ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

Enhanced Document Maintenance System with Efficient Citation Information

Ms.R.Niveditha
PG Scholar, Dept of Computer Science
M.Kumarasamy College of Engineering,Karur
nivedithara gupathi@gmail.com

Mrs.K.Makanyadevi
Assistant Professor, Dept of Computer Science
M.Kumarasamy College of Engineering,Karur
makanyadevik.cse@mkce.ac.in

Abstract— Increasing quantity of articles in online digital libraries or any other repositories induce need for proposing efficient knowledge discovery methods. Effective human-machine interface should be considered since it is important in information retrieval system. Documents in a corpus are connected by citations which get attention now-a-days. To make better citations topic distribution should be performed correctly. The two roles of document should be taken into account while performing topic distribution. One is, a document can act as that document itself in a document level. The other is, it can act as a citation for other documents in a corpus. An idea is to propose a document maintenance system using an efficient topic model named Bernoulli Process Topic model which satisfies all the above goals. Along with this an administrator role is considered for better performance.

Keywords—documents,citations,corpus,topic models

I. INTRODUCTION

The major hurdle in recent years is the way of handling huge documents in electronic form on many corpora. The documents should be stored, organized and retrieved properly. Machine – human interaction is needed for better usability. This becomes a more challenging task in the area of knowledge discovery of documents from huge corpora. Documents should be analyzed correctly and citations should be noted .Citation is nothing but references of a document.

Suppose, consider a person is willing to write an article. For that, he is reading ten articles. Based on the knowledge gained and his own idea, he is writing an article. In this case, the articles from which he gained knowledge become citations for his article. Handling this type of citations of a document is the main job happening in citation network[7]. For achieving this goal, many latent topic models were developed. These topic models used to collect relevant topics from documents and found matching among them. This helps in setting citations for the documents in a corpus.

Among many latent topic models, Probabilistic Latent Semantic Indexing [1] and Latent Dirichlet Allocation[2] are well-known models. After that Link-LDA, citation- influence model, copycat model, inherited topic model, author topic model and many other models are proposed overcoming drawbacks of the preceded models. Before analyzing the topic models we have to be clear in the two properties of citation network [7].

First, a document in a corpus has two levels: document level and citation level (Fig.1.). In document level, document acts as a document itself. In citation level, it acts as a citation of other documents.

Second, a multilevel hierarchical structure should be maintained in citation networks (Fig.2.). This means that the indirect references for a document from which that document only holds little information should also be considered as citation for that document.

The differentiation among the two roles of a document are failed to hold by many topic models. And also the multi-level structure of citation network is failed to capture by these models.

To overcome this limitation, a new generative topic model named Bernoulli Process Topic (BPT) model is used to create a document maintenance system[7]. This model helps in modeling documents in a corpus and also explicitly exploits the properties of citation networks. This model performs topic distribution of document in mixture of two contents: Content of document and related content of the document from other document.

The multi-level hierarchical structure of documents in a corpus are also achieved by performing random walk on directed graph by using random Bernoulli process. So that, the author who is uploading his document can get more citations that he cannot get during learning phase.

II. GENERAL STUDY

A. *Topic models*

Topic models are a type of statistical model which is familiar in the field of knowledge discovery, machine learning and natural language processing. Topic models are used to identify the occurrence of similar words in the collection of documents. For instance, the word “tree” will be present more times in a document describing nature and also in document describing about plant species [8]. The topic models used to model documents in citation networks are many in number. Based on some flaws on one topic model, others are proposed. New innovations also tend to the evolution of new topic models.

Now-a-days all information are becoming computerized and stored on web pages. Increasing number of online digital

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

libraries, scientific articles, magazines and social networks, many computational tools are needed to handle those vast collections. So far we are having two main ways to handle those collections – search and links. Although they are effective in their way, some advanced tools are needed to get information in the user desired fields. We cannot have human resource to read or align huge data on online which is large.

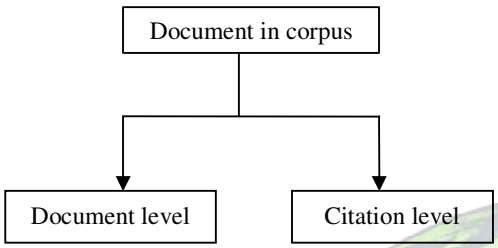

Fig. 1. First property of citation network: An illustration of two levels of document in a corpus

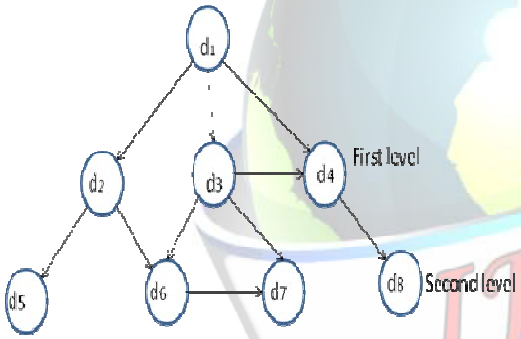

Fig. 2. Second property of citation network: An illustration of the multi-hierarchical structure of documents. Circles denote the documents and arrows denote the citation relationship.

To get rid of this hurdle, machine learning researchers have introduced *Probabilistic topic modeling*. Probabilistic topic modeling is considered as a suite of algorithms that provides full discovery and annotation of large set of documents with desired field information. These topics modeling algorithm used to analyze the topics and the background of that topic and the connectivity among them.

B. Evolution of topic models

Topic modeling algorithms got familiarized from 1998 by the description of some early topic models by Papadimitriou, Ragavan, Tamki and vempala .On later years, with the help of main cyst from earlier topic models, some advanced models are created. The first one is probabilistic semantic indexing (PLSI) discovered by Thomas Hofmann in 1999 [1].

After that ,by generalizing PLSI ,David Blei Andrew Ng and Michael I. Jordan discovered Latent Dirichlet Allocation (LDA) [2] in 2002.It is a widely used most common topic model. It represents a document with mixture of topics.

Following these two topic models, many others are invented by taking LDA as basic and performing some changes to overcome certain problems and to meet user requirements. Some are Pachinko allocation, Relational topic models etc..,

C. Generative topic model

Generative models correspond to probabilistic and statistical field. For given hidden patterns, it randomly generates data values. It also has the capability of specifying joint probability distribution based on label sequences by observing some information [10].These models are used commonly on machine learning to model documents directly or indirectly. Generative model is different from discriminative model. Discriminative model used to sample observed variables only based on some conditions that are predefined. But, the generative model is totally a probabilistic model which samples the observed information with maximum likelihood. And the main factor is that Generative model also considers the complex relationship among the observed information and target variables. Discriminative model fails to do this. Generative model models a document in two ways.

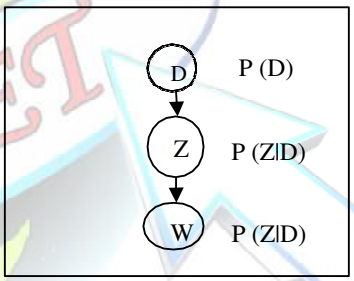

Fig. 3. Plate notation depicting modeling of documents by generative model (case A)

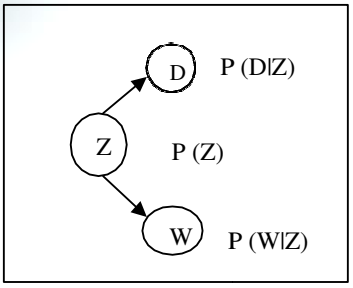

Fig. 4. Plate notation depicting modeling of documents by generative model (case B)

Case A

i. Consider a document

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

ii. Derive topics from document

iii. Derive

words from topics Case B

i. Consider a topic

ii. Derive a document from topic

iii. Derive words from topic

III. RELATED WORK

PLSI is an approach to achieve automatic indexing of documents in a corpus [1]. Document modeling in a corpus is done by using this topic model which considers collection of documents as combination of topics that are distributed over words. Probabilistic Latent Semantic Indexing is good at handling polysemy in words (same word having multiple meaning). For instance, the word "wood" gives different meaning in different documents. This word means to a piece of tree in a document describing tree. In a document having any english novels means the word "wood" denotes the area having trees. PLSI is having solid statistical foundation. It creates variation in the meaning of words and their usage in the document. The performance of PLSI is compared with LSI and other standard term matching algorithm It has better performance than two. But the main limitation in PLSI is that new documents that are not available in the training stage cannot be generated. Modeling increasing number of documents is not so easy and its computation cannot be traced because of its huge matrices. Generative model is nothing but it gives a clear explanation about how documents are distributed.

Like PLSI, In LDA also each document is composed of collection of topics and each topic contains collection of words. The main difference is that LDA uses Dirichlet prior to find the prior knowledge about the documents and the connectivity among them [2]. The topics are extracted from documents and words are generated according to that or else already generated words are classified under those topics[9].

Consider the following sentences

1. I **eat chicken** and **fruits**.
2. *Parrot* is a *pet*.
3. My *Cat* **drinks milk**.

Usually sentences are analyzed by LDA and relevant topics are generated. And words belonging to concern topics are put under those topics. Here two topics based on pets and food is created and they are named as Topic F and Topic P.The words "chicken", "fruits","eat","drinks" and "milk" goes under Topic F and the words "Parrot", "pet" and "cat" goes under Topic P. With the help of word count probability of words in a sentence is derived and the sentences are rated (as in Fig.5) by using that probability.

Sentence	Rating
1	100% of Topic F
2	100% of Topic P
3	30% of Topic P and 70% of Topic F

Fig. 5. A table representing rating of sentences in a document

Three steps should be followed in the working of Latent Dirichlet Allocation on collection of documents to model the topics and assigning words to topics. First step is that the algorithm should be trained based on our desire level. Some estimation can also be done to find topics having maximum likelihood among documents. Other than this normally, topics are created by general walkthrough on documents from top to bottom. In second step, the algorithm randomly assigns words to topics in temporary manner. The words occurring multiple times are stored in different topics. It does not consider the assignment of function words like "the", "in" ,"my" etc.., to any topics. In third step, the temporarily assigned topics are reconstructed by some updating process. Two criteria is chosen to update the topics. A word is taken and it is analyzed to answer two questions. How common is that word across topics? And how common are topics in the document? Finally topics are refilled with relevant words of the sentences in a document.

Although LDA is a model that is used commonly, it is having a flaw of not considering the properties of citation networks. It does not explicitly consider the correlation among documents in a corpus. Two different models are presented Pairwise-link-LDA and the link-PLSA-LDA models [3]. Pairwise-Link-LDA combines ideas of LDA and mixed membership models. It is expensive, because it models the documents without considering either presence or absence of link between the documents. Link-PLSA-LDA model overcomes this limitation by assuming that link structure in a bipartite graph. It combines LDA & PLSA models into a single graphical model. Many models proposed earlier models viz., relational topic model, pachinko model models different information like time, hierarchy of topics, author information etc..But they are failed to consider most important information named hyperlink or citation among documents.

The problem of Joint modeling of text and citation is addressed explicitly considering citations of documents improve power of prediction of the models [6]. Pairwise-Link-LDA is a combination of LDA and mixed membership model which is explained clearly in protein analysis. For a pair of protein documents topics z_d and $z_{d'}$ are derived which are distributed over words. The absence or presence of links among them is noted and a random process is generated based on that. Since the above model is not so scalable, Link-PLSA-LDA is generated. By retaining the scalability of Link-LDA model it considers multinomial sampling process for creation of citations. There is no need to compare every pair of documents in a corpus. It considers the link among documents in a bipartite graph. It helps in observing the variation among cited and citing documents. The drawback described in this paper is that topic distribution for citations are based on fixed

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

set of cited documents. Hence newly created documents are cited within the fixed set of documents.

After that two models are presented – copycat model and citation influence model [4].According to LDA, citations are not influenced by any topics. But in copycat model, a citing document is approximated by using weighted sum of documents in a corpus. Latent topic variables are used to handle semantics and polysemy of words in the topics. In each topic of citing document, there will be the presence of mixture of topics from already cited documents. The words of citing document are associated with the words of cited documents. For example, If we consider two documents 'A' and 'B' as citing documents and 'C' as cited document, the influenced topic mixtures of 'C' are also associated with topics of 'A' and 'B'. This association is named as topical atmosphere in citation environment.

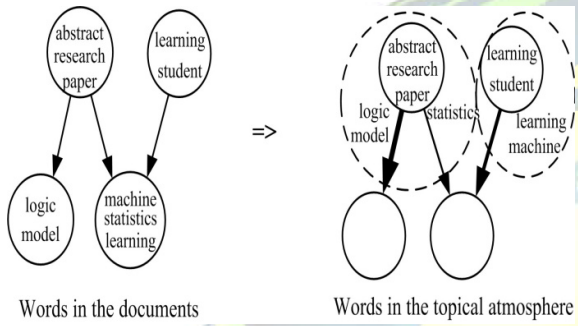

Fig. 6. Diagram depicting the topic influences among cited and citing documents

Two nodes d Ɛ D and c Ɛ C are represented in bipartite graph where the former denotes incoming link to documents and the later denotes outgoing link from the documents in citation graph. By performing association among citing and cited documents (Fig.6.), noise is introduced which may effect the prediction. It becomes hurdle to model evolving topics in copycat model. Hence in citation influence model, a citing document can derive topics from either the same citing document or from cited documents based on effective innovations to handle evolving topics. The limitation specified in this paper is that these models do not consider the multi-level hierarchical structure of citation networks.

Author topic model is a generative model and it is an extension of LDA which includes information about author. It models the documents in a corpus based on author's interest. A mixture of topics is distributed to multiple authors of a document [5]. This author topic model describes the author of the document and other authors who wrote documents similar to the current documents. Here modeling of documents in a corpus is done by considering both content of the document and authors of the document. Suppose consider some authors are willing to write a document .The author is chosen randomly for writing each word of the document. After assigning topics from topic distribution to authors words are generated according to those topics.

IV. PROPOSED WORK

The current trend for building a knowledge-based model is to capitalize on efforts made to design a pre existing well established models or citation (a reference system). A piece of schema relevant to the new application need a module and possibly a personalization is required with extra constraints with respect to the application under construction. Then management of a dataset using the resulting schema must be clear. Complex to analyze the presence of data in the critical environment as well as searching is defined in the means of existing data formulations. The extraction operation requires huge time to retrieve the data from the server, and it is difficult in cost worthy terms. Many security issues occur while maintaining the data in the remote place without any security measures. There are no approvals and acknowledgements from the server side to maintain the data and any user can upload their resources to the destination server. Any user can retrieve the information or data from the stuffed server. So, there are chances for many issues and it is difficult to identify the malwares or attackers. Many topic models are available for maintaining citation of documents in a corpus. Each model has its own advantages and disadvantages. In proposed work an efficient Bernoulli Process Topic Model is used. A document maintenance system is created with an authenticated administrator and registered authors. This uses Bernoulli Process Topic model for its efficient working.

A. Bernoulli Process Topic model

The models seen so far are useful in modeling documents in a corpus to make storage and retrieval easier. Every model takes different information like author information, influence information and others to model the documents. But all fails to handle document levels while distributing topics with newly generated words. BPT considers two level of document namely citation level and document level during topic distribution. It also considers multi-level hierarchical structure of citation networks. As a result, even the documents which are not available during training stage can be marked as citation for newly created documents. Hence it is clear that BPT satisfies the properties of citation networks. An experimental evaluation is done on Cora and Citeseer corpora which prove that BPT is an effective knowledge discovery model than others [7].

B. Architecture

The author who is ready to write a document will first refer to lots of document from huge repositories. The document that he/she refers to becomes citation of that document. Nowadays, referring related documents in a particular area is a very hectic task. For effective knowledge acquisition an effective schema is needed. For that new system architecture is proposed below. The document owner creates a document with proper title and tags. Then they used to upload that on citation system**.** Then, the citation system extracts the content of the uploaded document. Citation mean is calculated for the uploaded

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

document. Using that citation mean, reference documents are identified and their reference id is fetched. After fetching all related documents, citation mean is fixed for that uploaded document. That document is maintained with citation id in server that can be viewed by user and owner of document. The work of administrator is to view and maintain the document stored in system by its owner. The authentication of document and document owners are handled by this administrator.

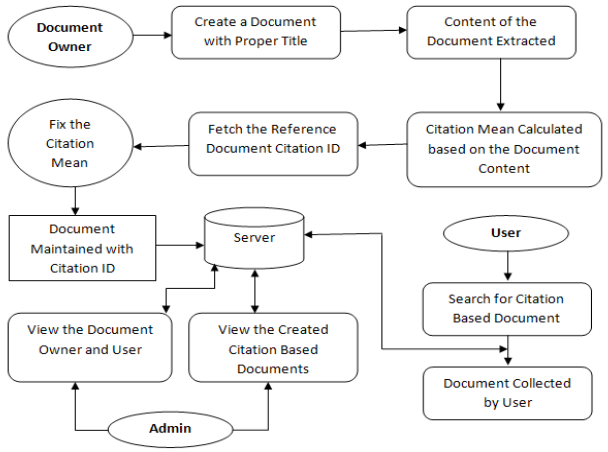

Fig. 7. Proposed Architecture for document maintenance system

a. Structured data storage and cluster handling

The structured data storage scheme fully describes about the flow of data structure maintenance and the concept of implicit data mining and document maintenance schema. Once the resource owner uploads the respective document it checks for the citation schema of the existing document for reference. If the document is presented in the database server, then the following documents is sequenced under the existing document otherwise it creates a new schema for the following document. so that the data into the database server is maintained in the structured manner. All the data into the server is based on the cluster format and provides the frequent access for the user search between the server and the data client (user).

b. Citation level maintenance and resource checking

A novel generative model is proposed for modeling documents linked by citations at a finer level by differentiating the two roles of each document. More specifically, the topic distribution of each document at the document level is a mixture of the topic distributions of the related documents at the citation level. The multi-level hierarchical structure of the citation network is captured by a random Bernoulli process which involves a random walk on a directed graph. In addition to the direct relations among documents, the indirect relations are also fully considered. The comprehensive evaluations are conducted to investigate the performance of the BPT model. The experimental results on the document modeling task demonstrate that the BPT model achieves a significant improvement over the state-of-the-art methods in the generalization performance. Moreover, the BPT model is applied to well-known corpora to discover the latent topics, cluster the documents, detect the trends of various research areas in computer science between 1991 and 1998, and investigate the interaction among the research areas. The comparisons against the state-of-the-art methods demonstrate the promising knowledge discovery capability of the BPT model.

c. Document identity based on citation logic

The document identity purely depends on the content of the document whatever the resource owner uploading into the server and the document content is extracted from the uploaded document and compared the citation with all the existing documents for further reference. Once the document content is matched with the other document identity then the following document is processed under the available document which illustrates the citation mean process exactly and more efficiently.

d. Global answering illustration

The comparisons on the Cora data set with the number of clusters ranging from 2 to 10, where the comparison methods are split into two groups for better illustration. The comparison shows that BPT has the best performance in terms of accuracy and achieves significant improvements in terms of information F1 score. The evaluations on the Cora also show that the relations among the documents do help in the document clustering. On the other hand, some comparison methods only have a good performance in terms of a certain metric. For example, NCut which is a representative spectral clustering method gives a good accuracy, but does not perform well in terms of informationf1 score. By examining the Cora corpus in details, it is found that the Cora data set is very unbalanced, which means that Ncut can obtain a good accuracy by assigning most of the documents to the clusters of large sizes, but the information F1 score is very low.

V. Empirical Results

In document modeling, perplexity is considered main measure on topic models. Lower the perplexity better is the generalization performance of topic models [2].

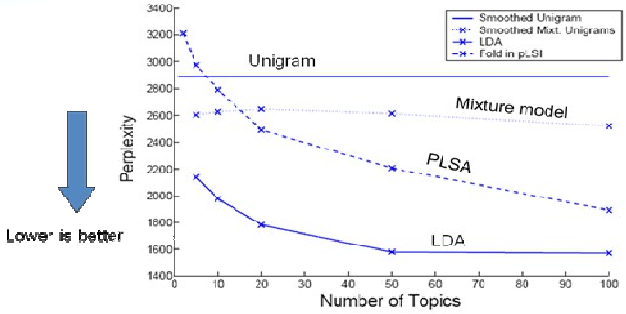

Fig. 8. Perplexity results on nematode corpora for LDA with other models

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 11, March 2016

Perplexity goes down as the likelihood of datasets gets increased. Perplexity of Latent Dirichlet allocation is compared with the unigram model, mixture model and PLSA model (Fig.8.). As a result LDA is having lower perplexity than others [2]. In Fig.9 perplexity of author topic model is compared with other topic model with about 400 topics from NIPS training sets [5]. In Fig.10 and Fig.11 perplexity of Bernoulli Process Topic model on Cora and Citeseer corpora is compared with LDA, Link-LDA, copycat model and citation influence model [7].

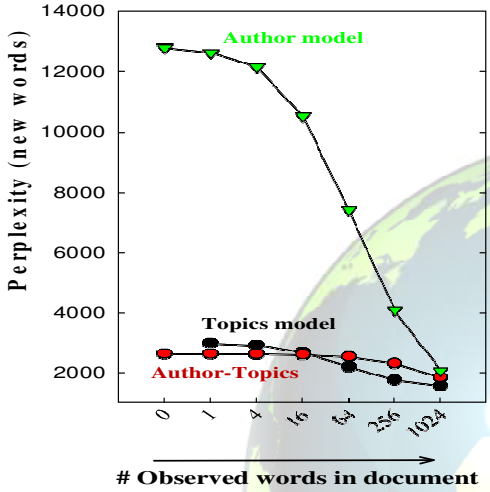

Fig.9. Perplexity of author topic model with other models

Fig.10.Perplexity measure of BPT model with other models on cora corpus

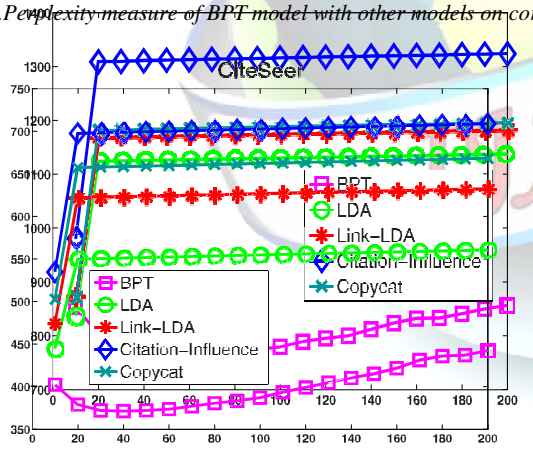

Fig.11. Perplexity measure of BPT model with other models on citeseer corpus

That graph depicts that perplexity measure of BPT is lower than others. So the performance of BPT is better than other models and accuracy is also achieved.

VI. CONCLUSION

All the topic models surveyed above are having efficiency on their own way. By including or improving some factors, some models other than these are also proposed. But, Bernoulli Process Topic model is considered good in handling properties of citation network and it is proved to be efficient by experimental evaluation in real time corpora. By using the same scalability of BPT, an efficient document maintenance system is created. This helps in maintaining documents with best citation information. Accompanying this system with efficient search engine is considered to be the future work.

VII. REFERENCES

[1] *T.Hofmann,"Probabilistic Latent Semantic Indexing "Proc. 22nd Ann. Int'l ACM Conf.Research and Development in Information Retrieval (SIGIR),pp.50-57,1999*

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

[3] *R.Nallapati,A.Ahmed,E.P.XingandW.W.Cohen,"Joint Latent Topic Models for Text and Citations," proc.14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD'08), pp.542-550,2008*

[4] *L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised Prediction of Citation Influences," Proc. 17th Int'l Conf. Machine Learning (ICML), pp. 233-240, 2007*

[5] *M. Rosen-Zvi, T.L. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," Proc. 20th Conf. Uncertainty in Artificial Intelligence (UAI '04), pp. 487-494, 2004*

[6] *D.A. Cohn and T. Hofmann, "The Missing Link—A Probabilistic Model of Document Content and Hypertext Connectivity," Proc. Advances in Neural Information Processing Systems (NIPS '00),pp. 430-436, 2000.*

[7] *Zhen Guo, Zhongfei (Mark) Zhang,Shengghuo Zhu,Yun Chi, and Yihong Gong,"A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks", IEEE Transactions on Knowledge and Data Engineering, vol.26,No. 4,April 2014.*

[8] *https://en.wikipedia.org/wiki/Topic_model*

[9] *https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation*

[10] *https://en.wikipedia.org/wiki/Generative_model*