Utilizing Citation Context in a Two-Level Topic Model for Knowledge Discovery

Lixue Zou, Li Wang, Xiwen Liu
National Science Library, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
E-mail: zoulx@mail.las.ac.cn, wangli@mail.las.ac.cn, liuxw@mail.las.ac.cn

***Abstract*—Knowledge discovery from academic articles has received increasing attention since full text has been made available by the development of the digital databases. In a corpus of scientific articles, documents are connected by citations and one document has two different parts in the corpus: citation context and autonomous text. We believe that the topic distributions of these two parts are different and related in a certain way. In the existing topic models, little effort is made to incorporate the citation context. In this paper, we propose a citation context topic model which considers the corpus at two levels: cited topic level and citing topic level, utilizing citation context extracted from the full text. Each document has two different representations in the latent topic space. We apply our model to a dataset of PubMed Central, where the full text is available from the XML data. The results clearly show that the citation context can help to discover the latent two-level topics and demonstrate a very promising knowledge discovery capability.**

Keywords-Topic model; Citation context; Knowledge Discovery; XML data.

I. INTRODUCTION

Proliferation of large electronic document collections in the recent past has posed several interesting challenges in knowledge discovery. Latent topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [1] and Latent Dirichlet Allocation (LDA) [2], have become very popular as completely unsupervised techniques for topic discovery in large document collections. These approaches model the co-occurrence patterns present in text and identify a probabilistic membership of the words and the documents in the lower-dimensional topic space [3].

Then, variants of PLSA and LDA allow incorporating more aspects of articles, and here we consider the citation information. As an extension of PLSA, Probabilistic Hypertext-Induced Topic Selection (PHITS) [4] proposed a topical clustering of citations in a manner similar to the topical clustering of words proposed in PLSA, while PLSA-PHITS [5] performed a simultaneous modeling of the citations associated with word occurrences. The Bayesian version of PHITS was proposed as mixed membership model and linked-LDA [6]. In addition, the Citation Network Topic Model (CNTM) [7] presented a non-parametric extension of a combination of the Poisson mixed-topic link model and the author-topic model.

There is also existing work on modeling citation influence. The Copycat and the Citation Influence Model (CIM) [8] introduced the influence parameter to determine how the cited papers are blended into the citing document. The Pairwise-Link-LDA model combines the ideas of LDA and Mixed Membership Block Stochastic Models, while the Link-PLSA-LDA model combines the LDA and PLSA models, assuming that the link structure is a bipartite graph [9]. Additionally, similar models include the Inheritance Topic Model (ITM) [10], the Bi-citation-LDA [11], the Bernoulli Process Topic (BPT) model [12], etc.

Although current citation related topic models are quantitatively successful in clustering the citations and in identifying the citation influence and transitive property, they overlook how those documents influenced the content of this document. That is, the process of incorporation of the citation information ignores the citation context in which that citation appeared in the document.

In our work, we present a two-level topic model utilizing the citation context in a document to discover the latent topic, called the citation context topic model. We define the citation context for a cited document as a bag of words that contains a certain number of words appearing before and after the citation's mention in the citing document. These words can help identify the major topics in the cited document. Moreover, the citation context does not necessarily portray the entire content of the cited document, but provides a description from the authors' perspective in relation to the citing document's topic. This allows us to identify both the cited topics and the autonomous topics.

The rest of this paper is organized as follows: In Section II, we describe the model and the data we dealt with for analysis. Then, Section III gives the experiments and results. Finally, in Section IV, we summarize this paper and prospect our future plans.

II. METHODS

We assume that when the authors write an article, they often reuse ideas and techniques from references, and then based on the inherited thoughts, they generate their own innovative ideas. Thus, each document can be separated into two parts, the citation context and the autonomous text. Moreover, the citation context can be reflected by the cited sentences, while the autonomous text is composed of words that appear outside the citation context. In other words, the topics of each document include two parts: the "citing

 ISBN: 978-1-61208-715-3

topics" from the document itself and the "cited topics" from the citations. Further, we let the citation context or the autonomous text choose to "generate" the topic of a word in the autonomous text by incorporating the Bernoulli distribution into the model, to handle the associations among the autonomous text and the citation context.

Our model assumes the following generative process for each document in the corpus: (1) for the citation context, choose a topic from the multinomial distribution of cited topic conditioned on the document, where the distribution parameter is drawn from a Dirichlet distribution; (2) for the autonomous text, toss a coin s ~ Bernoulli (λ), then if s=0, choose a topic from the multinomial distribution of cited topic; if s=1, choose a topic from the multinomial distribution of citing topic, where the distribution parameter is drawn from a Dirichlet distribution; (3) for each topic, choose a word which follows the multinomial distribution conditioned on the topic with the distribution parameter drawn from a Dirichlet distribution.

Similar to LDA, we also need to infer the posterior probability. Considering that the Markov Chain Monte Carlo sampling methods, such as Gibbs sampling, come with a theoretical guarantee of converging to the actual posterior distribution and the recent advances that make its fast computation feasible over a large corpus, we utilize Gibbs sampling as a tool to approximate the posterior distribution.

We performed our experiments on a dataset of PubMed Central [13], where the full text was extracted from the XML files. The dataset corresponds to brain aging and 246 articles that were cited for more than once were chosen for the test. We extracted one sentence surrounding the citation mentioned in the document as the citation context for each cited document.

For the preprocessing, firstly, we used the tokenization and lemmatization to extract and lemmatize words from the citation context and autonomous text, respectively. Then, we filtered out certain words that were stop words, common words and rare words. We define common words as words that appear in more than 80% of the publications, and rare words are words that occur less than 10 times. Finally, the vocabulary size was 2348 unique words.

III. RESULTS

For each paper, we extracted the citation context and its position in the full text. There were a total of 13858 cited documents with 16316 citation sentences in the collection of 246 articles. In these citation sentences, 8673 sentences were labeled with their location in the full text, which mainly contained four types, introduction (including background), methods, results, conclusion (including discussion). As shown in Figure 1, over two thirds of 8673 sentences were located in the introduction part and the conclusion part, at 30 percent and 38 percent, respectively, whereas those in methods and results made up 20 percent and 12 percent, respectively.

Then, we applied our topic model to the dataset with the number of topics fixed at 10. The parameter was also fixed. One main advantage of the model is the capacity of differentiating the two-level topics. For each paper, we can obtain the topic probabilities at the cited topic level and the citing topic level.

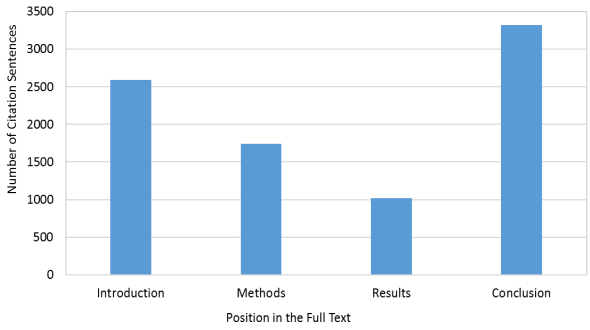

Figure 1. Distribution of citation context in the full text.

Four topics and the top ten words were selected from the output learned by our model, as illustrated in Table Ⅰ. The topic probability conditioned on the dataset has a high value on "memory" and "mitochondrion and damage" at the cited topic level, while "brain structure" and "dementia" have strong probability at the citing topic level.

TABLE I. DETECTION OF TWO-LEVEL TOPICS

Cited topics	Associated words
Memory	hippocampal (0.069), synaptic (0.054), learn (0.045), plasticity (0.034), receptor (0.034), signal (0.023), impairment (0.020), bdnf (0.015), rodent (0.014), channel (0.014)
Mitochondrion and Damage	mitochondrial (0.032), oxidative (0.030), damage (0.022), neurodegenerative (0.020), pathway (0.018), sirt (0.017), dna (0.016), signal (0.012), antioxidant (0.011), neurodegeneration (0.011)

Citing topics	Associated words
Brain Structure	cortex (0.071), pattern (0.038), cortical (0.0354), rest (0.031), connectivity (0.027), refrontal (0.022), stimulus (0.021), lobe (0.019), neural (0.019), gyrus (0.017)
Dementia	clinical (0.048), dementia (0.044), mild cognitive impairment (0.043), risk (0.033), atrophy (0.031), apoe (0.025), mri (0.025), hippocampal (0.024), diagnosis (0.016)

IV. DISCUSSION AND FUTURE WORK

In this paper, we propose a citation context topic model to jointly model the generation process of the autonomous text and citation context for each document to discover the two-level topics. The experiment results demonstrate the effectiveness of the model.

In the future, we will test the efficiency of our topic model on a large collection. Additionally, we want to compare with the state of art topic models and evaluate the likelihood performance and the link prediction task. Furthermore, the investigation of the various applications suggests the promising knowledge discovery capability of this model from the full text, such as topic evolution. Wewill couple our method with other bibliometric methods to portray the inherent dependence among topics in the topic evolution.

 ISBN: 978-1-61208-715-3

ACKNOWLEDGMENT

The research reported in this paper has been supported by the Knowledge Innovation Program of the Chinese Academy of Sciences.

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic analysis," Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999, pp. 289-296.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J Machine Learning Research Archive, 2003, vol. 3, pp. 993-1022.

[3] S. Kataria, P. Mitra, and S. Bhatia, "Utilizing context in generative bayesian models for linked corpus," Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010, pp. 1340-1345.

[4] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," Seventeenth International Conference on Machine Learning, 2000, pp. 167-174.

[5] D. Cohn and T. Hofmann, "The missing link: a probabilistic model of document content and hypertext connectivity," International Conference on Neural Information Processing Systems, 2001, pp. 409-415.

[6] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," Proceedings of the National Academy of Sciences of the United States of America, 2004, vol. 101, pp. 5220.

[7] K. W. Lim and W. Buntine, "Bibliographic Analysis with the Citation Network Topic Model," Proceedings of the Sixth Asian Conference on Machine Learning (ACML), 2014, pp. 142-158.

[8] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," International Conference on Machine Learning (ACM), 2007, pp. 233-240.

[9] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 542-550.

[10] Q. He, B. Chen, J. Pei, B. J. Qiu, P. Mitra, and C. L. Giles. "Detecting topic evolution in scientific literature: how can citations help," ACM, 2009, pp. 957-966.

[11] L. Huang , H. Liu, J. He, and X. Y. Du, "Finding Latest Influential Research Papers Through Modeling Two Views of Citation Links," Asia-pacific Web Conference, 2016, pp. 555-566.

[12] Z. Guo, Z. M. Zhang, S. H. Zhu, Y. Chi, and Y. H. Gong, "A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks," IEEE Transactions on Knowledge & Data Engineering, 2014, vol. 26, pp. 780-794.

[13] PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/ [accessed May, 2019]

 ISBN: 978-1-61208-715-3