OntoLDA: An Ontology-based Topic Model for Automatic Topic Labeling

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Mehdi Allahyari [a,*], Seyedamin Pouriyeh [a], Krys Kochut [a] and Hamid R. Arabnia [a]
[a] *Department of Computer Science, University of Georgia, Athens, GA, USA*
E-mail: {mehdi,pouriyeh,kochut,hra}@cs.uga.edu

Abstract. Topic models, which frequently represent topics as multinomial distributions over words, have been extensively used for discovering latent topics in text corpora. Topic labeling, which aims to assign meaningful labels for discovered topics, has recently gained significant attention. In this paper, we argue that the quality of topic labeling can be improved by considering ontology concepts rather than words alone, in contrast to previous works in this area, which usually represent topics via groups of words selected from topics. We have created: (1) a topic model that integrates ontological concepts with topic models in a single framework, where each topic is represented as a multinomial distribution over concepts and each concept is a multinomial distribution over words, and (2) a topic labeling method based on the ontological meaning of the concepts included in the discovered topics. In selecting the best topic labels, we rely on the semantic relatedness of the concepts and their ontological classifications. The results of our experiments conducted on two different data sets show that introducing ontological concepts as additional, richer features between topics and words and describing topics in terms of concepts offers an effective method for generating meaningful labels for the discovered topics.

Keywords: Statistical learning, Topic modeling, Topic model labeling, ontologies, Linked Open Data

1. Introduction

Topic models such as Latent Dirichlet Allocation (LDA) [5] have gained considerable attention, recently. They have been successfully applied to a wide variety of text mining tasks, such as word sense disambiguation [19,7], sentiment analysis [24], information retrieval [46] and others, in order to identify hidden topics in text documents. Topic models typically assume that documents are mixtures of topics, while topics are probability distributions over the vocabulary. When the topic proportions of documents are estimated, they can be used as the themes (high-level representations of the semantics) of the documents. Highest-ranked words in a topic-word distribution indicate the meaning of the topic. Thus, topic models provide an effective framework for extracting the latent semantics from unstructured text collections. For example, Table 1 shows the top words of a topic learned from a collection of computer science abstracts; the topic has been labeled by a human “relational databases”.

However, even though the topic word distributions are usually meaningful, it is very challenging for the users to accurately interpret the meaning of the topics based only on the word distributions extracted from the corpus, particularly when they are not familiar with the domain of the corpus. It would be very difficult to answer questions such as “What is a topic talking about?” and “What is a good enough label for a topic?”

Topic labeling means finding one or a few phrases that sufficiently explain the meaning of the topic. This

[*] Corresponding author. E-mail: mehdi@cs.uga.edu

Table 1

Example of a topic with its label.

Human Label: relational databases				
query	database	databases	queries	processing
efficient	relational	object	xml	systems

task, which can be labor intensive particularly when dealing with hundreds of topics, has recently attracted considerable attention.

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 1.

Within the Semantic Web, numerous data sources have been published as ontologies. Many of them are inter-connected 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 example, DBpedia [4] (as part of LOD) is a publicly available knowledge base extracted from Wikipedia in the form of an ontology of concepts and relationships, making this vast amount of information programmatically accessible on the Web.

The principal objective of the research presented here is to leverage and incorporate the semantic graph of concepts in an ontology, DBpedia in this work, and their various properties within unsupervised topic models, such as LDA. In our model, we introduce another latent variable called, *concept*, i.e. ontological concept, between topics and words. Thus, each document is a multinomial distribution over topics, where each topic is represented as a multinomial distribution over concepts, and each concept is defined as a multinomial distribution over words.

Defining the concept latent variable as another layer between topics and words has multiple advantages: (1) it gives us much more information about the topics; (2) it allows us to illustrate topics more specifically, based on ontology concepts rather than words, which can be used to label topics; (3) it automatically integrates topics with knowledge bases. We first presented the our ontology-based topic model, OntoLDA model, in [1] 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 ontology-based framework, demonstrating the effectiveness of our proposed model over two datasets.

Our contributions in this work are as follows:

1. We propose an ontology-based topic model, OntoLDA, which incorporates an ontology into the topic model in a systematic manner. Our model integrates the topics to external knowledge bases, which can benefit other research areas such as information retrieval, classification and visualization.
2. We introduce a topic labeling method, based on the semantics of the concepts that are included in the discovered topics, as well as ontological relationships existing among the concepts in the ontology. Our model improves the labeling accuracy by exploiting the topic-concept relations and can automatically generate labels that are meaningful for interpreting the topics.
3. We demonstrate the usefulness of our approach in two ways. We first show how our model can be exploited to link text documents to ontology concepts and categories. Then we illustrate automatic topic labeling by performing a series of experiments.

The paper is organized 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.

2. Background

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

2.1. Ontologies

Ontologies are fundamental elements of the Semantic Web and could be thought of knowledge represen-

[1] http://linkeddata.org/

tation methods, which are used to specify the knowledge shared among different systems. An ontology is referred to an "explicit specification of a conceptualization." [14]. 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 [12,18,17], text classification [2,29,8], word sense disambiguation [6,25,26], and others. See [38] for a comprehensive review of Semantic Web in data mining and knowledge discovery.

2.2. Probabilistic Topic Models

Probabilistic topic models are a set of algorithms that are used to uncover the hidden thematic structure from a collection of documents. The main idea of topic modeling is to create a probabilistic generative model for the corpus of text documents. In topic models, documents are mixture of topics, where a topic is a probability distribution over words. The two main topic models are Probabilistic Latent Semantic Analysis (pLSA) [16] and Latent Dirichlet Allocation (LDA) [5]. Hofmann (1999) introduced pLSA for document modeling. pLSA model does not provide any probabilistic model at the document level which makes it difficult to generalize it to model new unseen documents. Blei et al. [5] extended this model by introducing a Dirichlet prior on mixture weights of topics per documents, and called the model Latent Dirichlet Allocation (LDA). In this section we describe the LDA method.

The latent Dirichlet allocation (LDA) [5] is a generative probabilistic model for extracting thematic information (topics) of a collection of documents. LDA assumes that each document is made up of various topics, where each topic is a probability distribution over words.

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 of the corpus. A topic $z_j, 1 \leq j \leq K$ is represented as a multinomial probability distribution over the $V|$ words, $p(w_i|z_j)$, $\sum_i^V p(w_i|z_j) = 1$. LDA generates the words in a two-stage process: words are generated from topics and topics are generated by documents. More formally, the distribution of words given the document is calculated as follows:

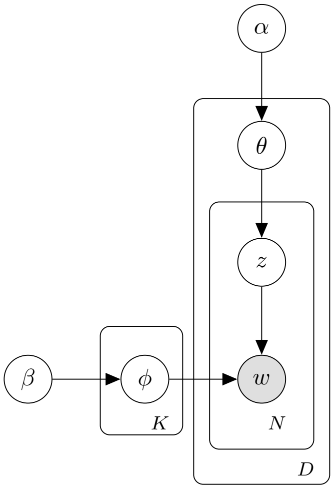


Fig. 1. LDA Graphical Model

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

The graphical model of LDA is shown in Figure 1 and the generative process for the corpus $\mathcal{D}$ is as follows:

1. For each topic $k \in \{1, 2, \ldots, K\}$, sample a word distribution $\phi_k \sim \mathrm{Dir}(\beta)$
2. For each document $d \in \{1, 2, \ldots, D\}$,

   (a) Sample a topic distribution $\theta_d \sim \mathrm{Dir}(\alpha)$
   (b) For each word w_n, where $n \in \{1, 2, \ldots, N\}$, in document d,

      i. Sample a topic $z_i \sim \mathrm{Mult}(\theta_d)$
      ii. Sample a word $w_n \sim \mathrm{Mult}(\phi_{z_i})$

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

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

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 [13]. 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.

3. 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. When the topic proportions of documents are estimated, they can be used as the themes (high-level semantics) of the documents. Top-ranked words in a topic-word distribution indicate the meaning of the topic.

For example, Table 2 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, it is very difficult for the users to accurately infer the meanings of the topics just from the top words, particularly when they are not familiar with 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 sufficiently explain the meaning of the topic. For instance, As shown in Table 2, topics do not have any labels, therefore they must be manually assigned. Topic labeling task can be labor intensive particularly when dealing with hundreds of topics. Table 3 illustrates the same topics that have been labeled (second row in the table) manually by a human.

Table 2

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

Table 3

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

Automatic topic labeling which aims to to automatically generate meaningful labels for the topics has recently attracted increasing attention [45,33,30,22,20]. Unlike previous works that have essentially concentrated on the topics learned from LDA topic model and represented the topics by words, we propose an ontology-based topic model, OntoLDA, 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 incorporate the semantic graph of concepts in an ontology (e.g., DBpedia) and their various properties with unsupervised topic models, such as LDA, in a principled manner and exploit this information to automatically generate meaningful topic labels.

4. 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 [15,27,41], word sense disambiguation [19,7], sentiment analysis [24,28], 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 [5,13]. This option is not satisfactory, because it necessitates significant perception to interpret the topic, particularly if the user is not familiar with the domain of the topic. For example, it would be very hard to infer the meaning of the topic shown in Table 1 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 [32,44]. This approach has disadvantages: (a) the labels are prone to subjectivity; and (b) the method can not be scale up, especially when dealing with massive number of topics.

Recently, automatic topic labeling has been an area of active research. [45] represented topics as multinomial distribution over n-grams, so top n-grams of a topic can be used to label the topic. Mei et al. [33] proposed an approach to automatically label the topics by converting 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 minimum KL divergence and the maximum mutual information with the topic to label the corresponding topic. [30] 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 find the best label according to a set of similarity measures.

Lau et al. [23] 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. [22] reused the features proposed in [23] 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. [31] proposed a topic labeling approach which enhances the labeling by using the sibling and parent-child relations between topics. They first generate a set of candidate labels by extracting meaningful phrases using Ngram Testing [11] 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. [20] 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.

Our work is different from all previous works in that we propose a topic model that integrates structured data with data-driven topics within a single general framework. Prior works basically focus on the topics learned via LDA topic model (i.e. topics are multinomial distribution over words) whereas in our model we introduce another latent variable called *concept* between topics and words, i.e., each document is a multinomial distribution over topics where each topic is represented as a multinomial distribution over concepts and each concept is defined as a multinomial distribution over words.

The hierarchical topic models, which represent correlations among topics, are conceptually related to our OntoLDA model. Mimno et al. [34] proposed the hPAM model that models a document as a mixture of distributions over super-topics and sub-topics, using a directed acyclic graph to represent a topic hierarchy. The OntoLDA model is different, because in hPAM, distribution of each super-topic over sub-topics depends on the document, whereas in OntoLDA, 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 OntoLDA, ontological concepts are n-grams, which makes them more specific and more meaningful, a key point in OntoLDA.

[2] http://dbpedia.org

[9,10] introduced topic models that combine concepts with data-driven topics. The key idea in their frameworks is that topics from the statistical topic models and concepts of the ontology are both represented by a set of "focused" words, i.e. distributions over words, and they use this similarity in their models. However, our OntoLDA model is different from these models in that they treat the concepts and topics in the same way, whereas in OntoLDA, concepts and topics form two distinct layers in the model.

5. Problem Formulation

In this section, we formally describe our model and its learning process. We then explain how to leverage the topic-concept distribution to generate meaningful semantic labels for each topic, in section 4. The notation used in this paper is summarized in Table 5.

Most topic models like LDA consider each document as a mixture of topics where each topic is defined as a multinomial distribution over the vocabulary. Unlike LDA, OntoLDA defines another latent variable called *concept* between topics and words, i.e., each document is a multinomial distribution over topics where each topic is a represented as a multinomial distribution over concepts and each concept is defined as a multinomial distribution over words.

The intuition behind our model is that using words from the vocabulary of the document corpus to represent topics is not a good way to convey the meaning of the topics. Words usually describe topics in a broad way while ontological concepts express the topics in a more focused way. Additionally, concepts representing a topic are semantically more closely related to each other. As an example, the first column of Table 4 lists a topic learned by standard LDA and represented by top words, whereas the second column shows the same topic learned by the OntoLDA model, which represents the topic using ontology concepts. From the topic-word representation we can conclude that the topic is about "sports", but the topic-concept representation indicates that not only the topic is about "sports", but more specifically about "American sports".

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

Table 4

Example of topic-word representation learned by LDA and topic-concept representation learned by OntoLDA.

LDA		**OntoLDA**	
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)

Table 5

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

Definition 1. (Concept): A *concept* in a text collection $\mathcal{D}$ is represented by c and defined as a multinomial 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 text collection $\mathcal{D}$ 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., $\{p(\phi_k|\theta_d)\}_{k\in K}$.

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.,

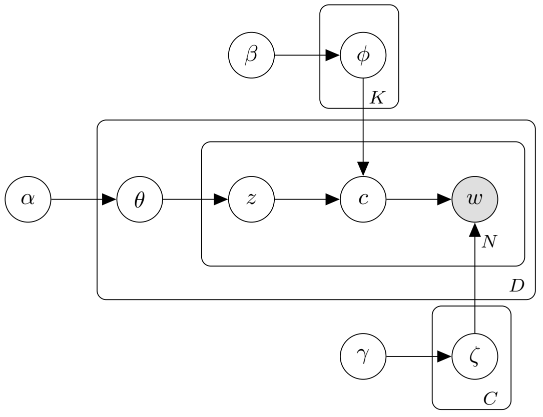


Fig. 2. Graphical representation of OntoLDA model

$\{\phi_1, \phi_2, \ldots, \phi_K\}$, where the number of topics, K, is specified by the user.

5.1. The OntoLDA Topic Model

The key idea of the OntoLDA topic model is to integrate ontology concepts directly with topic models. Thus, topics are represented as distributions over concepts, and concepts are defined as distributions over the vocabulary. Later in this paper, concepts will also be used to identify appropriate labels for topics.

The OntoLDA topic model is illustrated in Figure 2 and the generative process is defined as Algorithm 1.

Algorithm 1: OntoLDA Topic Model

```
foreach concept $c \in \{1, 2, \ldots, C\}$ do
    Draw a word distribution $\zeta_c \sim$ Dir($\gamma$)
end
foreach topic $k \in \{1, 2, \ldots, K\}$ do
    Draw a concept distribution $\phi_k \sim$ Dir($\beta$)
end
foreach document $d \in \{1, 2, \ldots, D\}$ do
    Draw a topic distribution $\theta_d \sim$ Dir($\alpha$)
    foreach word $w$ of document $d$ do
        Draw a topic $z \sim$ Mult($\theta_d$)
        Draw a concept $c \sim$ Mult($\phi_z$)
        Draw a word $w$ from concept $c, w \sim$
        Mult($\zeta_c$)
    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:

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

5.2. Inference using Gibbs Sampling

Since the posterior inference of the OntoLDA is intractable, we need to find an algorithm for estimating posterior inference. A variety of algorithms have been used to estimate the parameters of topic models, such as variational EM [5] and Gibbs sampling [13]. In this paper we will use collapsed Gibbs sampling procedure for OntoLDA topic model. Collapsed Gibbs sampling [13] is a Markov Chain Monte Carlo (MCMC) [39] algorithm which constructs a Markov chain over the latent variables in the model and converges to the posterior distribution after a number of iterations. In our case, we aim 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 [40] argue that in cases where latent variables are greatly related, blocked sampling boosts convergence of the Markov chain and decreases auto-correlation, as well.

We derive the posterior inference from Eq. 3 as follows:

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

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 [43], 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 [36] 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 OntoLDA 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}:

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

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

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

6. Concept-based Topic Labeling

The intuition behind our approach is that entities (i.e., ontology concepts and instances) occurring in the text along with relationships among them can determine the document's topic(s). Furthermore, the entities classified into the same or similar domains in the on-

Algorithm 2: OntoLDA Gibbs Sampling

```
Input  : A collection of documents $D$, number of topics $K$ and $\alpha, \beta, \gamma$
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 $\phi, \theta$ and $\zeta$ randomly;
if computing parameter estimation then
    initialize alpha parameters, $\alpha$, using Eq. 9;
end
$t \leftarrow 0$;
while $t < MaxIteration$ do
    foreach word $w$ do
        $c = \mathbf{c}(w)$ // get the current concept assignment
        $k = \mathbf{z}(w)$ // get the current topic assignment
        // Exclude the contribution of the current word w
        $n_w^{(c)} \leftarrow n_w^{(c)} - 1$;
        $n_c^{(k)} \leftarrow n_c^{(k)} - 1$;
        $n_k^{(d)} \leftarrow n_k^{(d)} - 1$ // w is a document word
        $(newk, newc)$ = sample new topic-concept and concept-word for word $w$ using Eq. 8;
        // Increment the count matrices
        $n_w^{(newc)} \leftarrow n_w^{(newc)} + 1$;
        $n_{newc}^{(newk)} \leftarrow n_{newc}^{(newk)} + 1$;
        $n_{newk}^{(d)} \leftarrow n_{newk}^{(d)} + 1$;
        // Update the concept assignments and topic assignment vectors
        $\mathbf{c}(w) = newc$;
        $\mathbf{z}(w) = newk$;
        if computing parameter estimation then
            update alpha parameters, $\alpha$, using Eq. 9;
        end
    end
    $t \leftarrow t + 1$;
end
```

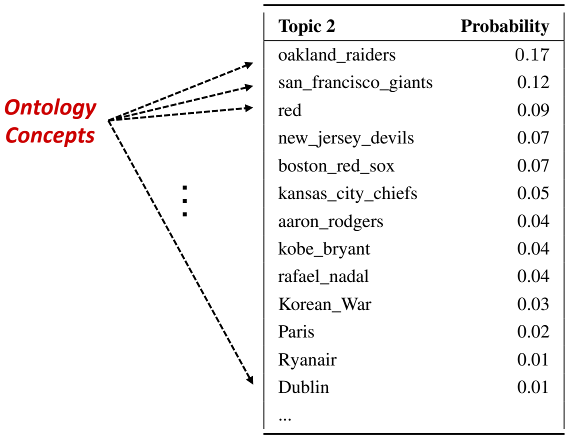
tology are semantically closely related to each other. Hence, we rely on the semantic similarity between the information included in the text and a suitable fragment of the ontology in order to identify good labels for the topics. Research presented in [2] use a similar approach to perform ontology-based 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.

Our approach focuses only on the ontology concepts and their class hierarchy as topic labels. Finding meaningful and semantically relevant labels for an identified topic ϕ involves four primary steps: (1) construction of the semantic graph from top concepts in the given topic; (2) selection and analysis of the thematic graph, a semantic graph's subgraph; (3) topic graph extraction from the thematic graph concepts; and (4) computation of the semantic similarity between topic ϕ and the candidate labels of the topic label graph.

6.1. Semantic Graph Construction

We use the marginal probabilities $p(c_i|\phi_j)$ associated with each concept c_i in a given topic ϕ_j and ex-


Ontology
Concepts
⋮

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 OntoLDA.

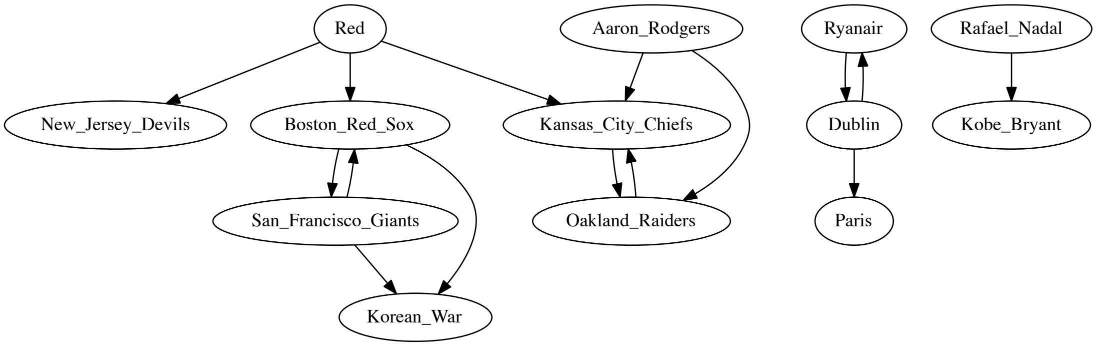


Fig. 4. Semantic graph of the example topic ϕ described in Fig. 3 with $|V^{\phi}| = 13$

tract the $\mathcal{K}$ concepts with the highest marginal probability to construct the topic's semantic graph. Figure 3 shows the top-10 concepts of a topic learned by OntoLDA.

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

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

6.2. Thematic Graph Selection

The selection of the thematic graph is based on the assumption that concepts under a given topic are closely associated in the ontology, whereas concepts from different topics are placed far apart, or even not connected at all. Due to the fact that topic models are statistical and data driven, they may produce topics that are not coherent. In other words, for a given topic that is represented as a list of $\mathcal{K}$ most probable concepts, there may be a few concepts which are not semantically close to other concepts and to the topic, accordingly. As a result, the topic's semantic graph may be composed of multiple connected components.

Definition 7. (Thematic graph): A *thematic graph* is a connected component of G^{ϕ}. In particular, 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.

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

6.3. Topic Label Graph Extraction

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

We determine the importance of concepts in a thematic graph not only by their initial weights, which are the marginal probabilities of concepts under the topic, but also by their relative positions in the graph. Here, we utilize the HITS algorithm [21] with the assigned initial weights for concepts to find the *authoritative concepts* in the dominant thematic graph. Subsequently, we locate the *central concepts* in the graph based on the geographical centrality measure, since these nodes can be identified 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 Figure 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" has far more relationships to other concepts in DBpedia).

From now on, we will simply write thematic graph when referring to the dominant thematic graph of a topic.

To extract the topic label graph for the core concepts CC^{ϕ}, we primarily focus on the ontology class structure, since we can consider the topic labeling as assigning class labels to topics. We introduce definitions similar to those in [20] for describing 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, due to the fact that increasing the distance further quickly leads to excessively general classes.

6.4. 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. [33] describe that the semantics of a topic should be interpreted based on two parameters: (1) distribution of the topic; and (2) the context of the topic. Our topic label graph for a topic ϕ is extracted, taking into account the topic distribution over the concepts as well as the context of the topic in the form of semantic relatedness between the concepts in the ontology.

In order to find the semantic similarity of a label ℓ in $\mathbf{G}_{cc^{\phi}}$ to a topic ϕ, we compute the semantic similarity between ℓ and all of the concepts in the core concept set CC^{ϕ}, rank the labels and then select the best labels for the topic.

A candidate label is scored according to three main objectives: (1) the label should cover *important concepts* of the topic (i.e. concepts with higher marginal probabilities); (2) the label should be specific (lower in the class hierarchy) to the core concepts; and (3) the label should cover the highest number of core concepts in $\mathbf{G}_{cc^{\phi}}$.

To compute the semantic similarity of a label to a concept, we first calculate the *membership score* and the *coverage score*. We have adopted a modified Vector-based Vector Generation method (VVG) described in [42] to calculate the membership score of a concept to a label.

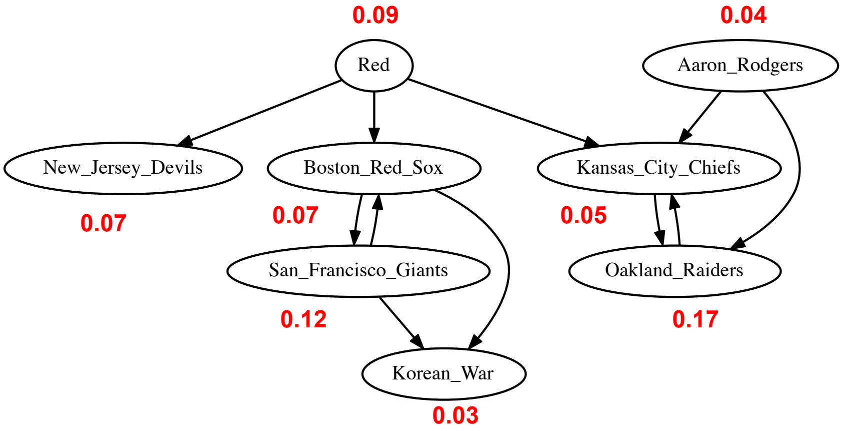


Fig. 5. Dominant thematic graph of the example topic described in Fig. 4

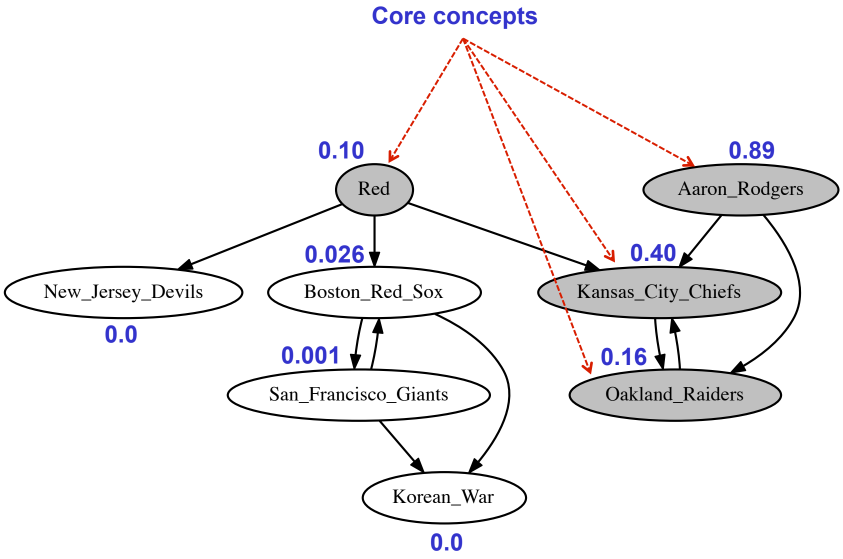
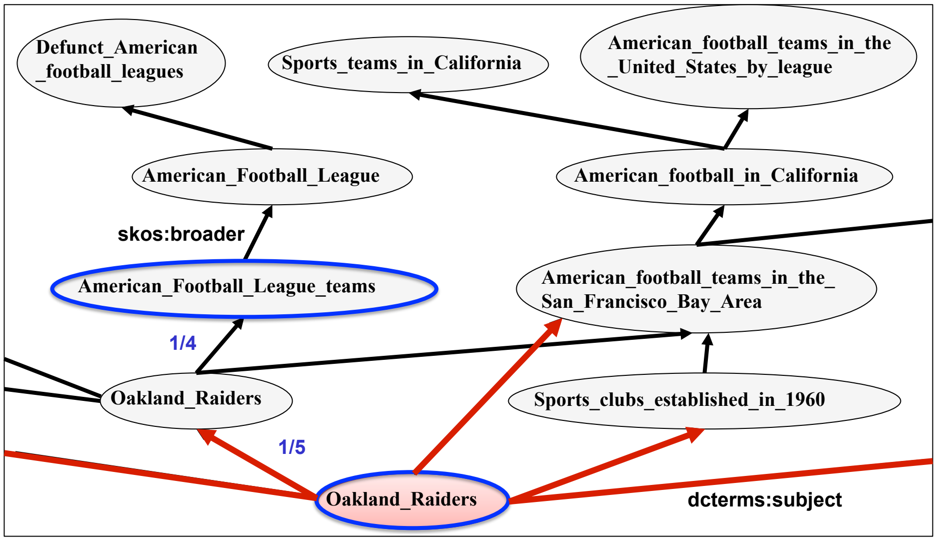


Fig. 6. Core concepts of the Dominant thematic graph of the example topic described in Fig. 5


Defunct_American _football_leagues
Sports_teams_in_California
American_football_teams_in_the _United_States_by_league
American_Football_League
American_football_in_California
skos:broader
American_Football_League_teams
American_football_teams_in_the_ San_Francisco_Bay_Area
1/4
Oakland_Raiders
Sports_clubs_established_in_1960
1/5
Oakland_Raiders
dcterms:subject

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

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

In the experiments described in this paper, we used DBpedia, an ontology created out of Wikipedia. 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 $E_l = \{e_1, e_2, \dots, e_m\}$ represents the set of all membership relationships forming the shortest path p from concept c_i to category C_j. Figure 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:

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

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

Table 6

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

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

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:

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

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 ranks a label node higher, if the label covers more important topical concepts, if it is closer to the core concepts, and if it covers more core concepts. Top-ranked labels are selected as the labels for the given topic. Table 6 illustrates a topic along with the top-10 generated labels using our ontology-based framework.

7. Experiments

In order to demonstrate the effectiveness of our OntoLDA method, utilizing ontology-based topic models, we compared it to one of the state-of-the-art traditional, text-based approaches described in [33]. We will refer to that method as Mei07.

We selected two different data sets for our experiments. First, we extracted the top-2000 bigrams using the N-gram Statistics Package [3]. Then, we tested the significance of the bigrams using the Student's T-Test, and extracted the top 1000 candidate bigrams $\mathcal{L}$. For each label $\ell \in \mathcal{L}$ and topic ϕ, we computed the score s, defined by the authors as:

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

where PMI is the point-wise mutual information between the label ℓ and the topic words w, given the document corpus D. We selected the top-6 labels as the labels of the topic ϕ generated by the Mei07 method.

7.1. Data Sets and Concept Selection

The experiments in this paper are based on two text corpora and the DBpedia ontology. The text collections are: the British Academic Written English Corpus (BAWE) [37], and a subset of the Reuters[3] news articles. BAWE contains $2,761$ documents of proficient university-level student writing that are fairly evenly divided into four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and

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

Physical Sciences) covering 32 disciplines. In this paper, we focused on the documents categorized as LIFE SCIENCES (covering Agriculture, Biological Sciences, Food Sciences, Health, Medicine and Psychology) consisting of $D = 683$ documents and $218,692$ words. The second dataset is composed of $D = 1,414$ Reuters news articles divided into four main topics: *Business*, *Politics*, *Science*, and *Sports*, consisting of $155,746$ words.

Subsequently, we extracted 20 major topics from each dataset using OntoLDA and, similarly, 20 topics using Mei07.

The DBpedia ontology created from the English language subset of Wikipedia includes over $5,000,000$ concepts. Using the full set of concepts included in the ontology is computationally very expensive. Therefore, we selected a subset of concepts from DBpedia that were relevant to our datasets. We identified $16,719$ concepts (named entities) mentioned in the BAWE dataset and $13,676$ in the Reuters news dataset and used these concept sets in our experiments.

7.2. Experimental Setup

We pre-processed the datasets by removing punctuation, stopwords, numbers, and words occurring fewer than 10 times in each corpus. For each concept in the two concept sets, we created a bag of words by downloading its Wikipedia page and collecting the text, and eventually, constructed a vocabulary for each concept set. Then, we created a $W = 4,879$ vocabulary based on the intersection between the vocabularies of BAWE corpus and its corresponding concept set. We used this vocabulary for experiments on the BAWE corpus. Similarly, we constructed a $W = 3,855$ vocabulary by computing the intersection between the Reuters news articles and its concept set and used that for the Reuters experiments. We assumed symmetric Dirichlet prior and set $\beta = 0.01$ and $\gamma = 0.01$. We ran the Gibbs sampling algorithm for 500 iterations and computed the posterior inference after the last sampling iteration.

7.3. Results

Tables 7 and 8 present sample results of our topic labeling method, along with labels generated from the Mei07 method as well as the top-10 words for each topic. For example, the columns with title “Topic 1” show and compare the top-6 labels generated for the same topic under Mei07 and the proposed OntoLDA method, respectively. We compared the top-6 labels and the top words for each topic are also shown in the respective Tables. We believe that the labels generated by OntoLDA are more meaningful than the corresponding labels created by the Mei07 method.

In order to quantitatively evaluate the two methods, we asked three human assessors to compare the labels. We selected a subset of topics in a random order and for each topic, the judges were given the top-6 labels generated by the OntoLDA method and Moi07. The labels were listed randomly and for each label the assessors had to choose between “Good” and “Unrelated”.

We compared the two different methods using the *Precision*$@k$, taking the top-1 to top-6 generated labels into consideration. Precision for a topic at top-k is defined as follows:

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

We then averaged the precision over all the topics. Figure 8 illustrates the results for each individual corpus.

The results in Figure 8, reveal two interesting observations: (1) in Figure 8(a), the precision difference between the two methods illustrates the effectiveness of our method, particularly for up to top-3 labels, and (2) the average precision for the BAWE corpus is higher than for the Reuters corpus. Regarding (1), our method assigns the labels that are more specific and meaningful to the topics. As we select more labels, they become more general and likely too broad for the topic, which impacts the precision. For the BAWE corpus as shown in 8(b), the precision begins to rise as we select more top labels and then starts to fall. The reason for this is that OntoLDA finds the labels that are likely too specific to match the topics. But, as we choose further labels ($1 < k \leq 4$), they become more general but not too broad to describe the topics, and eventually ($k > 4$) the labels become too general and consequently not appropriate for the topics. Regarding observation (2), the BAWE documents are educational and scientific, and phrases used in scientific documents are more discriminative than in news articles. This makes the constructed semantic graph include more inter-related concepts and ultimately leads to the selection of concepts that are good labels for the scientific documents, which is also discussed in [33].

Topic Coherence. In our model, the topics are represented over concepts. Hence, in order to compute the

Table 7

Sample topics of the BAWE corpus with top-6 generated labels for the Mei method and OntoLDA + 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

OntoLDA + 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

word distribution for each topic t under OntoLDA, we can use the following formula:

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

Table 9 shows three example topics from the BAWE corpus. Each "topic" column illustrates the top words from LDA and OntoLDA, respectively.

Based on Table 9, we can draw an interesting observation. Although both LDA and OntoLDA represent the top words for each topic, the ***topic coherence*** under OntoLDA is qualitatively better than LDA. For each topic we italicized and marked in red the wrong topical words. We can see that OntoLDA produces much better topics than LDA does. For example, "Topic 3" in Table 9 shows the top words for the same topic under standard LDA and OntoLDA. LDA did not perform well, as some words in most of the topics were considered as not relevant to the topic.

We performed quantitative comparison of the coherence of the topics created using OntoLDA and LDA, computing the *coherence score* based on the formula presented in [35]. This has become the most commonly used topic coherence evaluation method. 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 defined as:

Table 8
Sample topics of the Reuters corpus with top-6 generated labels for the Mei method and OntoLDA + 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

OntoLDA + 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

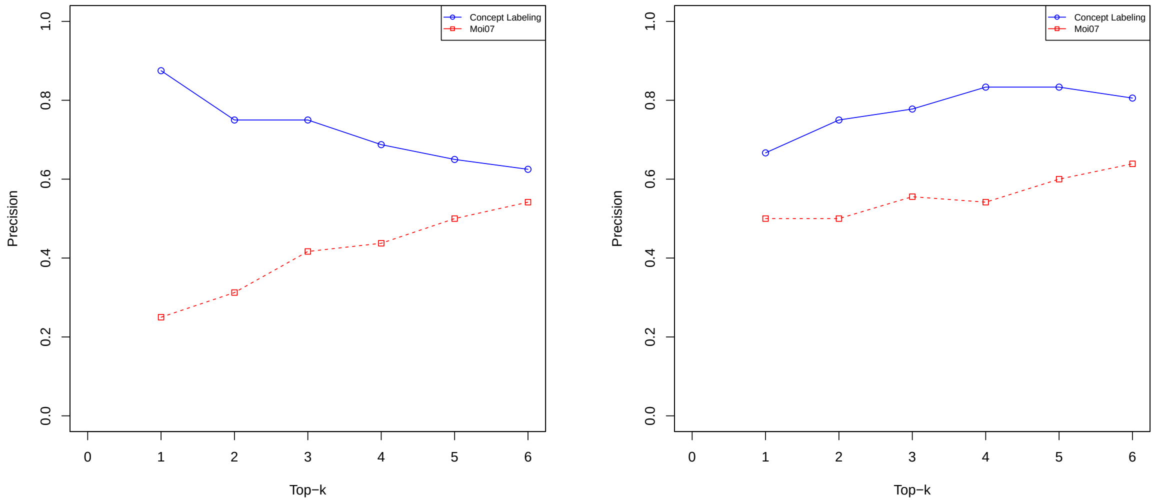
Topic top-10 words

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

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

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. It is demonstrated that the coherence score is highly consistent with human-judged topic coherence [35]. Higher coherence scores indicates higher quality of topics. The results are illustrated in Table 10.

As we mentioned before, OntoLDA represents each topic as a distribution over concepts. Table 11 illustrates the top-10 concepts of highest probabilities in the topic distribution under the OntoLDA framework for the same three topics (“topic 1”, “topic2” and “topic3”) of Table 9. Because concepts are more informative than individual words, the interpretation of top-



(a) Precision for Reuters Corpus

(b) Precision for BAWE Corpus

Fig. 8. Comparison of the systems using human evaluation

Table 9
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	**OntoLDA**	**LDA**	**OntoLDA**	**LDA**	**OntoLDA**	**LDA**	**OntoLDA**	**LDA**	**OntoLDA**
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

Table 10
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
OntoLDA	$\mathbf{-193.41}$	$\mathbf{-926.13}$	$\mathbf{-2474.70}$	$\mathbf{-206.14}$	$\mathbf{-1256.00}$	$\mathbf{-3213.00}$

Table 11
Example topics with top-10 concept distributions in OntoLDA 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

ics is more intuitive in OntoLDA than that of standard LDA.

8. Conclusions

In this paper, we presented OntoLDA, an ontology-based topic model, along with a graph-based topic labeling method for the task of topic labeling. Experimental results show the effectiveness and robustness of the proposed method when applied on different domains of text collections. The proposed ontology-based topic model improves the topic coherence in comparison to the standard LDA model by integrating ontological concepts with probabilistic topic models into a unified framework.

There are many interesting future extensions to this work. It would be interesting to define a global optimization scoring function for the labels instead of Eq. 17. Furthermore, how to incorporate the hierarchical relations as well as *lateral* relationships between the ontology concepts into the topic model, is also an interesting future direction.

References

[1] 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.

[2] 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*, pages 273–278. IEEE, 2014.

[3] 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*, pages 370–381, Mexico City, February 2003.

[4] 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*, 7(3):154–165, 2009.

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

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

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

[8] 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*, pages 1321–1330. ACM, 2011.

[9] 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*, pages 229–244. Springer, 2008.

[10] 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*, pages 1469–1470. ACM, 2008.

[11] 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*, pages 834–838. IEEE, 2006.

[12] S. Fodeh, B. Punch, and P.-N. Tan. On ontology-driven document clustering using core semantic features. *Knowledge and information systems*, 28(2):395–421, 2011.

[13] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[14] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5):907–928, 1995.

[15] 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, pages 385–394. ACM, 2014.

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

[17] A. Hotho, A. Maedche, and S. Staab. Ontology-based text document clustering. *KI*, 16(4):48–54, 2002.

[18] 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*, pages 389–396. ACM, 2009.

[19] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, 2014.

[20] 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*, pages 465–474. ACM, 2013.

[21] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[22] 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*, pages 1536–1545. Association for Computational Linguistics, 2011.

[23] 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*, pages 605–613. Association for Computational Linguistics, 2010.

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

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

[26] C. Li, A. Sun, and A. Datta. Tsdw: Two-stage word sense disambiguation using wikipedia. *Journal of the American Society for Information Science and Technology*, 64(6):1203–1223, 2013.

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

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

[29] Q. Luo, E. Chen, and H. Xiong. A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10):12708–12716, 2011.

[30] 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*, pages 1227–1232. IEEE, 2009.

[31] 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*, pages 2383–2386. ACM, 2012.

[32] 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*, pages 533–542. ACM, 2006.

[33] 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*, pages 490–499. ACM, 2007.

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

[35] 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*, pages 262–272. Association for Computational Linguistics, 2011.

[36] T. Minka. Estimating a dirichlet distribution, 2000.

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

[38] 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.

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

[40] 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)*, 28(1):4, 2010.

[41] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[42] 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*, pages 71–79. ACM, 2009.

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

[44] 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*, pages 424–433. ACM, 2006.

[45] 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*, pages 697–702. IEEE, 2007.

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