CitationLDA++: an Extension of LDA for Discovering Topics in Document Network

Thuc Nguyen
University of Information Technology
Ho Chi Minh, Vietnam
thucnt@uit.edu.vn

Phuc Do*
University of Information Technology
Ho Chi Minh, Vietnam
phucdo@uit.edu.vn

ABSTRACT

Along with rapid development of electronic scientific publication repositories, automatic topics identification from papers has helped a lot for the researchers in their research. Latent Dirichlet Allocation (LDA) model is the most popular method which is used to discover hidden topics in texts basing on the co-occurrence of words in a corpus. LDA algorithm has achieved good results for large documents. However, article repositories usually only store title and abstract that are too short for LDA algorithm to work effectively. In this paper, we propose CitationLDA++ model that can improve the performance of the LDA algorithm in inferring topics of the papers basing on the title or/and abstract and citation information. The proposed model is based on the assumption that the topics of the cited papers also reflects the topics of the original paper. In this study, we divide the dataset into two sets. The first one is used to build prior knowledge source using LDA algorithm. The second is training dataset used in CitationLDA++. In the inference process with Gibbs sampling, CitationLDA++ algorithm use topics distribution of prior knowledge source and citation information to guide the process of assigning the topic to words in the text. The use of topics of cited papers helps to tackle the limit of word co-occurrence in case of linked short text. Experiments with the AMiner dataset including title or/and abstract of papers and citation information, CitationLDA++ algorithm gains better perplexity measurement than no additional knowledge. Experimental results suggest that the citation information can improve the performance of LDA algorithm to discover topics of papers in the case of full content of them are not available.

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; Distributed programming languages; • **Information systems** → **Clustering**; **Document topic models**; • **Mathematics of computing** → *Gibbs sampling*; • **Applied computing** → *Document analysis*;

KEYWORDS

text mining, topic model, citation network analysis, document network, distributed computing

ACM Reference Format:
Thuc Nguyen and Phuc Do. 2018. CitationLDA++: an Extension of LDA for Discovering Topics in Document Network. In *The Ninth International Symposium on Information and Communication Technology (SoICT 2018), December 6–7, 2018, Danang City, Viet Nam.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3287921.3287930

*Corresponding Author

SoICT 2018, December 6–7, 2018, Danang City, Viet Nam

ACM ISBN 978-1-4503-6539-0/18/12…$15.00
https://doi.org/10.1145/3287921.3287930

1 INTRODUCTION

Latent Dirichlet Allocation (LDA) model [2] is the most popular model which is used to discover latent topics from text corpus. We have witnessed success of applying LDA in text mining for the last decades. LDA was also used in discovering research topics from corpus of papers. However, inferring topics from papers becomes critical task because of exponential increasing in amount of publications nowadays. Beside that, we cannot always get full content of papers but only have titles and abstracts of them together with citation information.

The LDA model and probabilistic topic models are based on word co-occurrence to infer latent topics of documents. Therefore, they cannot solve problems very well with short texts. To tackle this problem, most of LDA variants used extra knowledge to overcome limit of word co-occurrence information. Recent works are: Citation-LDA [12] which considered documents as bag of citation or Source-LDA [13] which used prior knowledge source.

Usually when writing an article, the authors will cite articles related to their research topic. Consequently, the use of citation information can provide more on article topics in addition to the content of the article itself. To overcome these inherent weaknesses and keep the advantages of both strategies, we propose a novel method CitationLDA++ which is capable of combining the textual content of the article and citations to address the short text problem in the LDA. Specifically, the CitationLDA++ model will use citation information in the process of labeling threads for words to make labeling more directional, which will help to speed up the process of Gibbs sampling.

The main contributions of our paper are:

- Propose unsupervised CitationLDA++ model to infer topics from papers leveraging title and/or abstract of papers and citation information.
- Suggest a method to incorporate additional knowledge source and citation network to enhance the Gibbs Sampling process for short text.
- Conduct the experiments using AMiner dataset on distributed computing environment.

The remaining of this paper is organized as follows:

- Section 2 - Related work section shows current trend of research to solve problem of the LDA.
- Section 3 - Background and Concepts section presents basic concepts and algorithm of the LDA model together with citation network.
- Section 4 - Method section presents our proposed approach to incorporate textual data and citation network in inferring hidden topics from short text.
- Section 5 - Experiments and Discussion
- Section 6 -Conclusion and Future work

2 RELATED WORK

Topic model is a machine learning method which help to discover themes of a collection of documents. Latent Dirichlet Allocation (LDA) [2] is the most popular topic model which is used in modeling text. Although the LDA model has been successfully used in many fields, there are many variants of the LDA model to incorporate more related information of tasks or to overcome its weakness especially for short text document. In this section, we briefly review several approaches to enhance LDA for short and linked text.

One of recent works which jointly models the document citation network and text content is the relational topic model [3]. The relational topic model aims to model data which are collections of words, and links between them. It tries to construct the latent structure that represents both the words of the documents and how they are connected. Besides the ability of summarizing documents, the model can be used to predict links between them. After that, H. Xia et al. [14] proposed Plink-LDA which uses words in a paper together with the ones in citation papers to calculate probability to re-assign topic of words. This approach uses citation information to get more words to infer word topic in a specific paper. In BPT model [7], the topic distributions of the documents are modeled at two levels (document level and citation level). While the Plink-LDA models topics of documents in one process, the BPT uses two separate generative processes. It firstly generates topics at citation level. Next, the topics at document level are modeled by leveraging the topics at citation level together with words of the document.

For the purpose of analyzing citation network to detect research topics and research themes, Wang et al. [12] proposed Citation-LDA which represents a research paper by a "bag of citations" and model such a "citation document" with a probabilistic topic model. This model represents a document as a mixture of topics and a topic as a mixture of citations. Because the Citation-LDA model uses citations as an observed variable, it can detect papers related to a specific topic.

In addition to works which aim to incorporate extra information to original LDA for specific task, there are also studies which want to address the weakness of LDA for short text. Twitter Topic Modeling by Tweet Aggregation [10] is a recent work in this approach. To tackle the problem with short text, the authors apply several pooling techniques to aggregate similar tweets into a larger documents based on sharing authors and hashtags. After that, these larger documents are fed into LDA algorithm to detect hidden topics. This method is considered to increase the topic coherence of topics which are discovered from short text.

Recent improvement on LDA is Source-LDA [13] which uses prior knowledge source to enhance original LDA. In this model, an existing labeled knowledge source is used to represent known potential topics in the form of frequency of words. Source-LDA incorporates prior knowledge to guide the topic modeling process to improve both the quality of the resulting topics and of the topic labeling. This approach makes the topic inference process consistent with existing knowledge. The main drawback of Source-LDA is that it is a semi-supervised algorithm comparing to unsupervised one in original LDA.

In this work, we propose a novel model, CitationLDA++, which tries to use both the idea of using prior knowledge to guide Gibbs sampling process in [13] and link information to address problem of short text in [10]. Experimental results show that CitationLDA++ gains better performance on perplexity measure.

3 BACKGROUND AND CONCEPTS

3.1 Latent Dirichlet Allocation

Topic model is a common approach to model textual data at document level. A document is assumed to draw its vocabulary from one or more topics. Topics are represented as probability distributions over the vocabulary, where differing topics give different words high probabilities. In other words, topic models are built around the idea that semantics of our documents are actually being governed by some hidden, or "latent", variables. As a result, the goal of topic modeling is to uncover these latent variables - topics - that shape the meaning of our document and corpus. LDA [2] is the most popular model of topic models. In LDA model, with the given parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{1}$$

The key inferential problem of LDA is Equation (1) cannot be computed directly. Therefore, the inference process of LDA is usually be conducted by Gibbs sampling which is an MCMC sampling method in which we construct a Markov chain which is used to sample from a desired joint (conditional) distribution [2]. Gibbs sampling assumes that we can compute conditional distributions of one variable conditioned on all other variables and sample exactly from these distributions [6].

The first step of the Algorithm 1 is to go through each document and randomly assign each word in the document to one of the K topics. This random assignment is already both the topic representations of all the documents and word distributions of all the topics, although it is not very good ones. The main work of Gibbs sampling process is improving these distributions by repeating to re-assign topic to the words until convergence as in Algorithm 1.

According to [6], we can approximate $\mathbb{P}(x_i|x_{-i}, y)$ in Algorithm 1 basing on Bayesian law as in Equation (2):

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{i,j}^{(w_i)} + \beta}{n_{i,j}^{(\cdot)} + \mathrm{W}\beta} \times \frac{n_{i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \tag{2}$$

where:

Algorithm 1: Original Gibbs Sampling Process.

Procedure ***Gibbs Sampling***
  Randomly initialize x_1,x_2,...,x_n.
  while *Not converged* **do**
    for *i=1, 2, ..., n* **do**
      Draw $x \sim \mathbb{P}(x_i|x_{-i}, y)$
      Fix $x_i = x$
    end
    $x^{(k)} = (x_1, x_2, ..., x_n)$
  end
end

- $P(z_i = j)$: the probability that i^{th} word is assigned to topic j.
- z_{-i}: topic assignments of all other words.
- $n_{i,j}^{(w_i)}$: the number of instances of word w_i assigned to topic j.
- $n_{i,j}^{(\cdot)}$: the total number of words assigned to topic j, excluding the current one.
- d_i: document containing w_i.
- $n_{i,j}^{(d_i)}$: the number of words from document d_i assigned to topic j, excluding the current one.
- $n_{-i,\cdot}^{(d_i)}$: the total number of words in document d_i, excluding the current one.
- W: the total number of words in the corpus.
- T: number of topics, equivalent of the K we defined earlier.

The Gibbs sampling algorithm requires large documents to work effectively because it is based on word co-occurrence to infer word-topic and topic-document distributions. Besides that, the sampling process of original Gibbs sampling will choose a random topic which have probability above a random probability to re-assign topic of current word. This makes Gibbs sampling process need a great number of iterations to be converged. In this work, we want to propose a novel method to enhance Gibbs sampling algorithm to work effectively with linked and short textual documents. Deriving from original Gibbs sampling, we propose to use citing topics of original paper to guide the re-assignment process in the Methodology section.

3.2 Measuring Topic Similarity

In probabilistic topic model approach, topics are represented as probability distributions over vocabulary W and captured by the matrix θ. Therefore, if we want to measure the similarity of any two topics, we can directly compare their word distributions from θ. The Kullback-Leibler (KL) divergence, a distance measure of two probability distributions, is often used to make such comparisons. However, KL divergence is not a metric because it does not satisfy neither the triangle inequality nor symmetry. Instead, Blei .et al used Hellinger distance to measure document similarity because Hellinger distance can not only measure the difference between two probability distributions as KL divergence but also is a metric [1].

In this work, we use Hellinger distance, a form of f-divergence, to measure the similarity between two topic distributions as in [9]. For two discrete distributions $P(p_1, ..., p_k)$ and $Q(q_1, ..., q_k)$, the Hellinger distance H is defined as Equation (3).

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \tag{3}$$

In the next section, we will use Hellinger distance to compare word-topic distribution of topics in prior knowledge with the one in the current model. So that, we can map the topics from prior knowledge to current topics. This action help us to incorporate prior knowledge to guide Gibbs sampling process.

3.3 Citation Network & Document Network

According to Egghe .et al [5], citation network is a directed graph where node is a paper (document) and link is a citation relationship. In this way, Chang .et al called linked documents from a collection D a document network [3]. A document network consists of a collection of words together with links between them as in Figure 1. In our corpus, each node contains title and abstract of papers and each edge is citation relation between them.

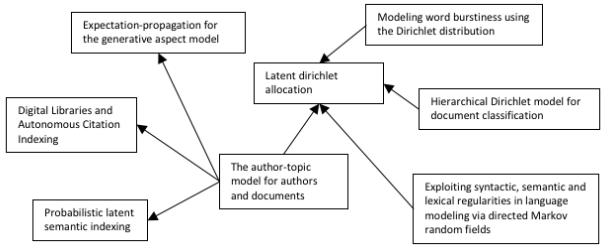

Figure 1: Document network.

From the graph theoretic perspective, citation networks can be considered as directed graphs with research paper in each node and citation relation for each edge:

- **Nodes:** research papers;
- **Edges:** citations of papers;
- **Edge directions:** in order to represent the direction of information flow, we denote the direction of edges from citing papers to cited papers;

Given a citation network as a directed graph G (V, E) where V is a set of nodes (papers), E is a set of directed links (citing relation). For each node (paper) n of V, we define the concept CitingSet(n) as follows:

$CitingSet(n) = \{m \in V |$ there is a direct link from n to m$\}$.

The CitingSet(n) is a set of papers to which paper n cites directly. They are papers which share the same concerns with the original paper. Figure 1 represents document network which are going to be used as data for CitationLDA++. With this definition,we assume that CitingSet of a paper can reveal its interested topics. This assumption is used to build CitationLDA++ model in Methodology section.

4 METHODOLOGY

In this section, we propose a novel model, CitationLDA++, which is an extension of the LDA generative model. Our method firstly

use original LDA to build prior knowledge from research papers. In the second stage, CitationLDA++ leverages citation information of document network to get topics of cited papers and use these information to guide the Gibbs sampling process. The relevant terms and concepts used in the following discussion are defined below.

4.1 Problem Definition

The LDA model represents each document as a bag of word. Each word is an element of $Dic = \{w_1, w_2, ..., w_V\}$ with $V = |Dic|$ and $D = \{d_1, d_2, ..., d_N\}$ is the corpus with $N = |D|$ documents, where each document $d_i = \{w_{i1}, w_{i2}, ..., w_{iq}\}$, $w_{ij} \in Dic$ and $q \leq V$.

Let $T = \{t_1, t_2, t_3, ..., t_K\}$ be a set of K topics which we want to discover from the corpus, where:

$t_i = \{wp_1, wp_2, wp_3, ..., wp_V\}$ with $wp_i \in [0, 1]$ and $\sum_{1}^{V} wp_i = 1$ is called a word distribution of Topic i of the corpus.

With above definition of topics, LDA discovers latent topics in documents and represents each ones as following:

$$TV_i = \{tp_{i1}, tp_{i2}, ..., tp_{ij}\} \tag{4}$$

, with $tp_i \in [0, 1], j \leq K$ and $\sum_{1}^{V} tp_i = 1$.

With the above corpus and integer K as input, the LDA model infers a representation of document as a distribution of K topics and each topic is a distribution of V word using Gibb sampling process as in Algorithm 1.

In order to tackle the problem of original LDA in case of linked short text, we use extra information as follows:

Definition 1 (knowledge base): A knowledge base is a collection of documents before a time in the corpus. The knowledge base is used to generate word-topic distribution by using original LDA which is used as prior knowledge for inferring process in CitationLDA++. If the size of knowledge base is VB and the number of topics is K, the prior knowledge will be a matrix size $(VB \times K)$ with each row is as in (4).

Definition 2 (citing topic distribution): Citing topics is a set of top-k topics which are related to citing set of original paper. For each paper in the corpus, we have a citing topic distribution which is inferred from corresponding citation network as follows:

$$\delta^{(d)} \leftarrow (x_1, x_2, ..., x_k) \tag{5}$$

where $x_i = \frac{N_i}{topk * N_{citing}}$

- N_i: number of papers in CitingSet which related to topic i
- $topk$: number of topics which used to represent a paper.
- N_{citing}: number of citations in document d.
- $\sum_{i=1}^{K} x_i = 1$

4.2 Proposed Model

CitationLDA++ model is an extension of original LDA model with a prior knowledge component to provide meta-data for Gibbs sampling process.

Figure 2 shows elements of CitationLDA++ with extra component showing the prior knowledge. The meaning of these elements is as follows:

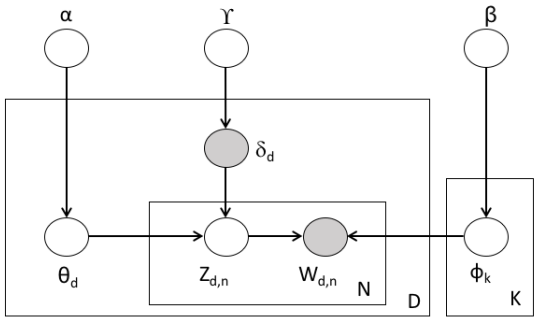

Figure 2: Plate notation for CitationLDA++.

- γ: the top-γ citing topics
- δ_d: citing topics after mapping from topics in prior knowledge to current word-topic distribution.

The remaining elements are the same as in original LDA model [2]. The prior-knowledge block in CitationLDA++ notation is used to provide extra information to approximate the probability of assigning topic j to word w as in (6).

$$P(z_i = j | z_{-i}, w, \delta) \propto \frac{n_{i,j}^{(w_i)} + \beta}{n_{i,j}^{(\cdot)} + \mathrm{W}\beta} \times \frac{n_{i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \times \left(\delta_j^{(d_i)} + \epsilon\right) \tag{6}$$

Equation (6) is derived from Equation (2) with an extra component to incorporate the citing topics distribution with original one, where:

- $\delta_j^{(d_i)}$: the probability document d_i related to topic j as in (5).
- ϵ: a very small positive number that allows for non-zero probability draws from the Dirichlet distribution.

While the probability of assignment word w_i to topic j, in original Equation (2), only depends on how much the word w_i related to topic j in entire corpus and how much document d_i related to topic j, Equation (6) uses an extra information, citing topic distribution, to estimate this probability. We propose (6) basing on an assumption that topics of cited papers are relevant to original one. Experimental results show that this assumption is hold because our proposed model get a lower perplexity measurement comparing to previous one in case of short text documents.

4.3 Inference Technique

In this section, we show how to leveraging citation information of document network to infer topics in CitationLDA++. We derive new inference algorithm basing on Gibbs sampling. The Gibbs sampling process of CitationLDA++ is analogous to original one which consists of decrementing the counts associated with a word, sampling the respective new topic assignment for the word, and incrementing the associated counts. However, we use the prior knowledge about topics of papers in the past and citation information to overcome the limit of word co-occurrence of short text.

Our Gibbs sampler is more complicated than original LDA. Firstly, we have to calculate the citing topics distribution of documents d_i using Algorithm 2 and using this information to calculate the probability to re-assign topic for word as in Equation (6).

Algorithm 2: Calculating citing topic distribution

Input: paper id, prior knowledge, document network, top-k and current topic-word distributuions
Output: citing topic distribution of paper
begin
  Get CitingSet of paper from document network
  foreach *paper p in CitingSet* **do**
    PriorList ← top-k topic of p from prior knowledge
    CurrentList ← Mapping(*PriorList*) (using (3))
  end
  Calculate δ from *CurrentList* using (5)
  return δ
end

Algorithm 2 firstly extract CitingSet (defining in Section 3.3) of a specific paper from document network and then get list of citing topics from prior knowledge. A set of citing topics of a papers is called CT_{prior}. However, the topic-word distributions of prior knowledge may differ from the current ones so the next action is mapping the CT_{prior} to current topic-word distributions. Each topic in the CT_{prior} is mapping to the most similar one by using Hellinger distance to compare the prior and current topic-word distribution. Finally, the citing topic distribution δ is calculated using Equation (5).

The second difference of our proposed process is the sampling action. While sampling action in original Gibbs sampling simply gets a random topic which have the probability above a random threshold to re-assign the word, CitationLDA++ sampling process leverages citation network and prior knowledge to get prospective topic list, CitingTopics. Next, CitationLDA++ firstly samples topic from CitingTopics to get topic for re-assignment. If it cannot get desired topic, the normal sample action is executed. The Gibbs sampling algorithm for CitationLDA++ is represented in Algorithm 3. With the two above improvements in sampling process of CitationLDA++, our model get a lower perplexity measurement as shown in experimental results.

Algorithm 3: CitationLDA++ sampling process

Input: words w of document d
and number of topics K
Output: topic assignments z and counts $n_{d,k}, n_{k,w}$ and n_k
begin
  randomly initialize z and increment counters
  $N = |d|$ //number of words in document d
  foreach *iteration* **do**
    Getting citing topic distributiions using Algorithm 2
    for $i = 0 \rightarrow N - 1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic} -= 1; n_{word,topic} -= 1; n_{topic} -= 1$
      for $k = 0 \rightarrow K - 1$ **do**
        Calculate $p(z = k|\cdot)$ using Equation (6)
      end
      $topic \leftarrow$ -1
      $topic \leftarrow$ sample from $p(z = k|\cdot)$ where $k \in CitingTopics$
      if *topic == -1* **then**
        $topic \leftarrow$ sample from $p(z|\cdot)$
      end
      $z[i] \leftarrow topic$
      $n_{d,topic} += 1; n_{word,topic} += 1; n_{topic} += 1$
    end
  end
  return $z, n_{d,k}, n_{k,w}, n_k$
end

5 EXPERIMENTS AND DISCUSSION

In this section, we conduct experiments using AMiner Citation Network Dataset which is available at https://aminer.org/citation [11]. This dataset consists of papers with title, abstract, authors, year, venue, and references. Our dataset is a subset of AMiner dataset which consists of 5,200 papers published in 2011 and before together with about nearly 49,000 citations. To prepare data for experiments, we extract content of papers consisting of title and abstract; and citation information from the dataset. The former information is used as textual corpus for LDA and CitationLDA++ model. The latter is used to build document network which provides extra knowledge for CitationLDA++.

Our experiments have two scenarios:

- The first scenario, discovering 20 hidden topics of 5,200 papers using original LDA and get the perplexity metric for later comparison.
- For the second scenario, we divide the same textual corpus into two dataset basing on publication year. The papers published before 2010 are fed into original LDA model to infer the hidden topics. We use document-topic distribution in the result of the LDA algorithm and document network to build prior knowledge for CitationLDA++. After that, we train the CitationLDA++ model using the generated prior knowledge and papers publish in 2010 and 2011.

For comparison, we conduct the experiments with the same basic parameters of LDA algorithm $\alpha = 0.5$, $\beta = 0.01$ and $K = 20$. These values get the best perplexity for LDA with our data set. For the CitationLDA++ algorithm, we use the third parameter $\gamma = 4$ to get the top-four topics related to each cited papers. The popular measurement to evaluate topic models is perplexity measurement. In probabilistic topic models, the perplexity is used to measure the uncertainty of a model in predicting some text. Therefore, the model with lower perplexity is the better one. The perplexity of the above experiment scenarios are computed as in [4] and are shown in Figure 3.

Experimental results show that document network together with prior knowledge about topics of cited papers can help to better result in inferring topics for short textual documents. In our dataset, we just have the title and/or the abstract of papers. The available textual data cannot provide enough word co-occurrence information for LDA algorithm to work effectively. In the contrast, CitationLDA++ leveraging information citation together with prior

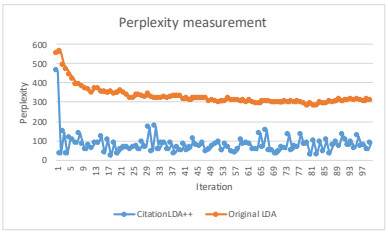

Figure 3: Perplexity of LDA and CitationLDA++.

knowledge about the topics of the cited papers to get meta-data about the topics of original paper. This helps CitationLDA++ to gain a lower uncertainty in predicting the topics.

Our approach is an unsupervised method comparing to a semi-supervised in Source-LDA [13] so our model does not require labeled data which is a limit of many supervised and semi-supervised method. While Plink-LDA [14] and Tweet aggregation [8] apply aggregation approach which using link information to get more words to infer topic, CitationLDA++ uses citation information together with prior-knowledge to tackle the problem of short text documents. The citing topic distribution used in CitationLDA++ provides useful information about topic of words in short text without the need of intermediate aggregated large text as in [14] and [10].

An obvious drawback of CitationLDA++ comparing to original LDA is execution time because our algorithm have an extra execution time to get meta-data in Gibbs sampling process. In big data era, traditional programming approach is overwhelmed. In particular, our implementation needs to process an enormous textual data together with a very large citation network. To overcome the weakness, we use a distributed computing approach, a modern approach in software engineering, to implement CitationLDA++ sampling algorithm. In our experiments, we use Apache Spark, a distributed computing framework [15]. This approach help us in processing document network with GraphX data type which supports graph distributed and parallel computation [16]. Our data model is demonstrated in Figure 4.

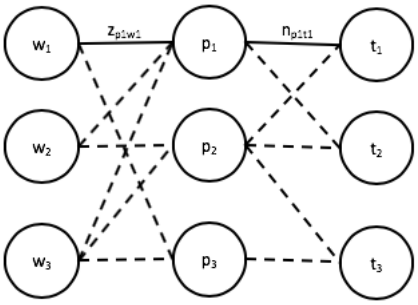

Figure 4: Data model for CitationLDA++ implementation.

In order to implement our algorithm using GraphX, we represent our data using graph structure with three types of nodes: *papers*, *words* and *topics*; and two types of edges: *belonging* and *concerning*. In Figure 4, words w_1, w_2 and w_3 belong to paper p_1, and the paper p_1 is concerned with topics t_1 and t_2. The "belonging"and "concerning"edges have attributes $z_{p_i w_j}$ and $n_{p_i t_j}$ respectively. The attribute $z_{p_i w_j}$ is a number indicating the topic which is assigned to the word w_j in the paper p_i while the number of cited papers of the paper p_i which are related to topic t_j is represented by $n_{p_i t_j}$.

The extreme performance cost of Algorithm 3 is calculating the probability to re-assign the topic for words using Equation (6). With proposed data model, Algorithm 3 can be implemented for very large document networks using distributed computing technology. With GraphX framework, we can distribute and parallelize the CitationLDA++ sampling process using cluster of computers. This implementation approach makes our algorithm available for very large document networks.

6 CONCLUSIONS & FUTURE WORK

Availability of enormous linked short text corpus nowadays leads to the need of novel methods to effective discovery topics automatically by incorporating the text and link information. In bibliographic analysis, citation network is an important source of information which can help to detect papers related to an interested research topic. Therefore, this attribute can help in inferring topics of linked textual document as in research papers. In this paper, we have proposed CitationLDA++ to discover the topics of papers leveraging document network. In order to incorporate citation information into inference process of CitationLDA++, we divide the corpus into two stages basing on publication time. The first one is used to generate prior knowledge using LDA model. The latter, which contains papers citing to the ones in the first, is learning dataset for CitationLDA++. During inferencing process of CitationLDA++, prior knowledge about topic distribution and citation information of a paper is used to approximate the probability of assigning a topic to words and guide the topic re-assignment of the words because interested topics are usually shared between the original and cited papers. This method helps to address the problem of limit word co-occurrence of short text document. Although CitationLDA++ use prior knowledge as Source-LDA [13], our model is unsupervised instead of semi-supervised in the other one. Comparing to Citation-LDA [12], our method incorporate both "bag of words" and "bag of citations" to discover topics of research papers. We conduct experiments with AMiner dataset which contains title/abstract and citation information of papers. Finally, we propose an approach to implement CitationLDA++ using GraphX distributed computing framework to tackle the problem of very large document networks. Experimental results with perplexity measurement show that leveraging citation information can help to address the problem of linked short text documents. We have only conducted experiments with research papers. For the next stage, we will do experiments with other linked textual corpora such as web pages, blogs .etc.

ACKNOWLEDGMENTS

This research is funded by the University of Information Technology -HCM under Grant No.: D1-2018-09.

REFERENCES

[1] David M Blei, John D Lafferty, et al. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1, 1 (2007), 17–35.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[3] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. 81–88.
[4] Kuan-Yu Menphis Chen and Yufei Wang. 2007. Latent dirichlet allocation. (2007).
[5] Leo Egghe and Ronald Rousseau. 1990. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers.
[6] Tom Griffiths. 2002. *Gibbs sampling in the generative model of Latent Dirichlet Allocation*. Technical Report.
[7] Zhen Guo, Zhongfei Mark Zhang, Shenghuo Zhu, Yun Chi, and Yihong Gong. 2014. A two-level topic model towards knowledge discovery from citation networks. *IEEE Transactions on Knowledge and Data Engineering* 26, 4 (2014), 780–794.
[8] Kar Wai Lim, Changyou Chen, and Wray Buntine. 2016. Twitter-network topic model: A full Bayesian treatment for social network and text modeling. *arXiv preprint arXiv:1609.06791* (2016).
[9] Arun S Maiya and Robert M Rolfe. 2014. Topic similarity networks: visual analytics for large document sets. *arXiv preprint arXiv:1409.7591* (2014).
[10] Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. 2017. Twitter Topic Modeling by Tweet Aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*. 77–86.
[11] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnet-Miner: Extraction and Mining of Academic Social Networks. In *KDD'08*. 990–998.
[12] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. 2013. Understanding evolution of research themes: a probabilistic generative model for citations. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1115–1123.
[13] Justin Wood, Patrick Tan, Wei Wang, and Corey Arnold. 2017. Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*. IEEE, 411–422.
[14] Huan Xia, Juanzi Li, Jie Tang, and Marie-Francine Moens. 2012. Plink-LDA: using link as prior information in topic modeling. In *International Conference on Database Systems for Advanced Applications*. Springer, 213–227.
[15] Reynold S Xin, Joseph E Gonzalez, Michael J Franklin, and Ion Stoica. 2013. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*. ACM, 2.
[16] Bo Zhao, Hucheng Zhou, Guoqiang Li, and Yihua Huang. 2018. ZenLDA: Large-scale topic model training on distributed data-parallel platform. *Big Data Mining and Analytics* 1, 1 (2018), 57–74.