Automatic Topic Labeling using Ontology-based Topic Models

Mehdi Allahyari
Computer Science Department
University of Georgia, Athens, GA
Email: mehdi@uga.edu

Krys Kochut
Computer Science Department
University of Georgia, Athens, GA
Email: kochut@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 and each concept are represented as a multinomial distribution over concepts and over words, respectively, 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 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, DBpedia ontology**

I. Introduction

Topic models such as Latent Dirichlet Allocation (LDA) [1] have gained considerable attention, recently. They have been successfully applied to a wide variety of text mining tasks, such as word sense disambiguation [2], sentiment analysis [3] 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 semantics) of the documents. Top-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.

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. For example, Table I 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".

TABLE I. Example of a topic with its label.

Human Label: relational databases						
query	database	databases	queries	processing	efficient	relational

Topic labeling means finding one or a few phrases that sufficiently explain the meaning of the topic. This task, which can be labor intensive particularly when dealing with hundreds of topics, has recently attracted considerable attention.

Within the Semantic Web, numerous data sources have been published as ontologies. Many of them are interconnected as Linked Open Data (LOD)[1]. 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.

Recently, automatic topic labeling has been an area of active research. [5] 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. [6] proposed an approach to automatically label the topics by converting the labeling problem to an optimization problem. Thus, for each topic a candidate label is chosen that has the minimum Kullback-Leibler (KL) divergence and the maximum mutual information with the topic. In [7], the authors proposed a method for topic labeling based on: (1) generating the label candidate set from topic's top-terms and titles of Wikipedia pages containing the topic's top-terms; (2) scoring and ranking the candidate labels and selecting the top-ranked label as the label of the topic. Mao et al. [8] proposed a topic labeling approach which enhances the label selection by using the sibling and parent-child relations between topics. In a more recent work, Hulpus et al. [9] addressed the topic labeling by relying on the structured data from DBpedia. The main idea is to construct a topic graph of concepts corresponding to topic's top-k words from the DBpedia, apply graph-based centrality algorithms to rank the concepts, and then select the most prominent concepts as labels of the topic.

Our principal objective is to incorporate the semantic graph of concepts in an ontology, DBpedia here, and their various properties within unsupervised topic models, such as LDA. Our work is different from all previous works in that they basically focus on the topics learned via LDA topic model (i.e. topics are multinomial distribution over words). In our model, we introduce another latent variable called, *concept*, between

[1] `http://linkeddata.org/`

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.

The hierarchical topic models, which represent correlations among topics, are conceptually related to our ***OntoLDA*** model. Mimno et al. [10] 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, sub-topics 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. [11] introduced topic models that combine concepts with data-driven topics. Unlike these models, in OntoLDA concepts and topics form two distinct layers in the model.

In this paper, we propose (1) 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 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. We show how our model can be used to link text documents to ontology concepts and categories, as well as automatic topic labeling by performing a series of experiments.

II. Problem Formulation

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 to represent topics is not a good way to convey the meaning of the topics. Words usually describe topics in a broad way while 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 II 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".

TABLE II. Topic-word and topic-concept representations.

Topic-word	Topic-concept	Probability
team	oakland raiders	(0.174)
est	san francisco giants	(0.118)
home	red	(0.087)
league	new jersey devils	(0.074)
games	boston red sox	(0.068)
second	kansas city chiefs	(0.054)

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

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

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

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

Topic Modeling: Given a collection of text documents, D, the task of *Topic Modeling* aims at discovering and extracting K topics, i.e., $\{\phi_1, \phi_2, \ldots, \phi_K\}$.

A. The OntoLDA Topic Model

The key idea of the OntoLDA 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 1 and the generative process is given as follows:

1) For each concept $c \in \{1, 2, \ldots, C\}$,
   (a) Draw a word distribution $\zeta_c \sim \text{Dir}(\gamma)$
2) For each topic $k \in \{1, 2, \ldots, K\}$,
   (a) Draw a concept distribution $\phi_k \sim \text{Dir}(\beta)$
3) For each document $d \in \{1, 2, \ldots, D\}$,
   (a) Draw a topic distribution $\theta_d \sim \text{Dir}(\alpha)$
   (c) For each word w of document d,
      i. Draw a topic $z \sim \text{Mult}(\theta_d)$
      ii. Draw a concept $c \sim \text{Mult}(\phi_z)$
      iii. Draw a word w from concept $c, w \sim \text{Mult}(\zeta_c)$

B. 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

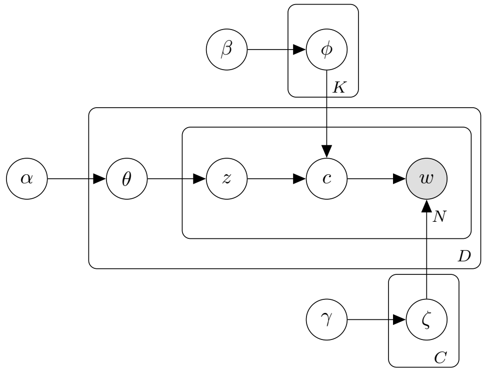


Fig. 1. Graphical representation of OntoLDA model

parameters of topic models, such as variational EM [1] and Gibbs sampling [12]. In this paper we will use collapsed Gibbs sampling procedure for OntoLDA topic model. Collapsed Gibbs sampling [12] is a Markov Chain Monte Carlo (MCMC) 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 $\boldsymbol{z}$ and $\boldsymbol{c}$ conditioned on observed words $\boldsymbol{w}$ and hyperparameters α, β and γ.

We derive the posterior inference as follows:

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

$$
P(z_i = k, c_i = c | w_i = w, \boldsymbol{z}_{-i}, \boldsymbol{c}_{-i}, \boldsymbol{w}_{-i}, \alpha, \beta, \gamma) \propto \\
\frac{n_{k,-i}^{(d)} + \alpha_k}{\sum_{k'} (n_{k',-i}^{(d)} + \alpha_{k'})} \times \frac{n_{c,-i}^{(k)} + \beta}{\sum_{c'} (n_{c',-i}^{(k)} + \beta)} \times \frac{n_{w,-i}^{(c)} + \gamma}{\sum_{w'} (n_{w',-i}^{(c)} + \gamma)} \tag{1}
$$

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 disregarded. Instead of using symmetric estimation of the parameters α, we use moment matching methods [13] to approximate these parameters.

III. 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 ontology 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. [14] use similar approach to do ontology-based text categorization. A ***topic label*** ℓ for topic ϕ is a sequence of words which is semantically meaningful and sufficiently explains the meaning of ϕ. To find meaningful and semantically relevant labels for an identified topic ϕ, our approach focuses only on the ontology concepts and their class hierarchy as topic labels, and 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.

A. 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 extract the $\mathcal{K}$ concepts with the highest marginal probability to construct the topic's semantic graph.

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

B. 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.

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. A ***dominant thematic graph*** for topic ϕ is a thematic graph with the largest number of nodes.

C. Topic Label Graph Extraction

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 [15] 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.

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. 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 [9] for describing the label graph and topic 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.

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.

D. Semantic Relevance Scoring Function

Mei et al. [6] 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 [16] to calculate the membership score of a concept to a label.

In the experiments described in this paper, we used DBpedia. 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*. 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:

$$m(e) = \frac{1}{N} \tag{2}$$

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

$$mScore(c_i, C_j) = \prod_{e_k \in E_l} m(e_k) \tag{3}$$

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

$$cScore(w_i, v_j) = \begin{cases} \frac{1}{d(c_i, C_j)} & \text{if there is a path from } c_i \text{ to } C_j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

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

$$SSim(c_i, \ell) = w(c_i) \cdot \Big[\lambda \cdot mScore(c_i, \ell) + (1 - \lambda) \cdot cScore(c_i, \ell)\Big] \tag{5}$$

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:

$$\begin{aligned} SSim(CC^{\phi}, \ell) = \frac{\lambda}{|CC^{\phi}|} \sum_{i=1}^{|CC^{\phi}|} w(c_i) \cdot mScore(c_i, \ell) \\ + (1 - \lambda) \sum_{i=1}^{|CC^{\phi}|} w(c_i) \cdot cScore(c_i, \ell) \end{aligned} \tag{6}$$

where λ is the smoothing factor to control the influence of the two scores. We used $\lambda = 0.8$ in our experiments. 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.

IV. Experiments

In order to demonstrate the effectiveness of our OntoLDA method, we compared it to one of the state-of-the-art traditional, text-based approaches described in [6]. We will refer to that method as Mei07. We selected two different data sets for our experiments. We extracted the top-2000 bigrams and 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:

$$s(\ell, \phi) = \sum_{w} \Big(p(w|\phi) PMI(w, \ell|D) \Big) \tag{7}$$

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.

A. 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) [17], and a subset of the Reuters[2] 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 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

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

TABLE III. SAMPLE BAWE TOPICS WITH TOP-5 GENERATED LABELS.

Mei07		OntoLDA + Concept Labeling	
Topic 1	**Topic 3**	**Topic 1**	**Topic 3**
rice production	cell lineage	agriculture	structural proteins
southeast asia	cell interactions	tropical agriculture	autoantigens
rice fields	somatic blastomeres	horticulture and gardening	cytoskeleton
crop residues	cell stage	model organisms	epigenetics
weed species	maternal effect	rice	genetic mapping

TABLE IV. SAMPLE REUTERS TOPICS W/ TOP-5 GENERATED LABELS.

Mei07		OntoLDA + Concept Labeling	
Topic 7	**Topic 8**	**Topic 7**	**Topic 8**
hockey league	mobile devices	national football league teams	investment banks
western conference	ralph lauren	washington redskins	house of morgan
national hockey	gerry shih	sports clubs established in 1932	mortgage lenders
stokes editing	huffington post	american football teams in maryland	jpmorgan chase
field goal	analysts average	green bay packers	banks established in 2000

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. Instead of using all 5 million DBpedia concepts, 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.

B. 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.

C. Results

Tables III and IV present sample results of our topic labeling method, along with labels generated from the Mei07 method. For example, the columns with title "Topic 1" show and compare the top-5 labels generated for the same topic under Mei07 and the proposed OntoLDA method, respectively. We compared the top-5 labels and the top words for each topic are shown in Table V. 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 Mei07. 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:

$$Precision@k = \frac{\text{\# of "Good" labels with rank} \leq k}{k} \qquad (8)$$

We then averaged the precision over all the topics. Figure 2 illustrates the results for each individual corpus. The results in Figure 2, reveal two interesting observations: (1) in Figure 2(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, 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 [6].

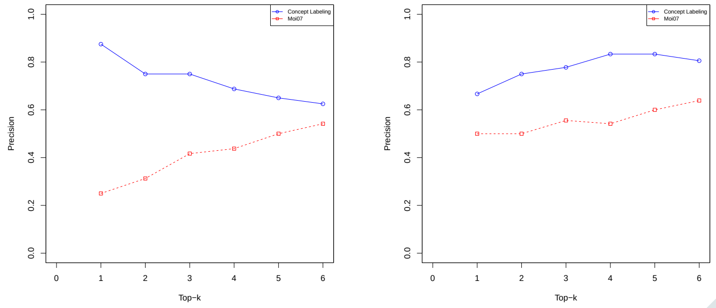
Topic Coherence. In our model, the topics are represented over concepts. Hence, in order to compute the word distribution for each topic t under OntoLDA, we can use the following formula:

$$\vartheta_t(w) = \sum_{c=1}^{\mathcal{C}} \Big(\zeta_c(w) \cdot \phi_t(c) \Big) \qquad (9)$$

Table V shows three example topics from the BAWE corpus. Each "topic" column illustrates the top words from LDA and OntoLDA, respectively. 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 V shows the top words for the same topic under standard LDA and OntoLDA. LDA did not perform well, as there are some words that are 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 [18]. 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:

$$C(\phi; V^{(\phi)}) = \sum_{t=2}^{T} \sum_{l=1}^{t-1} \log \frac{D(v_t^{(\phi)}, v_l^{(\phi)}) + 1}{D(v_l^{(\phi)})} \qquad (10)$$

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 [18]. Higher



(a) Precision for Reuters Corpus

(b) Precision for BAWE Corpus

Fig. 2. Comparison of the systems using human evaluation

TABLE V. TOP-10 WORDS FOR TOPICS FROM THE TWO DOCUMENT SETS. THE THIRD ROW PRESENTS THE MANUALLY GENERATED 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 VI. TOPIC COHERENCE ON TOP T WORDS.

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

coherence scores indicates higher quality of topics. The results are illustrated in Table VI. Table VII illustrates the concepts of highest probabilities in the topic distribution under the OntoLDA framework for the same three topics (“topic 1”, “topic2” and “topic3”) of Table V. Because concepts are more informative than individual words, the interpretation of topics is more intuitive in OntoLDA than standard LDA.

V. 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.

TABLE VII. TOPIC-CONCEPT DISTRIBUTION IN ONTOLDA.

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

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. 6. 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. It also would be interesting to investigate the use OntoLDA for other text mining tasks such as text classification.

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

[3] 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.

[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*, vol. 7, no. 3, pp. 154–165, 2009.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] T. Minka, “Estimating a dirichlet distribution,” 2000.

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

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

[16] 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.

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

[18] 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.