Ontology-Based Topic Labeling and Quality Prediction

Heidar Davoudi(✉) and Aijun An

Department of Electrical Engineering and Computer Science,
York University, Toronto, Canada
{davoudi,aan}@cse.yorku.ca

Abstract. Probabilistic topic models based on Latent Dirichlet Allocation (LDA) are increasingly used to discover hidden structure behind big text corpora. Although topic models are extremely useful tools for exploring and summarizing large text collections, most of time the inferred topics are not easy to understand and interpret by human. In addition, some inferred topics may be described by words that are not much relevant to each other and are thus considered low quality topics. In this paper, we propose a novel method that not only assigns a label to each topic but also identifies low quality topics by providing a reliability score for the label of each topic. Our rationale is that a topic labeling method cannot provide a good label for a low quality topic, and thus predicting label reliability is as important as topic labeling itself. We propose a novel measure (Ontology-Based Coherence) that can assess coherence of topics with respect to an ontology structure effectively. Empirical results on a real dataset and our user study show that the proposed predictive model using the defined measures can predict the label reliability better than two alternative methods.

Keywords: Topic modeling · Topic labeling · Labeling reliability

1 Introduction

Despite the fact that topic models (e.g., LDA [1]) are quite useful tools to explore and summarize the corpus, most of time the resulting latent (topic) space is not easily to interpret/evaluate by human. One approach to address this problem is to assign a label to each topic automatically [2–4]. However, most of automatic topic labeling methods do not provide information about the reliability for assigned labels. In other words, while resulting topics consist of some fused (i.e., more than one concept [5]) and random topics (i.e., no clear connection between the words [6]), automatic topic labeling schemes only deal with topic assignment and have nothing to do with these kinds of topics. Intuitively, *an automatic topic labeling method cannot assign a good label to a low quality topic.* As such, predicting reliability for the assigned labels is as important as the label assignment. Following this idea, we propose a novel automatic topic labeling

© Springer International Publishing Switzerland 2015
F. Esposito et al. (Eds.): ISMIS 2015, LNAI 9384, pp. 171–179, 2015.
DOI: 10.1007/978-3-319-25252-0_18

technique that not only assigns a label to each topic but also predicts the reliability of assigned labels. We define a novel coherence measure (Ontology-Based Coherence) based on the external ontology with a Directed Acyclic Graph (DAG) structure. The proposed predictive model utilizes this measure as the external labeling coherence criterion and average Normalized Pairwise Mutual Information (NPMI) as the internal topic coherence measure to predict the reliability of assigned labels. The major contributions of this study can be summarized as follows:

- Predicting topic labels and their reliabilities based on a DAG-structured ontology.
- Defining an appropriate measure to calculate the coherence of assigned label with respect to the external source (ontology).
- Designing a model based on both internal topic coherence (average NPMI) and the external labeling coherence (based on specificity) to calculate the reliability of predicted labeled.
- Empirical studies in a real dataset and respective user study show the predictive power of our proposed model.

2 Related Work

Topic labeling and evaluation tasks are conceptually related even though little studies have been done on topic labeling. For example, Mei et al. proposed a probabilistic framework for automatic topic labeling [2]. Their model extracted the meaningful phrases as the labels and then ranked them based on an scoring function. They defined the scoring function based on notion of good label (i.e., understandable, semantically relevant and distinguishable), but their model still cannot produce the labels which are the hypernyms of topic words [7]. Lau et al. [3] generated label candidates form Wikipedia (by querying the top topic words). They used a set of existing features (e.g., PMI, student t-test) and supervised and unsupervised methods to rank the candidates. In the unsupervised approach, each individual score (feature) between the label and topic words served as the ranking measure while in the supervised method, combination of features and users ratings were utilized to build a regression model. Meggati et al. [4] labeled topics based on alignment with hierarchal ontology (Google Directory). They used 6 similarity measures to find the best nodes in the tree. The final algorithm tried to find the best node in the tree by some heuristic rules even though the experiment was preliminary and did not show different aspects of the proposed algorithm. In another study, Newman et al. [7] conducted a large scale user study and investigated the correlation between PMI and human score for each topics. They showed that although top-10 topic words were reasonably good to be used as a part of label, they were not generally sufficient for the entire label. Musat et al. [8] suggested some coherence measures (based on coverage and specificity) with respect to an external ontology. The primary purpose of the paper was to evaluate the topics (and not labeling) and their measurements were

based on a tree structure ontology. While all these studies dealt with either evaluation or labeling tasks, our proposed method assigns labels and predicts their reliability based on the novel measures defined on a DAG structured ontology.

3 Problem Formulation

In this section, we formally define *topic*, *ontology*, and *topic labeling*. Given document collection D and vocabulary V, a topic t extracted by LDA is multinomial distribution over words. Usually top k probable words are considered as representative keywords for topic t.

Definition 1 (Topic): Topic t is defined by set $T_t = \{(w_{it}, p_{it}) : w_{it} \in V, p_{it} = P(w_{it}|t)\}$ where $P(w_{it}|t)$ is a multinomial distribution over words.

The basic idea behind topic models is that documents are mixtures of topics and each topic is a distribution over words.

Definition 2 (Topic Ontology): A topic ontology $O = (C, R)$ is a *Directed Acyclic Graph (DAG)*, where each vertex $c_i \in C$ represents a concept and each edge $r_{ij} \in R$ (from c_i to c_j) represents that concept c_j is a subtopic for concept c_i. Unlike a tree structure, an ontology may have more than one root (i.e., a node without parents). The root nodes can be considered as the most generic concepts.

Given top k keywords for topic t, a topic labeling procedure aims to assign the most semantically meaningful label to topic t.

Definition 3 (Topic Labeling): Suppose $Topics_U = \{T_t : t = 1, 2, ..., n\}$ represents the set of topics produced by the topic modeling method (i.e., LDA), topic labeling produces a set $Topics_L = \{(l_t, a_t) : t = 1, 2, ..., n\}$ where l_t is the label assigned to topic T_t and a_t is a reliability score for l_t.

4 The Proposed Technique

4.1 Labeling Score

We utilized Normalized *PMI* (*NPMI*), which has been used successfully in a wide range of applications [2,3,7,9], to define the coherence score between the top words (based on the probability of the words in the topic) in each topic and the concepts in the ontology. Suppose that p_{it} is the probability of w_{it} (the $i'th$ word in topic t) and c is a concept in the ontology $O = (C, R)$, the coherence score between topic t and concept c can be defined as follows:

$$Score(t, c) = \frac{1}{k \ |label(c)|} \sum_{i=1}^{k} \sum_{q \in label(c)} p_{it} \ NPMI(w_{it}, q) \tag{1}$$

where k is the number top words in each topic used for calculating the score. The label for topic t can be calculated as follows:

$$l_t = \arg\underset{c \in C}{max} \ Score(t, c) \tag{2}$$

Intuitively, we use the probability of each word in a topic as the weight in computation of semantics similarity between topics and ontology concepts. While LDA explores the co-occurrence between the words in a particular topic, the proposed score tries to measure the co-occurrence between the words in a topic and the words in a particular ontology.

4.2 Ontology-Based Specificity (OBS)

Intuitively, the specificity of a concept can be defined based on the ontology structure. Suppose $G = (C, R)$ is an ontology and $c \in C$ is a concept reachable from some roots by following a set of edges in R. We define Ontology-Based Specificity (OBS) of concept c as:

$$OBS(c) = \frac{1}{|R|} \sum_{i \in R} depth_i(c) \tag{3}$$

where $depth_i(c)$ is the length of path from root i to concept c added to 1 (OBS of each root node equals to 1).

4.3 Ontology-Based Coherence (OBC)

Assume $G = (C, R)$ is an ontology and $P(t) \subseteq C$ is the set of top m labels (i.e., concepts) for topic t (that is, for each $c \in P(t)$, $Score(t, c)$ is among the m highest scores for t). We define the cohesion of topic t based on the specificity of the most specific common ancestor of its top m labels in the ontology. The Ontology-Based Coherence (OBC) of topic t is defined as:

$$OBC(t) = \max_{cp \in CP(t)} OBS(cp) \tag{4}$$

where $CP(t)$ is a set of ancestors that are common among the members of $P(t)$:

$$CP(t) = \bigcap_{p \in P(t)} Ancestor(p) \tag{5}$$

Note that in case of $CP(t) = \emptyset$, the OBS measure is defined as 0.

4.4 Average NPMI

We use the average NPMI among the top k words in a topic as the internal measure of topic coherence. The average NPMI can be written as follows:

$$AvgNPMI(t) = \frac{1}{k\ (k-1)} \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} Npmi(w_{it}, w_{jt}) \tag{6}$$

where w_{it}/w_{jt} are the *i'th/j'th* words of topic t respectively.

Input: $Topics_U, O = (C, R)$ ▷ $Topics_U$ is a set of unlabeled topics ▷ O is an ontology
Output: $Topics_L$ ▷ $Topics_L$ is a set of labeled topics

1: **function** TopicLabeling($Topics_U, O$)
2: $Topics_L \leftarrow \emptyset$
3: **for** each topic $t \in Topics_U$ **do**
4: Calculate $Score(t, c)$ using Equation (1)
5: $l_t \leftarrow \arg\max_{c \in C} Score(t, c)$ ▷ Label assignment
6: $P(t) \leftarrow top\ m\ concepts\ c\ \in\ C\ based\ on\ Score(t, c)$
7: $f_1(t) \leftarrow OBC(t)$ ▷ (Equation 4)
8: $f_2(t) \leftarrow Avg\ NPMI(t)$ ▷ (Equation 6)
9: Calculate a_t using Equation (7) ▷ Label accuracy prediction
10: $Topics_L \leftarrow Topics_L \cup \{(l_t, a_t)\}$
11: **end for**
12: **Return** $Topics_L$
13: **end function**

Fig. 1. Proposed topic assignment algorithm

4.5 Building Predictive Model

We have defined two coherence measures as quality measures for a topic. In this section, we show how these measures can be combined to estimate the reliability of a topic and its label. Let $f_1(t) = OBC(t)$ and $f_2(t) = AvgNPMI(t)$. We can define the reliability score of the assigned label (l_t) to topic t using the linear combination of these features (*i.e.*, $\lambda f_1(t) + (1 - \lambda) f_2(t)$) where $\lambda \in [0, 1]$ can be set by the user to indicate the relative importance of each measure.

Alternatively, we can use a supervised approach that learns the importance of each measure using a set of training data if it is available (which is our method). The training data contain a set of topics, their labels, their OBCs, their $AveNPMI$s and the average user evaluation scores (given to each topic by a group of users, ranging from 0 to 1). We use the following logistic regression model to predict the reliability of the assigned label to topic t:

$$Pr(l_t\ is\ \text{“}reliable\text{”} | w_1, w_2, w_3) = \frac{1}{1 + \exp^{-(w_1 f_1(t) + w_2 f_2(t) + w_3)}} \tag{7}$$

The model parameters can be learned from the training data using the maximum likelihood estimation method [10].

Figure 1 illustrates the overall algorithm of our proposed method. Step 5 determines the best label for topic t according to score calculated in (step 4). In step 6, we find the top m concepts ($P(t)$) in the ontology for topic t, then the OBC and average PMI for topic (t) are calculated respectively (step 7–8). Finally, the accuracy of predicted label is calculated using the trained predictive model described by (7). In particular, our proposed labeling approach not only assigns a label to each topic but also estimates the reliability of labels.

5 Empirical Evaluation

5.1 Dataset and Experimental Setup

For evaluation, we use Dow Jones Intelligent Indexing System[1] as the ontology. The structure of Dow Jones ontology is similar to what is described in Definition 2. The corpus used in our experiments is 359145 on-line news published by The Globe and Mail[2] newspaper between the year 2010 to 2014. In the user-study, 5 users are hired and asked to classify each topic into tree levels (i.e., 1, 0.5 and 0 correspond to *good*, *neutral* and *bad*). For the predictive model training, average rates greater than/less than or equal to 0.5 are considered as good/bad labels respectively. Topics are extracted by Mallet software [11] using default parameter setting (k=19) and for OBS measure we considered top 3 words (m=3). The performance of predictive model evaluated using 10-fold cross validation method.

Table 1. Top-10 words of first five topics extracted from the globe and mail dataset

Topic 1	drug, health, drugs, people, disease, treatment, cases, hiv, virus, flu
Topic 2	debt, bond, bonds, credit, investors, interest, government, rates, yield, yields
Topic 3	event, events, royal, day, king, queen, prince, wedding, visit, couple
Topic 4	people, don, ve, lot, things, make, thing, kind, ll, doesn
Topic 5	hotel, travel, park, mountain, beach, resort, trip, town, visit, tourism

Table 2. Top labels of topics illustrated in Table 1

Human Rate	Avg NPMI	OBC	1'st Label	2'nd label	3rd label
0.90	**0.27**	**0.13**	Health	AIDS/HIV	Hepatitis
1.00	**0.24**	**0.38**	Debt	Corporate Debt	Government Debt
0.80	0.70	0.00	Calendar of Events	Festivals	Dance/Ballet
0.40	0.09	0.00	People Profile	Society/Community/Work	School
1.00	**0.21**	**0.25**	Travel	All Inclusive Resorts	Cruises

[1] http://new.dowjones.com/dj-intelligent-indexing/.
[2] http://www.theglobeandmail.com.

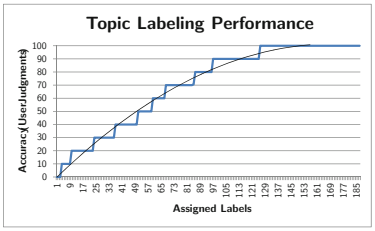


Fig. 2. Users judgments on assigned labels

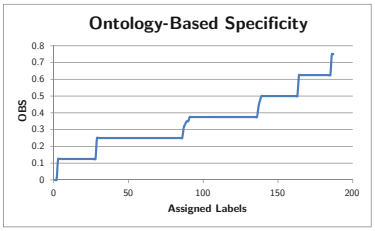
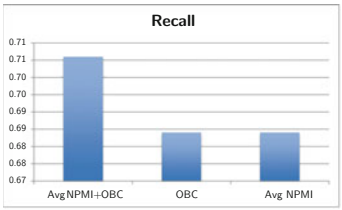
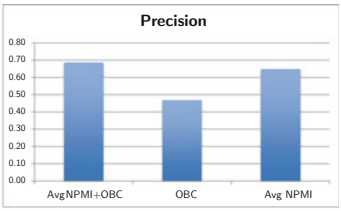


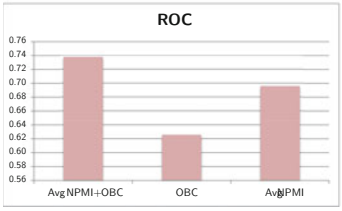
Fig. 3. OBS of assigned labels



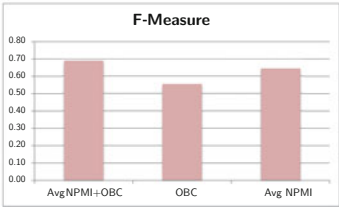
(a) Recall



(b) Precision



(c) ROC



(d) F-measure

Fig. 4. Performance of proposed predictive model

5.2 Experimental Results

Table 2 shows the top-3 labels assigned to the topics illustrated in Table 1 respectively. While topics 1, 2 and 5 have reasonably good labels (i.e., Health, Debt, Travel), topics 3 and 4 are not coherent enough and as a result assigned labels are not semantically meaningful. It can be seen that average NPMI and Ontology-Based Coherent (*OBC*) measures are quite appropriate measures to assess the coherence of topics and assigned labels (*OBC* is zero for topic 3 and 4) and are correlated to average human judgments (column 1). Figure 2 demonstrates the average users judgments (between 0 to 100 % corresponding to *bad* to *good* label) about the labels. We observe that only 27 % of assigned labels have less than 50 % accuracy. Moreover, 30 % of labeled is evaluated by all user as good labels. Figure 3 shows the assigned label specificities distribution according to the *OBS* measure. As it can be seen, proposed method assigns labels with variety of specificities and in fact, the label assignment method does not have a bias towards the

generic labels (generic labels have a good coverage and are hyperonym of many topics, so it is very likely to be chosen by coverage-based topic labeling methods). Figure 4 illustrates the performance of proposed label reliability prediction method based on different criteria. As it can be seen, designed predictive models using internal topic coherence (average NPMI) and external label assignment coherence (OBC) is consistently better than the case when one measure is utilized. While in terms of recall average PMI is slightly better than OBC (Fig. 4a), average NPMI precision is 18 % better than OBC precision (Fig. 4b). Interestingly, in both cases, proposed method (Average NPMI + OCE) is the best (with recall/precision equal to 0.71/0.69 respectively). Moreover, analysis using Roc curve demonstrates that proposed method (with ROC equal to 0.74) provides the best balance between true positive and false positive rates. Similarly, the best F-measure performance (at 0.69) signifies that best trade-off between precision and recall occurs in the proposed method.

6 Conclusions

In this paper, we proposed a novel approach for both topic labeling and labeling reliability prediction with respect to an external ontology. In particular, we defined some measures based on the notion of coherence which can serve as the both labeling and reliability assessment criteria. We showed that the logistic prediction model based on the combination of proposed measures can evaluate the quality of topics and their labels more effectively than using them individually.

Acknowledgement. This research is supported by the Center for Innovation in Information Visualization and Data Drive Design (CIVDDD), a CRD Grant from Natural Sciences and Engineering Research Council of Canada (NSERC) and The Globe and Mail. We thank The Globe and Mail for providing the dataset and ontology used in this research

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 490–499. ACM (2007)
3. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1536–1545. Association for Computational Linguistics (2011)
4. Magatti, D., Calegari, S., Ciucci, D., Stella, F.: Automatic labeling of topics. In: 2009 Ninth International Conference on Intelligent Systems Design and Applications. ISDA'09, pp. 1227–1232. IEEE (2009)
5. Chuang, J., Gupta, S., Manning, C., Heer, J.: Topic model diagnostics: assessing domain relevance via topical alignment. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 612–620 (2013)

6. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272. Association for Computational Linguistics (2011)
7. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, pp. 215–224. ACM (2010)
8. Musat, C., Velcin, J., Trausan-Matu, S., Rizoiu, M.A.: Improving topic evaluation using conceptual knowledge. In: 22nd International Joint Conference on Artificial Intelligence (IJCAI), vol. 3, pp. 1866–1871 (2011)
9. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics (2010)
10. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT Press, Cambridge (2012)
11. McCallum, A.K.: Mallet: a machine learning for language toolkit (2002). http://mallet.cs.umass.edu