

A Latent Topic Model for Linked Documents

Zhen Guo[†], Shenghuo Zhu[‡] Yun Chi[‡] Zhongfei (Mark) Zhang[†], Yihong Gong[‡]
[†]Computer Science Department, SUNY at Binghamton, Binghamton, NY 13905
[‡] NEC Laboratories America, Inc., 10080 N. Wolfe Rd. SW3-350, Cupertino, CA 95014
[†]{zguo,zhongfei}@cs.binghamton.edu, [‡]{zsh,ychi,ygong}@sv.nec-labs.com

ABSTRACT

Documents in many corpora, such as digital libraries and webpages, contain both content and link information. To explicitly consider the document relations represented by links, in this paper we propose a *citation-topic* (CT) model which assumes a probabilistic generative process for corpora. In the CT model a given document is modeled as a mixture of a set of topic distributions, each of which is borrowed (cited) from a document that is related to the given document. Moreover, the CT model contains a random process for selecting the related documents according to the structure of the generative model determined by links and therefore, the transitivity of the relations among documents is captured. We apply the CT model on the document clustering task and the experimental comparisons against several state-of-the-art approaches demonstrate very promising performances.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Retrieval models*

General Terms

Algorithms, Experimentation

Keywords

Topic model, document clustering

1. INTRODUCTION

One of the most fundamental problems in the information retrieval field is to characterize the content of documents. By capturing the essential characteristics in documents, one gives documents a new representation, which is often more parsimonious and less noise-sensitive. Among the existing methods that extract essential characteristics from documents, topic model plays a central role. Topic models extract a set of latent topics from a corpus and as a consequence represent documents in a new latent semantic space. One of the well-known topic models is the Probabilistic Latent Semantic Indexing (PLSI) model proposed by Hofmann [4]. In PLSI each document is modeled as a probabilistic mixture of a set of topics. Going beyond PLSI, Blei et al. [1] presented the Latent Dirichlet Allocation (LDA) model by incorporating a prior for the topic distributions of documents. In these probabilistic topic models, one assumption underpinning the generative process is that the

documents are independent. However, this assumption does not always hold true in practice, because documents in a corpus are usually related to each other in certain ways. Very often, one can explicitly observe such relations in a corpus, e.g., through the citations and co-authors of a paper. In such a case, these observations should be incorporated into topic models in order to derive more accurate latent topics that better reflect the relations among the documents.

In this paper, we propose a generative model, called the *citation-topic* (CT) model, for modeling linked documents that explicitly considers the relations among documents. In our model, the content of each document is a mixture of two sources: (1) the topics of the given document and (2) the topics of the documents that are related to (e.g., cited by) the given document. This perspective actually reflects the process of writing a scientific article: the authors probably first learn knowledge from the literature and then combine their own creative ideas with the learned knowledge to form the content of the paper. Furthermore, to capture the indirect relations among documents, our model contains a generative process to select related documents where the related documents are not necessarily directly linked to the given document. We apply the CT model to the document clustering task and the experimental comparisons against several state-of-the-art approaches demonstrate very promising performances.

2. CITATION TOPIC MODEL

Suppose that the corpus consists of N documents $\{d_j\}_{j=1}^N$ in which M distinct words $\{w_i\}_{i=1}^M$ occur. Each document d might have a set of citations C_d , and so the documents are linked together by these citations.

The CT model assumes the following generative process for each word w in the document d in the corpus.

1. Choose a related document c from $p(c|d, \Xi)$, a multinomial probability conditioned on the document d .
2. Choose a topic z from the topic distribution of the document c , $p(z|c, \Theta)$.
3. Choose a word w which follows the multinomial distribution $p(w|z, \Psi)$ conditioned on the topic z .

As a result, one obtains the observed pair (d, w) , while the latent random variables c, z are discarded. To obtain a document d , one repeats this process $|d|$ times, where $|d|$ is the length of the document d . The corpus is obtained once every document in the corpus is generated by this process, as shown in Fig. 1. In this generative model, the dimensionality K of the topic variable z is assumed known and the document relations are parameterized by an $N \times N$ matrix

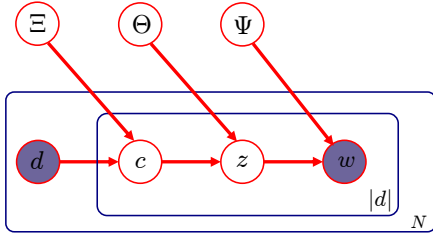


Figure 1: CT model using the plate notation.

Ξ where $\Xi_{lj} = p(c = l | d = j)$, which is computed from the citation information of the corpus.

Following the maximum likelihood principle, one estimates the parameters by maximizing the log-likelihood function

$$\mathcal{L} = \sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j) \log p(w_i | d_j) \quad (1)$$

where $n(w_i, d_j)$ denotes the number of the times w_i occurs in d_j . According to the above generative process, the log-likelihood function can be rewritten as the following equation

$$\mathcal{L} = \sum_{j=1}^N \sum_{i=1}^M n(w_i, d_j) \log \left\{ \sum_{l=1}^K \sum_{h=1}^N p(w_i | z_l) p(z_l | d_h) p(d_h | d_j) \right\} \quad (2)$$

The expectation-maximization (EM) algorithm can be applied to estimate the parameters.

The document relation matrix Ξ is computed from the citation information of the corpus. Suppose that the document d_j has a set of citations Q_{d_j} . A matrix \mathbf{S} is constructed to denote the direct relationships among the documents in this way: $S_{lj} = 1/|Q_{d_j}|$ for $d_l \in Q_{d_j}$ and 0 otherwise, where $|Q_{d_j}|$ denotes the size of the set Q_{d_j} . A simple method to obtain Ξ is to set $\Xi = \mathbf{S}$. However, this strategy only captures *direct* relations among the documents and overlooks *indirect* relationships. To capture this transitive property, we choose a related document by a random walk on the directed graph represented by \mathbf{S} . The probability that the random walk stops at the current node (and therefore chooses the current document as the related document) is specified by a parameter α . According to the properties of random walk, Ξ can be obtained by $\Xi = (\mathbf{I} - \alpha\mathbf{S})^{-1}$.

3. EXPERIMENTAL EVALUATIONS

In this section, we investigate the document clustering task on a standard dataset Cora with the citation information available. Cora [5] contains the papers published in the conferences and journals of the different research areas in computer science, such as artificial intelligence, information retrieval, and hardware. A unique label has been assigned to each paper to indicate the research area it belongs to. These labels serve as the ground truth in our performance studies. In the Cora dataset, there are 9998 documents where 3609 distinct words occur.

By representing documents in terms of latent topic space, topic models can assign each document to the most probable latent topic according to the topic distributions of the documents. For the evaluation purpose, we compare the CT model with the following representative clustering methods.

1. Traditional K-means.
2. Spectral Clustering with Normalized Cuts (Ncut) [6].
3. Nonnegative Matrix Factorization (NMF) [7].
4. Probabilistic Latent Semantic Indexing (PLSI) [4].

5. Latent Dirichlet Allocation (LDA) [1].

6. PHITS [2].

7. PLSI+PHITS, which corresponds to $\alpha = 0.5$ in [3].

We adopt the evaluation strategy in [7] for the clustering performance. The test data used for evaluating the clustering methods are constructed by mixing the documents from multiple clusters randomly selected from the corpus. The evaluations are conducted for different number of clusters K . At each run of the test, the documents from a selected number K of clusters are mixed, and the mixed document set, along with the cluster number K , is provided to the clustering methods. For each given cluster number K , 20 test runs are conducted on different randomly chosen clusters, and the final performance scores are obtained by averaging the scores over the 20 test runs.

The parameter α is simply fixed at 0.99 for the CT model. The accuracy comparisons with various numbers of clusters are reported in Fig. 2, which shows that CT has the best performance in terms of accuracy and the relationships among the documents do offer help in the document clustering.

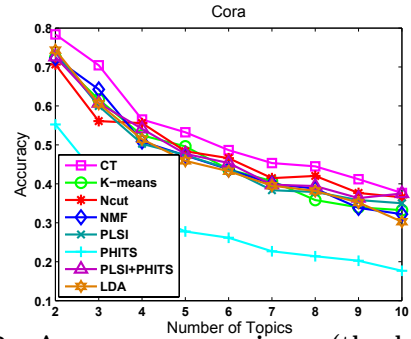


Figure 2: Accuracy comparisons (the higher, the better).

4. CONCLUSION

A novel probabilistic generative citation-topic (CT) model is presented in this paper. The model incorporates the relationships among documents and models each document as a distribution over a set of topics, which is a mixture of the distributions associated with the related documents. The experimental comparisons against state-of-the-art approaches demonstrate very promising performances.

5. ACKNOWLEDGEMENT

This work is supported in part by an internship at NEC Laboratories America, Inc. and the NSF (IIS-0535162, IIS-0812114).

6. REFERENCES

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, pages 993–1022, 2003.
- [2] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *ICML*, pages 167–174, 2000.
- [3] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2000.
- [4] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.
- [5] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163, 2000.
- [6] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.
- [7] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.