Plink-LDA: Using Link as Prior Information in Topic Modeling

Huan Xia[1], Juanzi Li[1], Jie Tang[1], and Marie-Francine Moens[2]

[1] Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200 A B-3001 Heverlee, Belgium
{xiahuan,ljz,tangjie}@keg.cs.tsinghua.edu.cn, sien.moens@cs.kuleuven.be

Abstract. Citations are highly valuable for analyzing documents and have been widely studied in recent years. Among the document modeling, the citations are treated as documents' attributes just like the words in the documents; or as the degrees in graph theory. These methods add citations into word sampling process to reform the document representation but they miss the impact of the citations in the generation of content. In this paper, we view the citations as the prior information which authors have had. In the generation of document, content of the document is split into two parts: the idea of the author and the knowledge from the cited papers. We proposed a prior information enabled topic model-PLDA. In the modeling, both the document and its citations play the important role of generating the topic layer. Our experiments on two linked datasets show that our model greatly outperforms basic LDA procedures on a clustering task while also maintaining the dependencies among documents. In addition, we also show the feasibility by the task of citation recommendation.

Keywords: Topic Modeling, LDA, Links, Prior Information, Plink-LDA.

1 Introduction

In recent years, social network such as Facebook, Twitter are growing rapidly. One important and essential part in these networks is the following relationship. The following relationship plays an important role in user generated contents, for users are strongly influenced by the posts which they follow. It is the posts which produce these comments. If there were no posts, such as "#911" shown in Figure 1, there would have been no comments about these events. Similarly, authors make a reference to other documents while they are writing a paper. People cite papers for they have gained some knowledge from the existed knowledge in the literature as shown in Figure 2. Both following relationships and references indicate not only topical similarity but also dependencies between different documents. Recent studies on how to use these links can be classified into two categories: one is that links are used as document attributes just like word appearances in the documents; the other is that links are used as degrees in graph theory. However, both of them cannot handle the

S.-g. Lee et al. (Eds.): DASFAA 2012, Part I, LNCS 7238, pp. 213–227, 2012.

problem of taking these two features, topical similarity and dependencies, into account simultaneously.

Topic model is a popular strategy to analyze the texts in the documents. Many derivations of topic models have been proposed to meet different requirements on the basis of LDA [1]. Among these models, researchers address the problem of how to integrate the links information into the model [2, 3, 4, 5, 6, 14]. However, links information is treated as attributes of documents in these models. As a result, the dependencies between documents are ignored. We think that both comments and references should be referred to because they have the influence on the generating of the content, so can we use the link information in another way to show both features? Usually, authors make a reference to other documents because what they talk about is closely related to what they cite, whether agreement or disagreement. So can we treat these links information as some kind of prior information to generate the document content? Driven by this, we address the problem of analyzing and using links in a different way. We hope that our new model can explicitly model the citations and words simultaneously and maintain both the topic similarity and dependencies, which makes our contribution in this paper:

1) We use link information which indicates the topical similarity and dependencies between documents as a kind of prior information in the generation of the document.

2) We propose a unified topic model which can model links and word simultaneously to address above problem.

3) We implement three experiments to evaluate the feasibility of our proposed model.

The rest of the paper is organized as follows. We give the formal description of our intuition in section 2. We describe our proposed model Plink-LDA by treating citations or following relationships as prior information in this paper. Plink-LDA can model the citations and content simultaneously. In order to evaluate the feasibility of our model, we do the experiments in section 4 and the experiment results show that our model outperforms the baseline. We discuss related work in section 5. Section 6 concludes the paper.

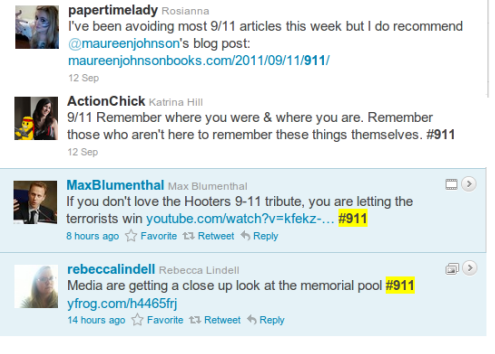

Fig. 1. User generated contents example. User comments on the event 911 ten year anniversaries extracted from twitter. Users express their views on this event. The views can be seen as exactly the reactions to the post #911.

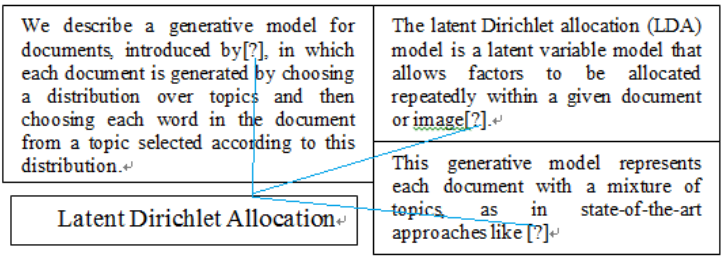

Fig. 2. Citations example. Contexts of papers which cite paper "Latent Dirichlet Allocation". These contexts are mostly introduction to or comparison with LDA. And they are highly related to each other for that they all cite the same paper.

2 Problem Definition

Latent Dirichlet Allocation (LDA) is a probabilistic model to analyze documents in the latent topic layer. It interprets the document generation as word sampling process from topics. For each word w_{di} in document d, a specific topic z_{d_i} is chosen from the document-specific distribution. Then w_{di} is generated according to the topic-specific multinomial distribution $\Phi_{z_{di}}$.

The generating process of LDA is very intuitive: first, the authors choose a topic, which concept the author wants to talk about; second, the authors choose the mostly used words for this topic to constitute the content of the document. However, this process models the word appearance only. In order to analyze the influence of the link, many derivations of LDA are proposed to model links with words simultaneously. All of these models have an underlying common view that integrating citation into model can make document distribution more precisely; for citations reveal some content that is not mentioned by words in the documents. Citations can make up this lost information which cannot be completely represented by words. Therefore, these models tend to place weight on some topics which not mentioned by the document but by its citations. So the documents will be fully represented. In these models, links are just like a special kind of word in the document and they are also generated from the document-specific distribution. Similarly, we formalized our idea in the topic pattern**:**

The author gets information z_i from a reference d_i

And he may have his own idea z after he gets many z_i

According to the mixture of z and z_i, all of the words of the document the author going to write are sampled to the topic distribution.

In this model, links are no longer sampled as results. Instead, they are the prior information of the documents. The citations make a change to the document-specific distribution which eventually reflects our idea. Before giving our model, we give some notations we are going to use in the following sections.

Definition 2.1. [Document]. The content of a document excluding the references. We use d_i to represent the i*th* document in the dataset.

Definition 2.2. [Citation (Link)]. The document's references. If document d cites another document d_i, we call d_i as a citation of document d. d is also noted as citing document and d_i as cited document. Both follow relationship on social network and references of documents are noted as citations here. We use c_i to represent the i*th* citation of a document.

Definition 2.3. [Related Documents]. Those documents which are talking about the same topic. Most parts of them are similar. The differences among them may only take a small part of their contents.

3 Methodology

3.1 1Intuition of Plink-LDA

To illustrate our model, we first look into some details of LDA model. LDA defines the following generating process for every document in a corpus [1]:

1. For each document d, draw a topic distribution $\theta \sim \mathrm{Dir}(\alpha)$;
2. For each word w_i in document d:
   a) Draw a topic $z_n \sim Multinomial(\theta)$.
   b) Sample a word w_i according to the multinomial condition probability distribution $\mathrm{p}(w_i|z_n, \beta)$.

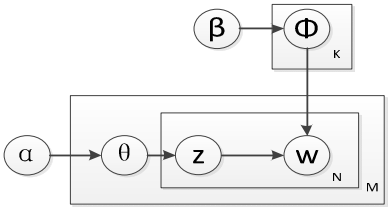

Fig. 3. Plate notation for LDA

α, β are hyper parameters for the specific collection. The probability of generating a word from a document d is:

$$\mathrm{P}(\mathrm{w}|\mathrm{d}, \theta, \Phi) = \sum_{z \in T} P(w|\, z, \Phi_z) P(z|d, \theta_d) \quad (1)$$

LDA model analyzes documents in three layers: word layer, topic layer and document layer. An informal interpretation is that: Documents generate topics and topics generate words. From the graphical representation, we can see that generating of topic layer is controlled only by the document which it belongs to.

Many derivations [4, 5, 7] of LDA integrate citations into the model during the word sampling process and topic layer is still controlled by the document itself in these models. Adding citations into word sampling process do reform the document

representation; however, sometimes these reformed distributions may not produce the expected results. For example, two documents may have same citation list but they talk about totally different solutions to one problem. So some parts of the two documents will be totally different from each other but they may just take a small place of the document content. Although it is these parts which distinguish them, their small content occupation may be ignored by their great similarity in citations.

To express this intuition more clearly, we can split document content into two parts: 1) the idea of the authors; 2) the knowledge learned in the existed literature. But it is not equally reflected in the words of the documents. All of the words are supposed to be related to the first part mentioned above in previous topic models. Citations which the authors refer to are not taken into account. They are just treated as attributes of documents just like word appearances. Actually, many words in the documents are generated by the topic of the citations. In order to reveal this, we assume that the generation for words should be controlled by both the document and its citations, and also the relation should be reflected in the model.

To reflect the intuition that citations have impact on the content constitution of the documents, we propose a model which utilizes citations as prior information. To reveal this change in utilizing citations, we modify the topic sampling process on the basis of LDA. The topic sampling is controlled by both document and its citations. We combine document and its citations' topic distributions together to generate the topic. So the generating of the topic layer is no longer controlled only by the document topic distribution only. Instead, both the document and its citations' document layer play the role of generating the topic layer.

3.2 Our Model

Based on the discussion in section 3.1, we propose the model to address the problem that using citations as prior information in LDA. First we give the plate notation graph and notations in the following:

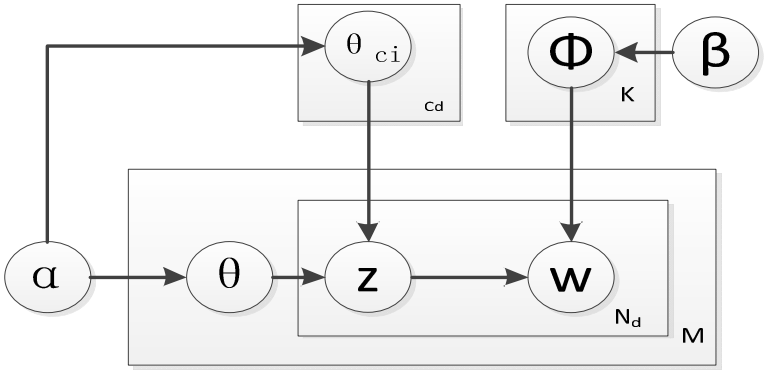

Fig. 4. Plate notation for Plink-LDA

For this model, the generative process is as follows:

1. For each document d, draw a document specific distribution θ.
2. For each word w_{di} in d:
   a) Randomly sample a citation inference c_i, then draw a document specific distribution θ_{c_i};
   b) Combine θ and θ_{c_i} by tuning parameter λ to generate a document distribution θ_c;
   c) Sample a topic z_i according to the combined topic distribution θ_c;
   d) Generate word w_{di} according to $P(w_{di}|z_i, \Phi)$.

As dedicated in Figure 4, topic sampling process has taken the citation into account. To show the influence, we make a linear combination of the document and its citations' topic distribution θ controlled by an tuning parameterλ. For the generating process of 2.c, the combined topic distribution is :

$$\theta_c = (1-\lambda)\theta + \lambda\theta_{c_i} \tag{2}$$

As to sampling a citation inference c_i, we take all the citations into account for the ease and adjust the weights by the parameter λ.

To estimate the parameters for this model, we take the widely used Gibbs sampling procedure to estimate the latent variable. We use the same sampling algorithm as that for LDA model with the posterior probability:

$$P\left(z_{d_i}|z_{d_{-i}}, w, c, \alpha, \beta, \lambda\right) \propto \frac{(1-\lambda n_{z_{d_i}}^{-d_i}) + \lambda n_{c_i} + \alpha_{z_{d_i}}}{\sum_z((1-\lambda)n_{dz} + \lambda n_{cz} + \alpha)} \times \frac{n_{w_i}^{-d_i} + \beta}{\sum_i\left(n_{w_i} + K\cdot\beta\right)} \tag{3}$$

where "-" indicates excluding that instance from counting. The notation is as follows:

Table 1. Notations for our model

Symbol	Description
z_{d_i}	topic i assigned to word w
$z_{d_{-i}}$	topic i assigned to word w excluding current instance
w	current word
c	citations of the document
$n_{z_{d_i}}^{-d_i}$	number of words assigned to topic i in document d excluding instance of word i
n_{dz}	number of words assigned to topic z in document d
n_{c_i}	number of words assigned to topic i in citations of document d
λ	the tuning parameter
$\alpha_{z_{d_i}}$	hyper parameter for each document, $\sum \alpha_{z_{d_i}} = \alpha$
α, β	hyper parameter for LDA model
$n_{w_i}^{-d_i}$	number of words assigned to topic i of all instances of word w excluding this instance
n_{w_i}	number of words assigned to topic i
k	number of word tokens

We notice that the difference in the posterior probability between LDA and our model is whether the instances of citations are counted. The instances of words in citations actually reveal its topic representation. In our model, those words strongly related to the topics of the citations are mainly generated by the citations topic distribution. The document topic distribution is modified to show the difference between its citations and itself. This change in topic distribution is supposed to discriminate the small difference between documents when most part contents of documents are similar. The similar dimensions caused by citations in topic space are removed or slightly reduced. The modified topic distribution in our model mainly focuses on the different parts of its content. As a result, it is capable of distinguishing those documents which are strongly related.

4 Experimental Design

4.1 Datasets

For our experiments, we used two standard linked data sets: Citeseer[1] and Cora[2], to evaluate our model.

Citeseer consists of 3312 scientific publications from six categories: Agents, Artificial Intelligence, Database, Human Computer Interaction, Machine Learning and Information Retrieval. The citation network consists of 4732 links. After stemming and removing stop words, 3703 unique words remain.

Cora is a dataset containing machine learning papers published in the conferences and journals of seven categories: Neural Networks, Rule Learning, Reinforcement Learning, Probabilistic Methods, Theory, Genetic Algorithms and Case Based. For each paper, there is a unique label to indicate which category it belongs to. The Cora dataset subset consists of 2708 scientific publications classified to one of seven classes. There are 5429 citations in the data set. After preprocessing, 1433 unique words remain.

4.2 Tasks and Evaluation

4.2.1 Clustering Performance

In this task, we measure how well our model performs after integrating links as prior information into document modeling. We do the clustering task and compare the results based on our model with LDA in terms of accuracy and recall number.

The experimental set-up is as follows. We first train Latent Dirichlet Allocation model on Citeseer and Cora datasets respectively. We use these model parameter results as our baseline. Then we model these two datasets based on our proposed model iteratively. To observe the impact of tuning parameter, we model the datasets with different tuning values for λ. After this, we utilize the model parameters to automatically cluster the documents in the two datasets. After clustering, we first decide which category these clusters belong to and then we define the accuracy for a cluster as follows:

[1] http://www.cs.umd.edu/~sen/lbc-proj/data/citeseer.tgz

[2] http://www.cs.umd.edu/~sen/lbc-proj/data/cora.tgz

$$cluster\ accuracy = \frac{number\ of\ documents\ belonging\ to\ the\ cluster\ category}{number\ of\ the\ cluster\ \ dcouments} \quad (4)$$

Then we calculate the accuracy for a dataset by combining cluster accuracies with their weights together.

Figure 5 and Figure 6 show the results of clustering on two linked datasets. For Cora and Citeseer, the model parameters, topic number, are set to 7 and 6 respectively. Zero for λ means LDA model without integrating link information.

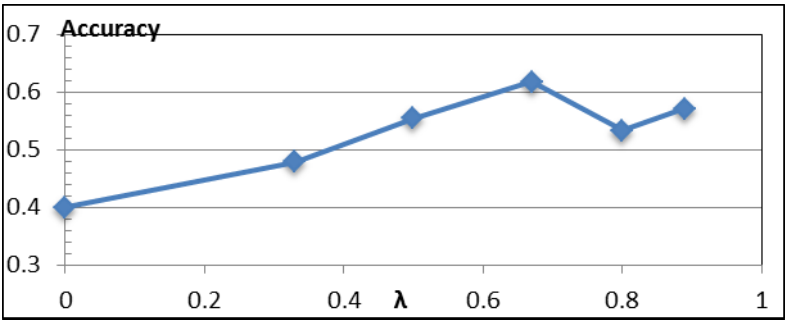

Fig. 5. Accuracy for Cora dataset. λ is set to 0, 0.33, 0.5, 0.67, 0.8,0.89 respectively. These values correspond to different ratios for existed literature and content of the document, which are 2:1, 1:1, 1:2, 1:4, 1:8.

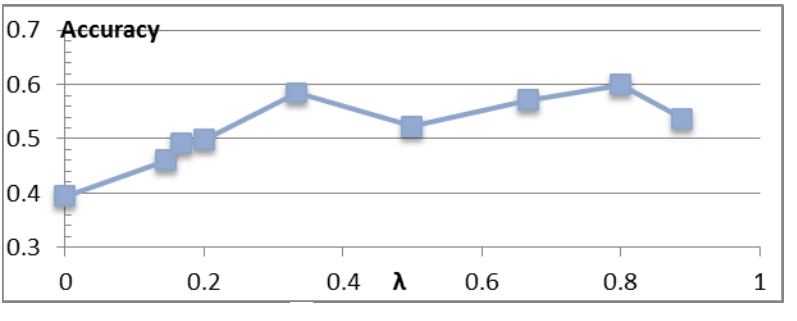

Fig. 6. Accuracy for Citeseer dataset. λ is set to 0, 0.145, 0.167, 0.2, 0.33, 0.5, 0.67, 0.8, 0.89 respectively. These values correspond to different ratios for existing literature and content of the document, which are 6:1, 5:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:8.

As depicted in Figure 5 and Figure 6, our model outperforms baseline greatly in all situations with different tuning parameters. Integrating link information into topic model reforms the document representation in latent semantic space.

Besides this, we can also observe the influence on model by the tuning parameter λ. Our experiments show that different datasets have different best tuning parameter values. We observe that there may exist two best tuning parameter values for a single dataset as shown in Figure 5 and Figure 6. It is interesting for it indicates the research's writing habits. The tuning parameter actually reveals the balance point between individuals and the population. The optimal tuning parameters correspond to the balance points for them. The first optimal value for λ is what we want to have. It indicates to what extent a researcher would gain knowledge from existed literature while doing research and writing papers. This value maintains researchers' individual characteristics and the population's features. For example, the best value for tuning parameter is 0.67 in Cora dataset. In other words, for every three words in the document of Cora, there is only one word that only talks about the authors' ideas without anything for the existed knowledge in the literature [Figure 7]. And for Citeseer, the value is 0.33. It is smaller than that in Cora dataset. Cora dataset consists of papers from machine learning area and it is reasonable for that papers of Cora dataset have more similar topics to each other and they have more coverage of each other. The second optimal value for λ corresponds to a different situation. In this situation, the authors are completely influenced by the existed literature. No innovations take place. To show this in the topic model, all of the documents are generated by the same topic distribution in the latent semantic space. That means all of the documents have the same document specific distributions over topics. Obviously, this is not we want to get for this would remove the diversity for the documents in the dataset.

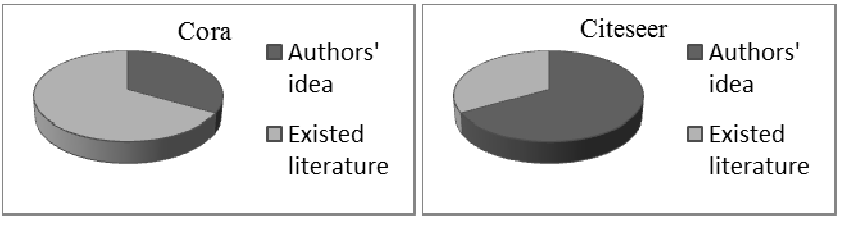

Fig. 7. Best tuning parameter for Cora and Citeseer datasets . The deep gray piece represents the authors' innovations. It indicates how different the author's idea from others works. The best tuning parameter maintains the diversity for the population and the characteristic for the individuals.

Although Cora and Citeseer contain several categories, these categories are also highly correlated to each other. For Cora, all of the documents belong to machine learning area. To see to what extent our model can distinguish them, we list the recall number in Table 2 and Table 3. Here, recall number is defined as the maximum number clustered in one category. Documents in these categories are more closely clustered compared with the procedure without utilizing the citations as prior information.

Table 2. Explicit recall number for Cora. C1 means represents category 1, etc. Row total means how many documents one category has. The star notation indicates the maximum recall number for a category.

	C1	C2	C3	C4	C5	C6	C7
Total	818	180	217	426	351	418	298
$\lambda = 0$	291	72	130	**285***	93	170	151
$\lambda = 0.33$	295	93	154	263	103	300	165
$\lambda = 0.5$	401	94	144	253	221	280	**187***
$\lambda = 0.67$	394	112	**157***	275	**228***	312	183
$\lambda = 0.8$	397	119	126	279	208	329	179
$\lambda = 0.89$	**416***	**123***	131	261	199	**337***	181

Table 3. Explicit recall number for Citeseer. C1 means represents category 1, etc. Row total means how many documents one category has. The star notation indicates the maximum recall number for a category.

	C1	C2	C3	C4	C5	C6
Total	596	668	701	249	508	590
$\lambda = 0$	187	326	210	19	318	246
$\lambda = 0.145$	350	329	211	45	303	288
$\lambda = 0.167$	401	403	409	66	112	235
$\lambda = 0.2$	249	434	397	69	275	226
$\lambda = 0.33$	430	345	406	58	318	**379***
$\lambda = 0.5$	349	**443***	414	20	279	226
$\lambda = 0.667$	374	399	**443***	**71***	349	256
$\lambda = 0.8$	435	393	370	61	**366***	362
$\lambda = 0.889$	**450***	371	267	63	263	366

We can observe that recall number are significantly improved after integrating link information into the model. The bold columns are the significant ones. However, we also find that there are limiting values for recall numbers. This is restricted by the dataset itself. High limiting recall percentage means that documents in the category are highly related to each other and closely located together in the semantic space. Low limiting recall percentage means that the documents in the category cover many topics and are not well classified. Besides, for each category of the two datasets, the best tuning parameters are different. This phenomenon reveals that each category has individual cluster aggregation characteristics.

4.2.2 Perplexity

In this part, we measure how well our model performs in terms of perplexity. Perplexity is an important measurement in information theory. It is a common way of evaluating language models. The lower the perplexity is, the better the model trains the dataset. The perplexity formula is as follows:

$$\text{perplexity} = \sum_{d \in D} 2^{-\sum_{i=1}^{N} \frac{1}{N} \log_2 p(w|d)} \quad (5)$$

where p(w|d) represents the probability of the document generating a specific word:

$$\text{p(w|d)} = \sum_{z \in Z} p(z|d)p(w|z) \quad (6)$$

Formula 5 list calculate the total perplexity for the corpus. To get average perplexity for each document, we have to divide them by dataset size which is 2708 for Cora and 3312 for Citeseer respectively.

Perplexities of the two datasets for all the models in section 4.2.1 are listed in Table 4.

Table 4. Perplexities for Cora and Citeseer datasets under different conditions. The bold values in Table 4 means that it achieve a lower perplexity than baseline. For Cora, we can see that, in all 5 situations, we have lower perplexities. And for Citeseer, there are also 5 situations when we get a lower perplexity. For both datasets, when take the first optimal value for λ, which is 0.667 for Cora and 0.33 for Citeseer, we also get a lower perplexity than baseline.

λ	*Cora*	*Citeseer*
0	223070.4	190888.6
0.145		190972.7
0.167		***189718.6***
0.2		***188660.7***
0.33	***218606***	***190091.4***
0.5	***220083.9***	***189109***
0.667	***219184.8***	***187625.4***
0.8	***218958.2***	191009.1
0.889	***218178.2***	190909.4

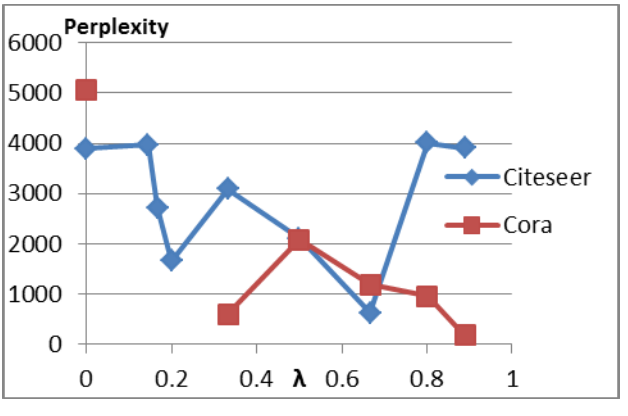

Fig. 8. Perplexities for Cora and CiteSeer datasets. To depict more clearly, we have already minus 218000 for Cora and 187000 respectively from the perplexities.

4.2.3 Citation Recommendation

We also manually evaluate the document recommendation performance of our model. The crucial part of recommending documents is to measure the similarity between documents. For example, we take a paper titled "Modeling Risk from a Disease in Time and Space" from the Cora dataset. This paper is mainly about Bayes network and Markov chain Monte Carlo (MCMC) methods cover most part of it.

Table 5. Example of recommending citations.For the paper titled "Modeling Risk from a Disease in Time and Space" in Cora, several citations recommend by our model are listed.

Modeling Risk from a Disease in Time and Space	
Citations	Recommended documents
1. Bayesian Dynamic Factor Models and Portfolio Allocation	1. On MCMC sampling in hierarchical longitudinal models
2. Bayesian Analysis of Agricultural Field Experiments	2. Exact bound for the convergence of metropolis chains
3. Hierarchical Spatio-Temporal Mapping of Disease Rates	3. A simulation approach to convergence rates for Markov chain Monte Carlo algorithms

We can represent this particular paper, its citations and recommended documents in the following composition chart shown in Figure 9. Usually papers consist of composition 1 style and composition 2 style are strongly correlated due to their highly similar compositions. Composition 3 style is less likely regarded as strongly related to composition 1 by this criterion, although its main part concerns the same topic, such as MCMC in this example. As discussed above, our model slightly removes the common parts, which is the population features, from its distribution. As a result, composition 3 would be more related to the refined distribution of composition 1. The recommended documents listed in the right part of Table 5 are strongly related to MCMC and recommending them as similar documents is reasonable.

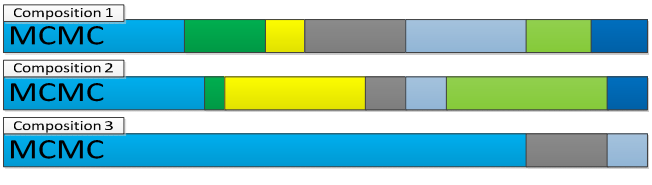

Fig. 9. Different types of compositions for documents. Different colors mean different topics. And the length of box indicate the topic weight in the distribution.

5 Related Work

Research on how to utilize links can be categorized into two groups. The first one is that links are utilized as degrees in graph theory. The other one is that links are treated as attributes of documents just like word appearance in the document. Trevor

Strohman [3] did a survey on the impact of all attributes in the documents. Among these attributes, such as publication year, text similarity, co-citation coupling, Katz distance and citation count, text similarity and Katz distance play the most important part. They are the two key attributes for a document. Much work has been done to integrate these two parts together to help the research.

On the basis of LDA, which analyze the document content in a low latent topic space, many derivations of LDA are proposed to tackle this problem. Cohn and Hoffman [4] proposed an extension to the pLSA [6] model, which called PHITS. Citations are modeled with words simultaneously and they are treated equally. Both of them share the same latent topic distribution. The intuition is that topic related documents have more intersection not only of words but also of citations. So citations or hyperlinks will be helpful in modeling the documents more precisely, which will eventually improve the performance using these distributions over latent topics. Link-LDA model [5] is very similar to PHITS. Erosheva et al developed PHITS by replacing pLSA with LDA. Reference sampling process is exactly the same to word sampling process. Both PHITS and Link-LDA model treat citations as word appearance. The generation process is completely guaranteed by the document specific topic distribution. They are all treated as observations while maximizing the likelihood function. However, documents dependences which revealed by citations are ignored in these models. Therefore, some other models were proposed to address this problem. Nallapati proposed Pairwise link-LDA and Link-PLSI-LDA [7] to tackle the document dependency problem. In this model, citations are guaranteed by document pair's topic distribution. For each pair of documents, it is treated as presence or absence of a citation which depends on a Bernoulli random variable. To explicitly consider the document relations represented by citations, Guo et al [2] proposed CT model which assumes a probabilistic generative process for corpora. Word sampling process in this model is completely controlled by the topic distribution of its citations. So the original content of the document itself is ignored. This perspective of treating citations can greatly reveal the document relations among them. Tang and Zhang [8] proposed a two layer Restricted Boltzmann Machines to model the links and word simultaneously. Links and words are linked together by a layer in the undirected graphical model.

Besides these topic models, many non-topical procedures are proposed. Qi He [9, 10] proposed another representation for document by utilizing the links information. They represent documents by its citation information. Citation information are actually manually generated contents by different researchers to describe a certain document. Therefore, citation information according to context are ideal for representing the documents by less words. Aya [11] proposed a machine learning algorithm to understand the motivation for the citations. Huang [12] investigated the effect citation contexts have when applied to clustering citations into topics and Ritachie [13] extensively investigated the impact of various citation context extraction methods.

Our procedure is different from previous procedures on how to treat links. We treat link as prior information in another way instead of word appearances. By doing this, we can maintain the dependencies while we model the documents. Both topical similarities between and dependencies documents are reflected in our proposed Plink-LDA model, which in turn promotes the performance. We compare our results on

dataset Cora with Zhen Guo [2] procedure. Both of our procedures outperform the baseline, LDA procedure. The accuracy is around 40% for LDA, 47% for their procedure and 62% for our Plink-LDA model which is shown in Figure 5.

6 Conclusion

In this paper, we explore the feasibility of utilizing citation information in another way. We propose a model which models citations and words simultaneously. In our model, citations are no longer regarded as observations but prior information. We evaluate this model and the results show that it is feasible. Besides, the proposed model can find the researchers' writing habits in the dataset.

In the future, we plan to explore the problem how to determine the tuning parameters automatically for different datasets, such as using EM algorithms.

Acknowledgements. The work is supported by the Natural Science Foundation of China No. 61035004, No. 60973102 and THU-NUS NExT Co-Lab.

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Guo, Z., Zhu, S., Chi, Y., Zhang, Z., Gong, Y.: A latent topic model for linked documents. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, pp. 720–721. ACM (2009), 978-1-60558-483-6
3. Strohman, T., Bruce Croft, W., Jensen, D.: Recommending citations for academic papers. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, Amsterdam, The Netherlands, July 23-27, pp. 705–706. ACM (2007), 978-1-59593-597-7
4. Cohn, D.A., Hofmann, T.: The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity. In: Advances in Neural Information Processing Systems, Denver, CO, USA, vol. 13, pp. 430–436. MIT Press (2000), Papers from Neural Information Processing Systems (NIPS) (2000)
5. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences, Sci. USA 101, 5220–5227 (2004)
6. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, Berkeley, CA, USA, August 15-19, pp. 50–57. ACM (1999)
7. Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint Latent Topic Models for Text and Citations. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, pp. 542–550. ACM (2008), 978-1-60558-193-4
8. Tang, J., Zhang, J.: A Discriminative Approach to Topic-based Citation Recommendation. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 572–579. Springer, Heidelberg (2009), ISSN: 978-3-642-01306-5

9. He, Q., Pei, J., Kifer, D., Mitra, P., Lee Giles, C.: Context-aware Citation Recommendation. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, pp. 421–430. ACM (2010), 978-1-60558-799-8
10. He, Q., Kifer, D., Pei, J., Mitra, P., Lee Giles, C.: Citation Recommendation without Author Supervision. In: Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, pp. 755–764. ACM (2011), 978-1-4503-0493-1
11. Aya, S., Lagoze, C., Joachims, T.: Citation Classification and its Applications. In: Proceedings of the 2005 International Conference on Knowledge Management, ICKM 2005, North Carolina, USA, October 27-28, pp. 287–298 (2005)
12. Huang, S., Xue, G.-R., Zhang, B., Chen, Z., Yu, Y., Ma, W.-Y.: TSSP: A Reinforcement Algorithm to Find Related Papers. In: 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), Beijing, China, September 20-24, pp. 117–123. IEEE Computer Society (2004), 0-7695-2100-2
13. Ritchie, A.: Citation context analysis for information retrieval. PhD thesis, University of Cambridge (2008)
14. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-Link LDA: Joint models of topic and author community. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, p. 84. ACM (2009), 978-1-60558-516-1