbf{w}, \alpha, \beta, \gamma) &= \frac{P(\mathbf{z}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \gamma)}{P(\mathbf{w}|\alpha, \beta, \gamma)} \\
&\propto P(\mathbf{z}, \mathbf{c}, \mathbf{w}|\alpha, \beta, \gamma) \\
&= P(\mathbf{z}) P(\mathbf{c}|\mathbf{z}) P(\mathbf{w}|\mathbf{c})
\end{aligned} \tag{4}
$$

where

$$
P(\mathbf{z}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\right)^D \prod_{d=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_k^{(d)} + \alpha)}{\Gamma(\sum_{k'}(n_{k'}^{(d)} + \alpha))} \tag{5}
$$

$$
P(\mathbf{c}|\mathbf{z}) = \left(\frac{\Gamma(C\beta)}{\Gamma(\beta)^C}\right)^K \prod_{k=1}^{K} \frac{\prod_{c=1}^{C} \Gamma(n_c^{(k)} + \beta)}{\Gamma(\sum_{c'}(n_{c'}^{(k)} + \beta))} \tag{6}
$$

$$
P(\mathbf{w}|\mathbf{c}) = \left(\frac{\Gamma(V\zeta)}{\Gamma(\zeta)^V}\right)^C \prod_{c=1}^{C} \frac{\prod_{w=1}^{V} \Gamma(n_w^{(c)} + \zeta)}{\Gamma(\sum_{w'}(n_{w'}^{(c)} + \zeta))} \tag{7}
$$

where $P(\mathbf{z})$ is the probability of the joint topic assignments $\mathbf{z}$ to all the words $\mathbf{w}$ in corpus $\mathcal{D}$. $P(\mathbf{c}|\mathbf{z})$ is the conditional probability of joint concept assignments $\mathbf{c}$ to all the words $\mathbf{w}$ in corpus $\mathcal{D}$, given all topic assignments $\mathbf{z}$, and $P(\mathbf{w}|\mathbf{c})$ is the conditional probability of all the words $\mathbf{w}$ in corpus $\mathcal{D}$, given all concept assignments $\mathbf{c}$.

For a word token w at position i, its full conditional distribution can be written as:

$$
\begin{aligned}
&P(z_i = k, c_i = c|w_i = w, \mathbf{z}_{-i}, \mathbf{c}_{-i}, \mathbf{w}_{-i}, \alpha, \beta, \gamma) \propto \\
&\frac{n_{k,-i}^{(d)} + \alpha_k}{\sum_{k'}(n_{k',-i}^{(d)} + \alpha_{k'})} \times \frac{n_{c,-i}^{(k)} + \beta}{\sum_{c'}(n_{c',-i}^{(k)} + \beta)} \times \\
&\frac{n_{w,-i}^{(c)} + \gamma}{\sum_{w'}(n_{w',-i}^{(c)} + \gamma)}
\end{aligned} \tag{8}
$$

where $n_w^{(c)}$ is the number of times word w is assigned to concept c. $n_c^{(k)}$ is the number of times concept c occurs under topic k. $n_k^{(d)}$ denotes the number of times topic k is associated with document d. Subscript $-i$ indicates the contribution of the current word w_i being sampled is removed from the counts.

In most probabilistic topic models, the Dirichlet parameters α are assumed to be given and fixed, which still produce reasonable results. But, as described in [44], that asymmetric Dirichlet prior α has substantial advantages over a symmetric prior, we have to learn these parameters in our proposed model. We could use maximum likelihood or maximum a posteriori estimation to learn α. However, there is no closed-form solution for these methods and for the sake of simplicity and speed we use moment matching methods [45] to approximate the parameters of α. In each iteration of Gibbs sampling, we update

$$
\begin{aligned}
mean_{dk} &= \frac{1}{N} \times \sum_{d} \frac{n_k^{(d)}}{n^{(d)}} \\
var_{dk} &= \frac{1}{N} \times \sum_{d} (\frac{n_k^{(d)}}{n^{(d)}} - mean_{dk})^2 \\
m_{dk} &= \frac{mean_{dk} \times (1 - mean_{dk})}{var_{dk}} - 1 \\
\alpha_{dk} &\propto mean_{dk} \\
\sum_{k=1}^{K} \alpha_{dk} &= exp(\frac{\sum_{k=1}^{K} log(m_{dk})}{K - 1})
\end{aligned} \tag{9}
$$

For each document d and topic k, we first compute the sample mean $mean_{dk}$ and sample variance var_{dk}. N is the number of documents and $n^{(d)}$ is the number of words in document d.

Algorithm 2 shows the Gibbs sampling process for our KB-LDA model.

After Gibbs sampling, we can use the sampled topics and concepts to estimate the probability of a topic given a document, θ_{dk}, probability of a concept given a topic, ϕ_{kc}, and the probability of a word given a concept, ζ_{cw}:

$$
\theta_{dk} = \frac{n_k^{(d)} + \alpha_k}{\sum_{k'}(n_{k'}^{(d)} + \alpha_{k'})} \tag{10}
$$

$$
\phi_{kc} = \frac{n_c^{(k)} + \beta}{\sum_{c'}(n_{c'}^{(k)} + \beta)} \tag{11}
$$

$$
\zeta_{cw} = \frac{n_w^{(c)} + \gamma}{\sum_{w'}(n_{w'}^{(c)} + \gamma)} \tag{12}
$$

Algorithm 2: KB-LDA Gibbs Sampling

Input : A collection of documents D, number of topics K and α, β, γ
Output: $\zeta = \{p(w_i|c_j)\}$, $\phi = \{p(c_j|z_k)\}$ and $\theta = \{p(z_k|d)\}$, i.e. concept-word, topic-concept and document-topic distributions

```
/* Randomly, initialize concept-word assignments for all word tokens, topic-concept assignments for all
   concepts and document-topic assignments for all the documents                                        */
initialize the parameters φ, θ and ζ randomly;
if computing parameter estimation then
    initialize alpha parameters, α, using Eq. 9;
end
t ← 0;
while t < MaxIteration do
    foreach word w do
        c = c(w) // get the current concept assignment
        k = z(w) // get the current topic assignment
        // Exclude the contribution of the current word w
        n_w^(c) ← n_w^(c) − 1;
        n_c^(k) ← n_c^(k) − 1;
        n_k^(d) ← n_k^(d) − 1 // w is a document word
        (newk, newc) = sample new topic-concept and concept-word for word w using Eq. 8;
        // Increment the count matrices
        n_w^(newc) ← n_w^(newc) + 1;
        n_newc^(newk) ← n_newc^(newk) + 1;
        n_newk^(d) ← n_newk^(d) + 1;
        // Update the concept assignments and topic assignment vectors
        c(w) = newc;
        z(w) = newk;
        if computing parameter estimation then
            update alpha parameters, α, using Eq. 9;
        end
    end
    t ← t + 1;
end
```

VI. Concept-based Topic Labeling

The key idea behind our model is that entities that are included in the text document and their inter-connections can specify the topic(s) of the document. Additionally, the entities of the ontology that are categorized into the same or similar classes have higher semantic relatedness to each other. Therefore, in order to recognize good topics labels, we count on the semantic similarity between the entities included in the text document and a suitable portion of the ontology. Research presented in [16] use a similar approach to perform Knowledge-based text categorization.

Definition 5. (Topic Label): A *topic label* ℓ for topic ϕ is a sequence of words which is semantically meaningful and sufficiently explains the meaning of ϕ.

KB-LDA highlights the concepts of the ontology and their classification hierarchy as labels for topics. To find representative labels that are semantically relevant for a discovered topic ϕ, KB-LDA involves four major steps: 1) constructs the semantic graph from top concepts from topic-concept distribution for the given topic; 2) selects and analyzes the thematic graph, a semantic graph's subgraph; 3) extracts the topic graph from the thematic graph concepts; and 4) computes the semantic similarity between topic ϕ and the candidate labels of the topic label graph.

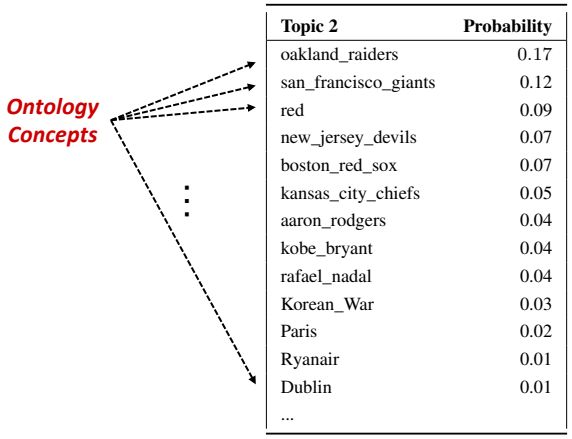
A. Semantic Graph Construction

In the proposed model, we compute the marginal probabilities $p(c_i|\phi_j)$ of each concept c_i in a given topic ϕ_j. We then, and select the $\mathcal{K}$ concepts having the highest marginal probability in order to create the topic's semantic graph. Fig. 3 illustrates the top-10 concepts of a topic learned by KB-LDA.

Definition 6. (Semantic Graph): A *semantic graph* of a topic ϕ is a labeled graph $G^{\phi} = \langle V^{\phi}, E^{\phi} \rangle$, where V^{ϕ} is a set of labeled vertices, which are the top concepts of ϕ (their labels are the concept labels from the ontology) and E^{ϕ} is a set of edges $\{\langle v_i, v_j \rangle$ with label r, such that $v_i, v_j \in V^{\phi}$ and v_i and v_j are connected by a relationship r in the ontology$\}$.

For instance, Fig. 4 shows the semantic graph of the example topic ϕ in Fig. 3, which consists of three sub-graphs (connected components).

Even though the ontology relationships are directed in G^{ϕ}, in this paper, we will consider the G^{ϕ} as an undirected graph.



Topic 2	Probability
oakland_raiders	0.17
san_francisco_giants	0.12
red	0.09
new_jersey_devils	0.07
boston_red_sox	0.07
kansas_city_chiefs	0.05
aaron_rodgers	0.04
kobe_bryant	0.04
rafael_nadal	0.04
Korean_War	0.03
Paris	0.02
Ryanair	0.01
Dublin	0.01
...	

Fig. 3. Example of a topic represented by top concepts learned by KB-LDA.

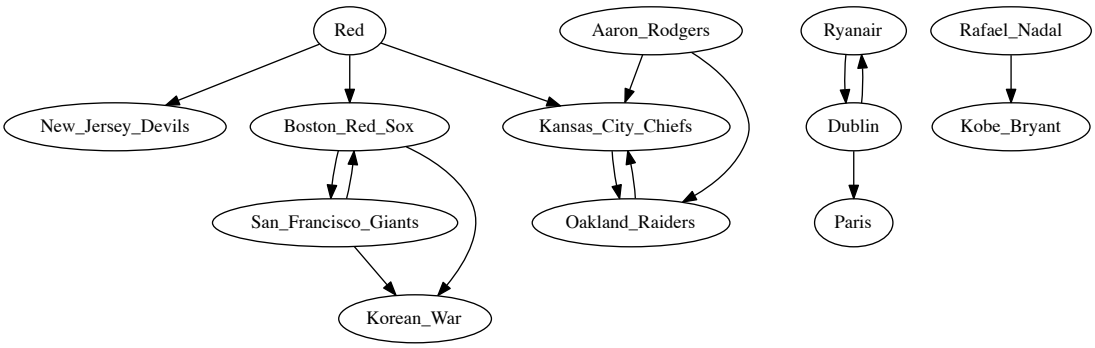


Fig. 4. Semantic graph of the example topic ϕ described in Fig. 3 with $|V^{\phi}| = 13$.

B. Thematic Graph Selection

In our model, we select the thematic graph assuming that concepts under a given topic are semantically closely related in the ontology, whereas concepts from varying topics are located far away, or even not connected at all. We need to consider that there is a chance of generating incoherent topics. In other words, for a given topic that is represented as a list of $\mathcal{K}$ concepts with highest probabilities, there may be a few concepts, which are not semantically close to other concepts and to the topic. It consequently can result in generating the topic's semantic graph that may comprise multiple connected components.

Definition 7. (Thematic graph): A *thematic graph* is a connected component of G^{ϕ}. Particularly, if the entire G^{ϕ} is a connected graph, it is also a thematic graph.

Definition 8. (Dominant Thematic Graph): A thematic graph with the largest number of nodes is called the *dominant thematic graph* for topic ϕ.

Fig. 5 depicts the dominant thematic graph for the example topic ϕ along with the initial weights of nodes, $p(c_i|\phi)$.

C. Topic Label Graph Extraction

The idea behind a topic label graph extraction is to find ontology concepts as candidate labels for the topic.

The importance of concepts in a thematic graph is based on their initial weights, which are the marginal probabilities of concepts under the topic, and their relative positions in the graph. Here, we apply Hyperlink-Induced Topic Search algorithm, HITS algorithm, [46] with the assigned initial weights for concepts to find the *authoritative concepts* in the dominant thematic graph. Ultimately, we determine the *central concepts* in the graph based on the geographical centrality measure, since these nodes can be recognized as the thematic landmarks of the graph.

Definition 9. (Core Concepts): The set of the the most authoritative and central concepts in the dominant thematic graph forms the *core concepts* of the topic ϕ and is denoted by CC^{ϕ}.

The top-4 core concept nodes of the dominant thematic graph of example topic ϕ are highlighted in Fig. 6. It should be noted that "Boston_Red_Sox" has not been selected as a core concept, because it's score is lower than that of the concept "Red" based on the HITS and centrality computations ("Red"

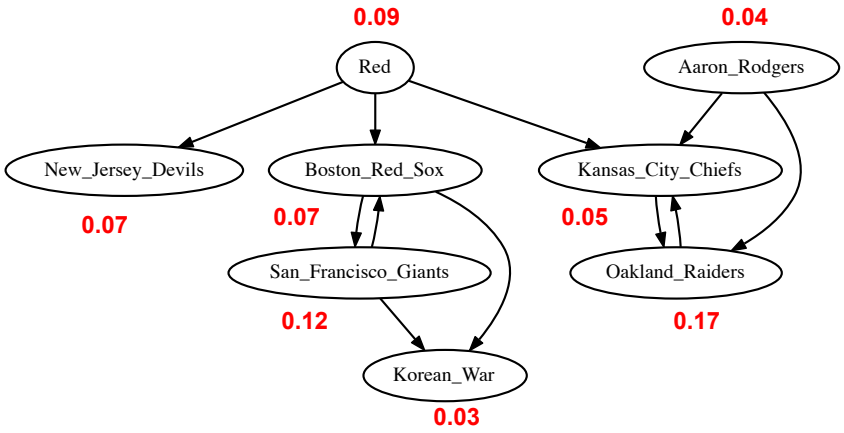


Fig. 5. Dominant thematic graph of the example topic described in Fig. 4.

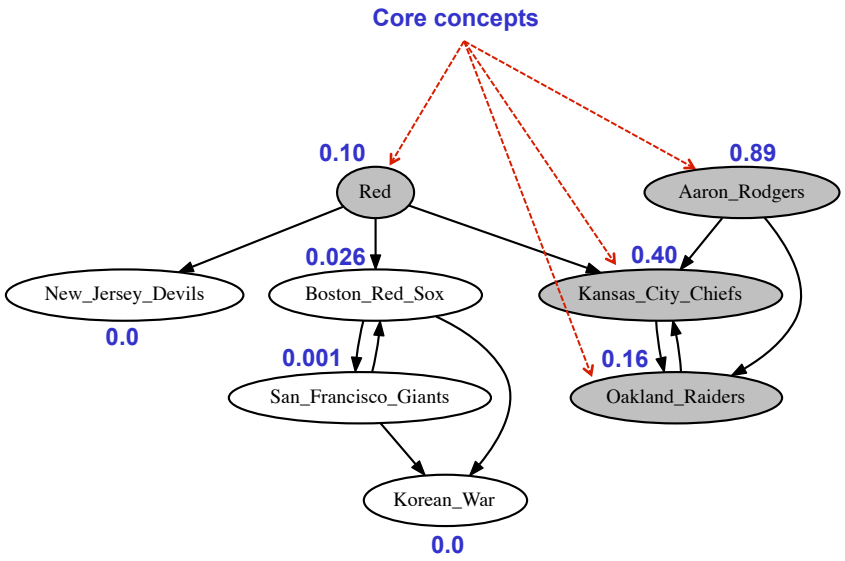


Fig. 6. Core concepts of the Dominant thematic graph of the example topic described in Fig. 5.

has far more relationships to other concepts in DBpedia).

From now on, we refer the dominant thematic graph of a topic as the thematic graph.

To exploit the topic label graph for the core concepts CC^{ϕ}, we primarily consider on the ontology class hierarchy (structure), since we can concentrate the topic labeling as assigning class labels to topics. We present definitions similar to those in [29] for representing the label graph and topic label graph.

Definition 10. (Label Graph): The *label graph* of a concept c_i is an undirected graph $G_i = \langle V_i, E_i \rangle$, where V_i is the union of $\{c_i\}$ and a subset of ontology classes (c_i's types and their ancestors) and E_i is a set of edges labeled by *rdf:type* and *rdfs:subClassOf* and connecting the nodes. Each node in the label graph excluding c_i is regarded as a *label* for c_i.

Definition 11. (Topic Label Graph): Let $CC^{\phi} = \{c_1, c_2, \ldots, c_m\}$ be the core concept set. For each concept $c_i \in CC^{\phi}$, we extract its *label graph*, $G_i = \langle V_i, E_i \rangle$, by traversing the ontology from c_i and retrieving all the nodes laying at most three hops away from C_i. The *union* of these

graphs $\boldsymbol{G}_{cc^{\phi}} = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$ where $\mathbf{V} = \bigcup V_i$ and $\mathbf{E} = \bigcup E_i$ is called the *topic label graph*.

It should be noted that we empirically restrict the ancestors to three levels, because expanding the distance causes undesirable general classes to be included in the graph.

D. Semantic Relevance Scoring Function

In this section, we introduce a semantic relevance scoring function to rank the candidate labels by measuring their semantic similarity to a topic.

Mei et al. [26] consider two parameters to interpret the semantics of a topic, including: 1) distribution of the topic; and 2) the context of the topic. Proposed topic label graph for a topic ϕ is exploited, utilizing the distribution of the topic over the set of concepts plus the context of the topic in the form of semantic relatedness between the concepts in the ontology.

To determine the semantic similarity of a label ℓ in $\mathbf{G}_{cc^{\phi}}$ to a topic ϕ, the semantic similarity between ℓ and all of the concepts in the core concept set CC^{ϕ} is computed and then ranked the labels and finally, the best representative labels for the topic is selected.

Scoring a candidate label is based on three primary goals: 1) the label should have enough coverage *important concepts* of the topic (concepts with higher marginal probabilities); 2) the generated label should be more specific to the core concepts (lower in the class hierarchy); and ultimately, 3) the label should cover the highest number of core concepts in $\mathbf{G}_{cc^{\phi}}$.

In order to calculate the semantic similarity of a label to a concept, the fist step is calculating the *membership score* and the *coverage score*. The modified Vector-based Vector Generation method (VVG) described in [47] is selected to compute the membership score of a concept to a label.

In the experiments, we used DBpedia, an ontology created out of Wikipedia knowledge base. All concepts in DBpedia are classified into DBpedia categories and categories are inter-related via subcategory relationships, including *skos:broader*, *skos:broaderOf*, *rdfs:subClassOf*, *rdfs:type* and *dcterms:subject*. We rely on these relationships for the construction of the label graph. Given the topic label graph $\mathbf{G}_{cc^{\phi}}$ we compute the similarity of the label ℓ to the core concepts of topic ϕ as follows.

If a concept c_i has been classified to N DBpedia categories, or similarly, if a category C_j has N parent categories, we set the weight of each of the membership (classification) relationships e to:

$$m(e) = \frac{1}{N} \tag{13}$$

The *membership score*, $mScore(c_i, C_j)$, of a concept c_i to a category C_j is defined as follows:

$$mScore(c_i, C_j) = \prod_{e_k \in E_l} m(e_k) \tag{14}$$

where, $E_l = \{e_1, e_2, \ldots, e_m\}$ represents the set of all membership relationships forming the shortest path p from concept c_i to category C_j. Fig. 7 illustrates a fragment of the label graph for the concept *"Oakland_Raiders"* and shows how its membership score to the category *"American_Football_League_teams"* is computed.

The *coverage score*, $cScore(c_i, C_j)$, of a concept c_i to a category C_j is defined as follows:

$$cScore(w_i, v_j) = \begin{cases} \frac{1}{d(c_i, C_j)} & \text{if there is a path from } c_i \text{ to } C_j \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The *semantic similarity* between a concept c_i and label ℓ in the topic label graph $\mathbf{G}_{cc^{\phi}}$ is defined as follows:

$$SSim(c_i, \ell) = w(c_i) \times \Big(\lambda \cdot mScore(c_i, \ell) + (1 - \lambda) \cdot cScore(c_i, \ell)\Big) \tag{16}$$

where, $w(c_i)$ is the weight of the c_i in $\mathbf{G}_{cc^{\phi}}$, which is the marginal probability of concept c_i under topic $\phi, w(c_i) = p(c_i|\phi)$. Similarly, the semantic similarity between a set of core concept CC^{ϕ} and a label ℓ in the topic label graph $\mathbf{G}_{cc^{\phi}}$ is defined as:

$$\begin{aligned} SSim(CC^{\phi}, \ell) = \frac{\lambda}{|CC^{\phi}|} \sum_{i=1}^{|CC^{\phi}|} w(c_i) \cdot mScore(c_i, \ell) \\ + (1 - \lambda) \sum_{i=1}^{|CC^{\phi}|} w(c_i) \cdot cScore(c_i, \ell) \end{aligned} \tag{17}$$

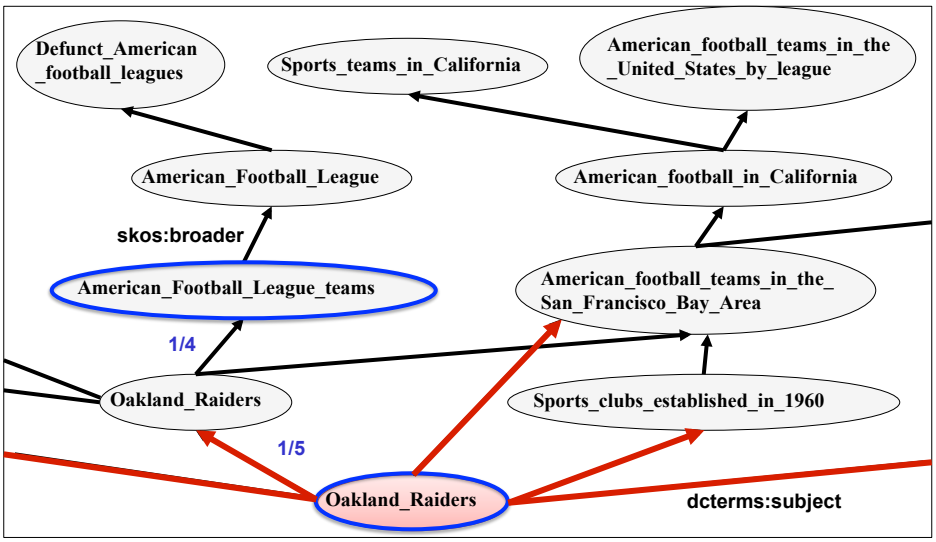
where, λ is the smoothing factor to control the influence of the two scores. We used $\lambda = 0.8$ in our experiments. It should be noted that $SSim(CC^{\phi}, \ell)$ score is not normalized and needs to be normalized. The scoring function aims to satisfy the three criteria by using concept *weight*, *mScore* and *cScore* for first, second and third objectives respectively. This scoring function works based on coverage of topical concepts. It ranks a label node higher, if the label covers more important topical concepts, It means that closing to the core concepts or covering more core concepts are the key points in this scenario. Top-ranked labels are selected as the labels for the given topic. Table VI shows a topic with the top-10 generated labels using our Knowledge-based framework.

VII. Experiments

In order to evaluate the proposed model, KB-LDA, we checked the effectiveness of the model against the one of the state-of-the-art text-based techniques mentioned in [26]. In this paper we call their model Mei07.

In our experiment we choose the DBpedia ontology and two text corpora including a subset of the Reuters[3] news

[3] http://www.reuters.com/



$$mScore(Oakland\_Raider, American\_Football\_League\_teams) = \frac{1}{5} \times \frac{1}{4} = 0.05$$

Fig. 7. Label graph of the concept *"Oakland_Raiders"* along with its *mScore* to the category *"American_Football_League_teams"*.

TABLE VI. EXAMPLE OF A TOPIC WITH TOP-10 CONCEPTS (FIRST COLUMN) AND TOP-10 LABELS (SECOND COLUMN) GENERATED BY OUR PROPOSED METHOD

Topic 2	Top Labels
oakland_raiders	National_Football_League_teams
san_francisco_giants	American_Football_League_teams
red	American_football_teams_in_the_San_Francisco_Bay_Area
new_jersey_devils	Sports_clubs_established_in_1960
boston_red_sox	National_Football_League_teams_in_Los_Angeles
kansas_city_chiefs	American_Football_League
nigeria	American_football_teams_in_the_United_States_by_league
aaron_rodgers	National_Football_League
kobe_bryant	Green_Bay_Packers
rafael_nadal	California_Golden_Bears_football

articles and the British Academic Written English Corpus (BAWE) [48]. More details about the datasets are available in [11]. At the fist step, we extracted the top-2000 bigrams by applying the N-gram Statistics Package [49]. Then, we checked the significance of the bigrams performing the Student's T-Test technique, and exploited the top 1000 ranked candidate bigrams $\mathcal{L}$. In the next step, we calculated the score s for each generated label $\ell \in \mathcal{L}$ and topic ϕ. The score s is defined as follows:

$$s(\ell, \phi) = \sum_{w} \Big(p(w|\phi) PMI(w, \ell|D) \Big) \tag{18}$$

where, PMI is defined as point-wise mutual information between the topic words w and the label ℓ, given the document corpus D. The top-6 labels as the representative labels of the topic ϕ produced by the Mei07 technique were also chosen.

A. Experimental Setup

The experiment setup including pre-processing and the processing parameters presented in details in [11].

B. Results

Tables VII and VIII shows sample results of our method, KB-LDA, along with the generated labels by the Mei07 approach as well as the top-10 words for each topic. We compared the top words and the top-6 labels for each topic and illustrated them in the respective tables. The tables confirm our believe that the labels produced by KB-LDA are more representative than the corresponding labels generated by the Mei07 method. In regards to quantitative evaluation for two aforementioned methods three human experts are asked to compare the generated labels and choose between "Good" and "Unrelated" for each one.

We compared the two different methods using the *Precision@k*, by considering the top-1 to top-6 generated labels. The Precision factor for a topic at top-k is represented as follows:

$$Precision@k = \frac{\text{\# of "Good" labels with rank} \leq k}{k} \tag{19}$$

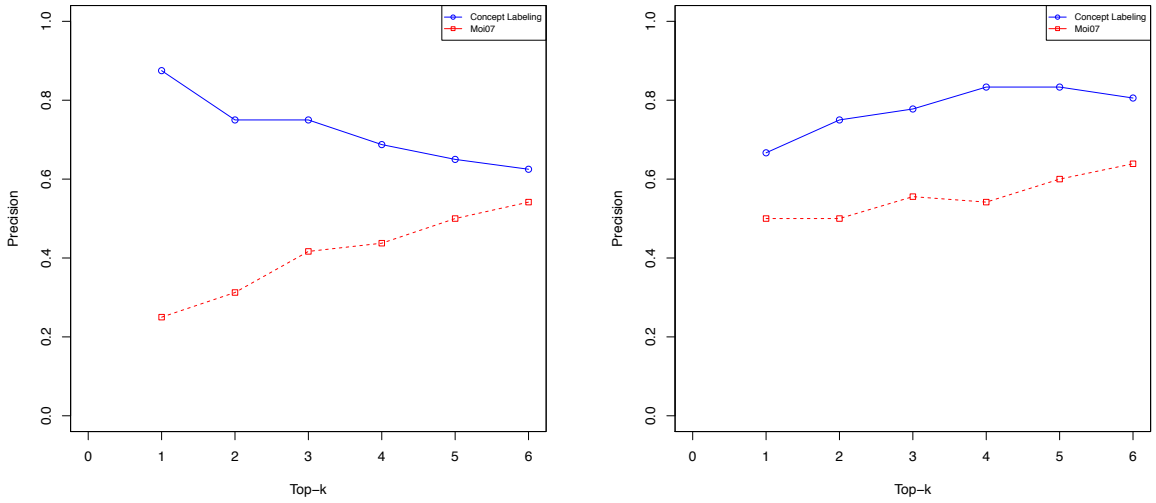
Fig. 8 illustrates the averaged the precision over all the topics for each individual corpus.

TABLE VII. SAMPLE TOPICS OF THE BAWE CORPUS WITH TOP-6 GENERATED LABELS FOR THE MEI METHOD AND KB-LDA + CONCEPT LABELING, ALONG WITH TOP-10 WORDS

Mei07				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
rice production	cell lineage	nuclear dna	disabled people	mg od
southeast asia	cell interactions	eukaryotic organelles	health inequalities	red cells
rice fields	somatic blastomeres	hydrogen hypothesis	social classes	heading mr
crop residues	cell stage	qo site	lower social	colorectal carcinoma
weed species	maternal effect	iron sulphur	black report	cyanosis oedema
weed control	germline blastomeres	sulphur protein	health exclusion	jaundice anaemia
KB-LDA + Concept Labeling				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
agriculture	structural proteins	bacteriology	gender	aging-associated diseases
tropical agriculture	autoantigens	bacteria	biology	smoking
horticulture and gardening	cytoskeleton	prokaryotes	sex	chronic lower respiratory
model organisms	epigenetics	gut flora	sociology and society	inflammations
rice	genetic mapping	digestive system	identity	human behavior
agricultur in the united kingdom	teratogens	firmicutes	sexuality	arthritis
Topic top-10 words				
Topic 1	**Topic 3**	**Topic 12**	**Topic 9**	**Topic 6**
soil	cell	bacteria	health	history
water	cells	cell	care	blood
crop	protein	cells	social	disease
organic	dna	bacterial	professionals	examination
land	gene	immune	life	pain
plant	acid	organisms	mental	medical
control	proteins	growth	medical	care
environmental	amino	host	family	heart
production	binding	virus	children	physical
management	membrane	number	individual	information

TABLE VIII. SAMPLE TOPICS OF THE REUTERS CORPUS WITH TOP-6 GENERATED LABELS FOR THE MEI METHOD AND KB-LDA + CONCEPT LABELING, ALONG WITH TOP-10 WORDS

Mei07				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
hockey league	mobile devices	upgraded falcon	investment bank	russel said
western conference	ralph lauren	commercial communications	royal bank	territorial claims
national hockey	gerry shih	falcon rocket	america corp	south china
stokes editing	huffington post	communications satellites	big banks	milk powder
field goal	analysts average	cargo runs	biggest bank	china sea
seconds left	olivia oran	earth spacex	hedge funds	east china
KB-LDA + Concept Labeling				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
national football league teams	investment banks	space agencies	investment banking	island countries
washington redskins	house of morgan	space organizations	great recession	liberal democracies
sports clubs established in 1932	mortgage lenders	european space agency	criminal investigation	countries bordering the philippine sea
american football teams in maryland	jpmorgan chase	science and technology in europe	madoff investment scandal	east asian countries
american football teams in virginia	banks established in 2000	organizations based in paris	corporate scandals	countries bordering the pacific ocean
american football teams in washington d.c.	banks based in new york city	nasa	taxation	countries bordering the south china sea
Topic top-10 words				
Topic 20	**Topic 1**	**Topic 18**	**Topic 19**	**Topic 3**
league	company	space	bank	china
team	stock	station	financial	chinese
game	buzz	nasa	reuters	beijing
season	research	earth	stock	japan
football	profile	launch	fund	states
national	chief	florida	capital	south
york	executive	mission	research	asia
games	quote	flight	exchange	united
los	million	solar	banks	korea
angeles	corp	cape	group	japanese



(a) Precision for Reuters Corpus

(b) Precision for BAWE Corpus

Fig. 8. Comparison of the systems using human evaluation.

TABLE IX. EXAMPLE TOPICS FROM THE TWO DOCUMENT SETS (TOP-10 WORDS ARE SHOWN). THE THIRD ROW PRESENTS THE MANUALLY ASSIGNED LABELS

BAWE Corpus						**Reuters Corpus**			
Topic 1		**Topic 2**		**Topic 3**		**Topic 7**		**Topic 8**	
AGRICULTURE		MEDICINE		GENE EXPRESSION		SPORTS-FOOTBALL		FINANCIAL COMPANIES	
LDA	**KB-LDA**	**LDA**	**KB-LDA**	**LDA**	**KB-LDA**	**LDA**	**KB-LDA**	**LDA**	**KB-LDA**
soil	soil	*list*	history	cell	cell	game	league	company	company
control	water	history	blood	cells	cells	team	team	million	stock
organic	crop	patient	disease	*heading*	protein	season	game	billion	buzz
crop	organic	pain	examination	*expression*	dna	players	season	business	research
heading	land	examination	pain	*al*	gene	left	football	executive	profile
production	plant	diagnosis	medical	*figure*	acid	time	national	revenue	chief
crops	control	*mr*	care	protein	proteins	games	york	shares	executive
system	environmental	*mg*	heart	genes	amino	*sunday*	games	companies	quote
water	production	problem	physical	gene	binding	football	los	chief	million
biological	management	disease	treatment	*par*	membrane	*pm*	angeles	customers	corp

By considering the results in Fig. 8, two interesting observations are revealed including: 1) in Fig. 8a for up to top-3 labels, the precision difference between the two methods demonstrates the effectiveness of our method, KB-LDA; and 2) the BAWE corpus shows the higher average precision than the Reuters corpus. More explanations are available in [11].

Topic Coherence. In our model, KB-LDA, the topics are defined over concepts. Therefore, to calculate the word distribution for each topic t under KB-LDA, we can apply the following equation:

$$\vartheta_t(w) = \sum_{c=1}^{\mathcal{C}} \Big(\zeta_c(w) \cdot \phi_t(c)\Big) \tag{20}$$

Table IX illustrates the top words from LDA and KB-LDA approaches respectively along with three generated topics from the BAWE corpus.

As Table IX demonstrates that the ***topic coherence*** under KB-LDA is qualitatively better than LDA. The wrong topical words for each topic in Table IX are marked in red and also italicized.

We also calculate the *coherence score* in order to have a quantitative comparison of the coherence of the topics generated by KB-LDA and LDA based on the equation defined in [50]. Given a topic ϕ and its top T words $V^{(\phi)} = (v_1^{(\phi)}, \cdots, v_T^{(\phi)})$ ordered by $P(w|\phi)$, the coherence score is represented as:

$$C(\phi; V^{(\phi)}) = \sum_{t=2}^{T} \sum_{l=1}^{t-1} \log \frac{D(v_t^{(\phi)}, v_l^{(\phi)}) + 1}{D(v_l^{(\phi)})} \tag{21}$$

where, $D(v)$ is the document frequency of word v and $D(v, v')$ is the number of documents in which words v and v' co-occurred. Higher coherence scores shows the higher quality

TABLE X. EXAMPLE TOPICS WITH TOP-10 CONCEPT DISTRIBUTIONS IN KB-LDA MODEL

Topic 1		Topic 2		Topic 3	
rice	0.106	hypertension	0.063	actin	0.141
agriculture	0.095	epilepsy	0.053	epigenetics	0.082
commercial agriculture	0.067	chronic bronchitis	0.051	mitochondrion	0.067
sea	0.061	stroke	0.049	breast cancer	0.066
sustainable living	0.047	breastfeeding	0.047	apoptosis	0.057
agriculture in the united kingdom	0.039	prostate cancer	0.047	ecology	0.042
fungus	0.037	consciousness	0.047	urban planning	0.040
egypt	0.037	childbirth	0.042	abiogenesis	0.039
novel	0.034	right heart	0.024	biodiversity	0.037
diabetes management	0.033	rheumatoid arthritis	0.023	industrial revolution	0.036

TABLE XI. TOPIC COHERENCE ON TOP T WORDS. A HIGHER COHERENCE SCORE MEANS THE TOPICS ARE MORE COHERENT

	BAWE Corpus			Reuters Corpus		
T	**5**	**10**	**15**	**5**	**10**	**15**
LDA	-223.86	-1060.90	-2577.30	-270.48	-1372.80	-3426.60
KB-LDA	$\mathbf{-193.41}$	$\mathbf{-926.13}$	$\mathbf{-2474.70}$	$\mathbf{-206.14}$	$\mathbf{-1256.00}$	$\mathbf{-3213.00}$

of topics. The coherence scores of two methods on different datasets are illustrated in Table XI.

As we mentioned before, KB-LDA defines each topic as a distribution over concepts. Table X illustrates the top-10 concepts with higher probabilities in the topic distribution under the KB-LDA approach for the same three topics, i.e."topic 1", "topic2", and "topic3" of Table IX.

VIII. CONCLUSIONS

In this paper, we presented a topic labeling approach, KB-LDA, based on Knowledge-based topic model and graph-based topic labeling method. The results confirm the robustness and effectiveness of KB-LDA technique on different datasets of text collections. Integrating ontological concepts into our model is a key point that improves the topic coherence in comparison to the standard LDA model.

In regards to the future work, defining a global optimization scoring function for the labels instead of (17) is a potential candidate for future extensions. Moreover, how to integrate *lateral* relationships between the ontology concepts with the topic models as well as the hierarchical relations are also other interesting directions to extend the proposed model.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] A. Lazaridou, I. Titov, and C. Sporleder, "A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations." in *ACL (1)*, 2013, pp. 1630–1639.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *ArXiv e-prints*, 2017.

[4] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, no. 3, pp. 423–469, 2014.

[5] J. L. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation." in *EMNLP-CoNLL*. Citeseer, 2007, pp. 1024–1033.

[6] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 178–185.

[7] E. D. Trippe, J. B. Aguilar, Y. H. Yan, M. V. Nural, J. A. Brady, M. Assefi, S. Safaei, M. Allahyari, S. Pouriyeh, M. R. Galinski, J. C. Kissinger, and J. B. Gutierrez, "A Vision for Health Informatics: Introducing the SKED Framework.An Extensible Architecture for Scientific Knowledge Extraction from Data," *ArXiv e-prints*, 2017.

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text Summarization Techniques: A Brief Survey," *ArXiv e-prints*, 2017.

[9] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Computers and Communications (ISCC), 2017 IEEE Symposium on*. IEEE, 2017, pp. 204–207.

[10] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.

[11] M. Allahyari and K. Kochut, "Automatic topic labeling using ontology-based topic models," in *14th International Conference on Machine Learning and Applications (ICMLA), 2015*. IEEE, 2015.

[12] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International journal of human-computer studies*, vol. 43, no. 5, pp. 907–928, 1995.

[13] S. Fodeh, B. Punch, and P.-N. Tan, "On ontology-driven document clustering using core semantic features," *Knowledge and information systems*, vol. 28, no. 2, pp. 395–421, 2011.

[14] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 389–396.

[15] A. Hotho, A. Maedche, and S. Staab, "Ontology-based text document clustering," *KI*, vol. 16, no. 4, pp. 48–54, 2002.

[16] M. Allahyari, K. J. Kochut, and M. Janik, "Ontology-based text classification into dynamically defined topics," in *IEEE International Conference on Semantic Computing (ICSC), 2014*. IEEE, 2014, pp. 273–278.

[17] Q. Luo, E. Chen, and H. Xiong, "A semantic term weighting scheme for text categorization," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12 708–12 716, 2011.

[18] L. Cai, G. Zhou, K. Liu, and J. Zhao, "Large-scale question classification in cqa by leveraging wikipedia semantic knowledge," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1321–1330.

[19] C. Boston, H. Fang, S. Carberry, H. Wu, and X. Liu, "Wikimantic: Toward effective disambiguation and expansion of queries," *Data & Knowledge Engineering*, vol. 90, pp. 22–37, 2014.

[20] C. Li, A. Sun, and A. Datta, "A generalized method for word sense disambiguation based on wikipedia," in *Advances in Information Retrieval*. Springer, 2011, pp. 653–664.

[21] ——, "Tsdw: Two-stage word sense disambiguation using wikipedia," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 6, pp. 1203–1223, 2013.

[22] P. Ristoski and H. Paulheim, "Semantic web in data mining and knowledge discovery: A comprehensive survey," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2016.

[23] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[25] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 697–702.

[26] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 490–499.

[27] D. Magatti, S. Calegari, D. Ciucci, and F. Stella, "Automatic labeling of topics," in *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*. IEEE, 2009, pp. 1227–1232.

[28] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1536–1545.

[29] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 465–474.

[30] S. Hingmire and S. Chakraborti, "Topic labeled text classification: a weakly supervised approach," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 385–394.

[31] J. Li, C. Cardie, and S. Li, "Topicspam: a topic-model based approach for spam detection." in *ACL (2)*, 2013, pp. 217–221.

[32] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.

[33] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 375–384.

[34] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 533–542.

[35] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433.

[36] J. H. Lau, D. Newman, S. Karimi, and T. Baldwin, "Best topic word selection for topic labelling," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 605–613.

[37] X.-L. Mao, Z.-Y. Ming, Z.-J. Zha, T.-S. Chua, H. Yan, and X. Li, "Automatic labeling hierarchical topics," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2383–2386.

[38] J. Chen, J. Yan, B. Zhang, Q. Yang, and Z. Chen, "Diverse topic phrase extraction through latent semantic analysis," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 834–838.

[39] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with pachinko allocation," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 633–640.

[40] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers, "Modeling documents by combining semantic concepts with unsupervised statistical learning," in *The Semantic Web-ISWC 2008*. Springer, 2008, pp. 229–244.

[41] C. Chemudugunta, P. Smyth, and M. Steyvers, "Combining concept hierarchies and statistical topic models," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 1469–1470.

[42] C. P. Robert and G. Casella, *Monte Carlo statistical methods*. Citeseer, 2004, vol. 319.

[43] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 1, p. 4, 2010.

[44] H. M. Wallach, D. Minmo, and A. McCallum, "Rethinking lda: Why priors matter," 2009.

[45] T. Minka, "Estimating a dirichlet distribution," 2000.

[46] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[47] M. Shirakawa, K. Nakayama, T. Hara, and S. Nishio, "Concept vector extraction from wikipedia category network," in *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication*. ACM, 2009, pp. 71–79.

[48] H. Nesi, "Bawe: an introduction to a new resource," *New trends in corpora and language learning*, pp. 212–28, 2011.

[49] S. Banerjee and T. Pedersen, "The design, implementation, and use of the Ngram Statistic Package," in *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2003, pp. 370–381.

[50] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.