Unsupervised Learning from 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. In a traditional topic model which plays an important role in the unsupervised learning, the link information is either totally ignored or treated as a feature similar to content. We believe that neither approach is capable of accurately capturing the relations represented by links. To address the limitation of traditional topic models, in this paper we propose a citation-topic *(CT) model that explicitly considers the document relations represented by links. In the CT model, instead of being treated as yet another feature, links are used to form the structure of the generative model. As a result, 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. We apply the CT model to several document collections and the experimental comparisons against state-of-the-art approaches demonstrate very promising performances.*

I. Introduction

Unsupervised learning from documents is a fundamental problem in machine learning, which aims at modeling the documents and providing a meaningful description of the documents. There has been comprehensive research on the unsupervised learning and the latent topic models play a central role among the existing methods. One of the well-known topic models is the Probabilistic Latent Semantic Indexing (PLSI) model proposed by Hofmann [7]. Going beyond PLSI, Blei et al. [2] presented the Latent Dirichlet Allocation (LDA) model by incorporating a prior for the topic distributions of the 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 observe such relation in a corpus, e.g., through the citations and co-authors of a paper or through the content similarities among documents. In such a case, these observations should be incorporated into the topic model in order to derive more accurate latent topics.

There have been several recent studies that attempt to combine both content and links in a corpus of linked documents. For example, Cohn et al. [4] applied the PLSI model twice, one on content and another on links, and combined the two in a linear fashion. As another example, Zhu et al. [13] fused content and links into a single objective function for optimization. The above studies, however, have one common weak point—they treated links as a feature in a similar way to the content features. Such a *yet-another-feature* approach to treat links ignored two important properties of links. First, links are used to represent relations; and it is the relations represented by the links, not the links themselves, that are important to a topic model. For example, in the previous paragraph we have cited the PLSI and the LDA papers—we have done so not because that our paper is about latent topic models and so we have an obligation to cite these two papers; instead, we have cited these two papers to criticize the prior arts and to position our work. In this sense, it is the *content* of those two papers that play a crucial role, and the citation links are just tokens used to index into those two papers. As a result, we argue that a topic model that treats these tokens (i.e., the links) as features in a peer-level to content is unnatural and questionable.

The second property of links that is ignored by the above studies is that the relations represented by links are often transitive. We again use the paper citation example: the content of a paper is not only likely to be influenced by the papers in its reference list (e.g., the LDA paper [2] that we just cited), but also likely to be influenced, probably to a lesser degree, by the

papers that are not directly cited but cited by the papers in the reference list (e.g., the paper by Nigam et al. on mixture of unigrams, which was cited in the LDA paper but is not cited here for the sake of argument). We believe that a topic model that fails to capture such indirect relation is flawed.

In this paper, we propose a generative model, called the *citation-topic* (CT) model, for modeling linked documents that explicitly addresses the above two properties of links. 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. 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.

The CT model can be applied to a lot of applications such as organizing and indexing documents, classifying or clustering the documents, etc. In this paper, we focus on the document clustering task due to space limitation. We apply the CT model to several document collections and the experimental comparisons against state-of-the-art approaches demonstrate very promising performances.

II. Related Work

PLSI [7] is a well-known topic model towards document modeling which treats the documents as mixtures of the topics and each topic as a multinomial distribution over the words. Beyond PLSI, LDA model [2] is a parametric empirical Bayes model which introduces a Dirichlet prior for the topic distributions of the documents. One difficulty in LDA is that the posterior distribution is intractable for exact inference and thus an approximation inference algorithm has to be considered. Although LDA might obtain a better parameter estimation by treating the topic distributions of the documents as the random variables due to the introduction of a prior, we treat the topic distributions of the documents as fixed parameters for simplicity because we focus on the model itself in this paper. PLSI and LDA led to other variants of topic model [3], [4], [10].

III. Citation Topic Model

The *citation-topic* (CT) model is a generative probabilistic model of a corpus that fully takes into consideration the citation information among the documents. The key feature that distinguishes CT model from the existing topic models is that the relationships among the documents are modeled by another generative process such that the topic distribution of each document

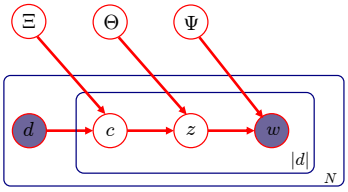

Figure 1. CT model using the plate notation.

is a mixture of the distributions associated with the related documents.

A. Generative Process

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

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

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 document relations are parameterized by an $N \times N$ matrix $\mathbf{\Xi}$ where $\Xi_{lj} = p(c = l|d = j)$, which is computed from the citation information of the corpus. Marginalizing over the latent variables c and z leads to the following distribution

$$p(w|d) = \sum_{c,z} p(c|d)p(z|c)p(w|z) \tag{1}$$

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

$$\mathcal{L} = \sum_{j=1}^{N}\sum_{i=1}^{M} n(w_i, d_j) \log p(w_i|d_j) \tag{2}$$

where $n(w_i, d_j)$ denotes the number of the times w_i occurs in d_j. Substituting Eq. (1) into Eq. (2) results in the following equation

$$\mathcal{L} = \sum_{j=1}^{N}\sum_{i=1}^{M} n(w_i, d_j) \log\{\sum_{l=1}^{K}\sum_{h=1}^{N} p(w_i|z_l)p(z_l|d_h)p(d_h|d_j)\} \tag{3}$$

B. Document Relation Matrix

The document relation matrix $\mathbf{\Xi}$ 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 $\mathbf{\Xi}$ is to set $\mathbf{\Xi} = \mathbf{S}$.

However, this strategy only captures *direct* relations among the documents and overlooks *indirect* relationships. As we have discussed in the introduction,

indirect relationships are also important because of the transitivity of relations. To capture this transitive property, we assume the following generative process for generating a related document c from the given document d_j.

1) Let $l = j$.
2) Choose $t \sim$ Bernoulli(α).
3) If $t = 1$, choose $h \sim$ Multinomial($\mathbf{S}_{.,l}$), where $\mathbf{S}_{.,l}$ denote the l-th column of $\mathbf{S}$; let $l = h$, and return to the step 2.
4) If $t = 0$, let $c = d_l$.

The above generative process can be understood intuitively in terms of a random walk on the directed graph represented by $\mathbf{S}$. The parameter α determines the probability that the random walk stops at the current node (and therefore chooses the topic distribution of the current document). Thus, the indirect relationships among the documents are captured in the above process.

As a result of the above generative process, $\mathbf{\Xi}$ can be obtained according to the following theorem. The proof is omitted due to the space limitation.

Theorem 1. *The probability matrix $\mathbf{\Xi}$ is given as follows*

$$\mathbf{\Xi} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{S})^{-1} \tag{4}$$

When $\mathbf{\Xi}$ is an identity matrix (or equivalently, when $\alpha = 0$), the relationships among the documents are not considered at all and CT reduces to PLSI [7]. Thus, PLSI is a special case of CT when $\alpha = 0$.

IV. Parameter Estimation

The expectation-maximization (EM) [5] algorithm is the standard approach for maximum likelihood estimation in latent variable models. The difficulty when implementing the EM algorithm is that a four-dimensional matrix is required in the E-step because of two latent variables and the current computing power and memory cannot afford it. However, the EM algorithm can be implemented without explicitly performing the E-step by using the nonnegative matrix factorization (NMF) technique. The simple multiplicative rules can be derived formally by converting Eq. (3) to an NMF problem, as shown in the following theorems. The proofs are skipped due to the space limitation. The close relationship between maximum likelihood and NMF is first studied by Gaussier et al. [6].

Theorem 2. *Maximization of Eq. (3) is equivalent to a nonnegative matrix factorization problem under the generalized KL divergence [1]*

$$\begin{aligned} \min_{\mathbf{\Psi},\mathbf{\Theta}} \quad & GL(\mathbf{A}, \mathbf{\Psi\Theta\Xi}) \\ s.t. \quad & \mathbf{\Psi}^\top\mathbf{1} = \mathbf{1}, \mathbf{\Theta}^\top\mathbf{1} = \mathbf{1} \end{aligned} \tag{5}$$

where $\mathbf{\Psi} \in \mathbb{R}_+^{M\times K}$, $\mathbf{\Theta} \in \mathbb{R}_+^{K\times N}$, $A_{ij} = n(w_i, d_j)$, $GL(\mathbf{X}, \mathbf{Y}) = \sum_{i,j}(X_{ij}\log\frac{X_{ij}}{Y_{ij}} - X_{ij} + Y_{ij})$, *and* $\mathbf{\Xi}^\top\mathbf{1} = \mathbf{1}$.

Theorem 3. *The divergence $GL(\mathbf{A}, \mathbf{\Psi\Theta\Xi})$ in Eq. (5) is nonincreasing under the update rules*

$$\begin{aligned} \Psi_{ik} &\leftarrow \beta_k\, \Psi_{ik}\, [\mathbf{B\Xi}^\top\mathbf{\Theta}^\top]_{ik} \\ \Theta_{kl} &\leftarrow \gamma_l\, \Theta_{kl}\, [\mathbf{\Psi}^\top\mathbf{B\Xi}^\top]_{kl} \end{aligned} \tag{6}$$

where $\mathbf{B}$ *is a matrix with the entry* $\mathrm{B}_{ij} = \mathrm{A}_{ij}/[\mathbf{\Psi\Theta\Xi}]_{ij}$ *and* β_k, γ_l *are the normalizing coefficients to make the normalization constraints satisfied. The divergence is invariant under these updates if and only if* $\mathbf{\Psi}$ *and* $\mathbf{\Theta}$ *are at a stationary point of the divergence.*

V. Experimental Evaluations

CT is a probabilistic model towards document modeling, and so it can be applied to lots of applications such as organizing, classifying, clustering, or searching a collection of documents. In this section, we only report the evaluation on the document clustering task due to space limitation. The document clustering task is performed on two corpora: Cora and CiteSeer. Cora [9] and CiteSeer[1] are the standard datasets with citation information available. In the Cora dataset there are 9998 document with 3609 unique words and 10 clusters. We use a subset of CiteSeer which has 2361 documents with 889 unique words and 22 clusters.

A. Evaluation Metrics

The two widely used metrics to measure clustering performance are accuracy (AC) and normalized mutual information (NMI). Suppose that $\mathbf{t}$ and $\mathbf{g}$ are the cluster labels (obtained by a certain clustering algorithm) and the ground truth labels, where $\mathbf{t}_i$ and $\mathbf{g}_i$ are the labels for document d_i. The best mapping function π from $\mathbf{t}$ to $\mathbf{g}$ can be found by the Hungarian algorithm [8]. The accuracy is defined by $\mathrm{AC} = \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{g}_i, \pi(\mathbf{t_i}))$ where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise.

The normalized mutual information which measures the clustering performance from the viewpoint of information theory is defined by $\mathrm{NMI} = \mathrm{MI}(t, g)/\max(\mathrm{H}(t), \mathrm{H}(g))$, where t and g are the random variables corresponding to the cluster distributions of $\mathbf{t}, \mathbf{g}$, respectively; $\mathrm{MI}(t, g)$ is the mutual information between the random variables t and g; $\mathrm{H}(t)$ is the entropy of the random variable t.

One disadvantage of NMI is that it only considers the maximum of the entropies and the smaller one does not contribute at all. A more reasonable metric should take into account both entropies. Inspired by the F1 score measure used to measure the classification performance, we propose the Information F1 score (IF1)

[1]http://citeseer.ist.psu.edu/

which is the harmonic mean of Information Recall (IR) and Information Precision (IP).

$$\text{IR} = \frac{\text{MI}(t,g)}{\text{H}(g)} \qquad \text{IP} = \frac{\text{MI}(t,g)}{\text{H}(t)} \qquad \text{IF1} = \frac{2 * \text{IR} * \text{IP}}{\text{IR} + \text{IP}}$$

B. Performance Comparisons

By representing the documents in terms of latent topic space, the topic models can assign each document to the most probable latent topic according to the topic distributions of the documents. To demonstrate how our method improves the clustering performance in comparison to the state-of-the-art clustering methods, we compare the CT model with the following representative clustering methods: K-means, Normalized Cuts (Ncut) [11], Nonnegative Matrix Factorization (NMF) [12], PLSI [7], LDA [2], PHITS [3], and PLSI+PHITS [4].

We adopt the evaluation strategy in [12] 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, are 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.

For all the datasets, α in Eq. (4) is simply fixed at 0.99. Fig. 2 reports the accuracy (first row) and information F1 score (second row) comparisons on three datasets with various numbers of clusters. In summary, the comparisons on the Cora dataset (first column) show that CT has the best performance in terms of accuracy and achieves significant improvements in terms of information F1 score. On the CiteSeer corpus (second column) CT obtains a better performance than other methods in most cases.

VI. Conclusion

The CT model incorporates the relationships among the 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.

Acknowledgements

This work is supported in part by an internship at NEC Laboratories America, Inc. and NSF (IIS-0535162, IIS-0812114, IIS-0956924).

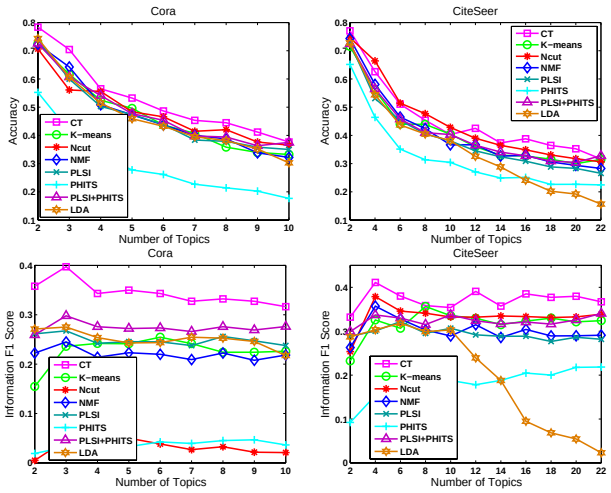

Figure 2. Accuracy and information F1 score comparisons (the higher, the better).

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning)*. Springer, 2006.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, pages 993–1022, 2003.

[3] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *ICML*, pages 167–174, 2000.

[4] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2000.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[6] É. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *SIGIR*, pages 601–602, 2005.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[8] L. Lovasz and M. D. Plummer. *Matching Theory (North-Holland mathematics studies)*. Elsevier Science Ltd, 1986.

[9] 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.

[10] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2004.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[12] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.

[13] S. Zhu, K. Yu, Y. Chi, and Y. Gong. Combining content and link for classification using matrix factorization. In *SIGIR*, pages 487–494, 2007.