Topic-sensitive Influential Paper Discovery in Citation Network

Xin Huang, Chang-an Chen, Changhuan Peng, Xudong Wu, Luoyi Fu, and Xinbing Wang(✉)

Shanghai Jiao Tong University, Shanghai, China
hxin18@gmail.com,{chen-chang-an,598095762,xudongwu,yiluofu,xwang8}@sjtu.edu.cn

Abstract. Discovering important papers in different academic topics is known as topic-sensitive influential paper discovery. Previous works mainly find the influential papers based on the structure of citation networks but neglect the text information, while the text of documents gives a more precise description of topics. In our paper, we creatively combine both topics of text and the influence of topics over citation networks to discover influential articles. The observation on three standard citation networks shows that the existence of citations between papers is related to the topic of citing papers and the importance of cited papers. Based on this finding, we introduce two parameters to describe the topic distribution and the importance of a document. We then propose MTID, a scalable generative model, which generates the network with these two parameters. The experiment confirms superiority of MTID over other topic-based methods, in terms of at least 50% better citation prediction in recall, precision and mean reciprocal rank. In discovering influential articles in different topics, MTID not only identifies papers with high citations, but also succeeds in discovering other important papers, including papers about standard datasets and the rising stars.

Keywords: Citation network · Generative model · Academic recommendation

1 Introduction

In academic research, the prior arts are essential for the future works. One bottleneck in research is that as the amount of available scientific literatures on the Internet becomes larger, it would be increasingly difficult for researchers to identify the masterpiece among numerous papers. This problem becomes even more complicated given the fact that important papers are only influential in one or several domains of knowledge. As a result, how to effectively identify the milestone papers in different academic topics is a crucial task in data mining.

The goal of finding important papers in different academic topics is to discover documents which are of great significance in a specific topic. In the researches of citation network, most works try to use the network to discover the interaction[3, 9] and the evolution of topics[5, 17] in the collection of documents, while little attention is paid to finding influencers in different academic topics. Among limited number of works, which indeed focus on influential paper discovery in citation network, Wang et al.[17] adopt Latent Dirichlet Allocation(LDA)[1] to generate citation networks. They view the reference of a paper

as a "bag of citation" and learn the topic-document distribution from the citation network. This distribution describes the importance of documents in a particular topic. Lu et al.[6] extend the method by taking into account additional factors that influence the importance of papers, such as authorship and published venues. The model proposed by Lu et al.[6] could discover the important papers for different topics, authors and venues.

In spite of [6, 17] mentioned above, finding influential nodes in citation networks remains an open problem. One direction is to add topics of textual information into influencer detection. The existing works only take into consideration the network structure and ignore the text. Although [6, 17] use "topic" in the discription of their methods, the topic defined in [6, 17] is actually a cluster of documents. He et al.[4] describe this kind of topic as "DocTopic", which could be simultaneously related to distinct "WordTopics", i.e. topics extracted from the text. Thus, the topics described in [6, 17] are too general but imprecise. The other direction is to determine the important factors that affect papers to cite. As for this direction, we conduct an observation on three standard citation networks. The result shows that whether one paper cites another depends on the topic of the citing paper and the importance of the cited paper.

In this paper, we study the problem of influential paper discovery in citation networks. One feature that distinguishes our method from other related works is that we solve this problem by covering both directions mentioned in the previous discussion. During this process, a challenge is to precisely describe the factors that affect papers to cite. To accomplish this, we introduce two parameters to represent the topic distribution and the importance of a document. Based on these two parameters, we generate citation networks. While we defer a more detailed description of our methods in section 4, we would like to point out that our method succeeds in adding the topic of papers into the generation of citation networks. During this process, the importance of papers is learnt from the data. The topic of node could be obtained by topic modeling, e.g. LDA, or any other methods that transform a document into a topic vector. Another advantage is that our learning schema could be separated into a set of independent convex optimization problems. This propriety indicates that our model is scalable and easy to initiate. The following three aspects are our core contributions.

- We conduct an empirical observation based on three standard datasets: AAN, DBLP and ACM. There are two fundamental conclusions. One reveals that papers with similar topic distribution are likely to cite similar papers as references. The other states that papers with high citations are likely to be selected as the papers to refer.
- We propose a new, robust model: Model for Topic-sensitive Influential paper Discovery(MTID), which is parallelizable and compatible to all methods representing documents with topic vectors. MTID is inspired by our observation and models the citation network with two parameters of papers: 1)An importance parameter, $\mathbf{M}$, that captures the importance of cited papers 2)A topic parameter, $\mathbf{N}$, which describes the topic distribution of citing papers.
- We evaluate our model on citation prediction and influential paper discovery. The first part proves that our model outperforms other topic-based citation

prediction methods with an improvement over 50% in recall, precision and mean reciprocal rank. In the second part, we not only effectively identify the papers with high citations, but also succeed in discovering other important papers such as papers about standard datasets and the rising stars for an academic topic. Taking advantage of this property, we further apply our model to an academic recommender system.

The rest of this paper is organized as follows: Section 2 summarizes the related work. Section 3 presents our empirical observations about how topics of documents influence their citations. Section 4 introduces the MTID along with the learning method. Then, we report the experimental results and the model applications in Section 5. Finally, the conclusion is presented in Section 6.

2 Related Work

In recent years, with increasing number of digital libraries such as ACM Digital Library[1] and DBLP[2] come into use. There is a growing interest on the analysis of citation networks in the research community.

As the citation network is a kind of network with rich textual information, one important direction in the study of citation network is to use the network structure in understanding the topic of text. In this direction, some researchers extend the classical topic model to joint versions in order to model both text and citation for documents. These works succeed in enhancing the quality of topics[14] and reflecting the interaction among topics[3, 9]. Others make use of the characteristic that papers could only cite papers published earlier to study the evolution of topics in academic fields[5, 17].

Another important direction is to detect the influential papers in the citation network. To assess the importance of papers, ranking algorithms such as PageRank and its variants are applied[12, 16]. These methods, however, detect the general influential papers in citation networks and ignore the topic context of documents. In fact, as a document only contains information in one or several knowledge domains, the influential papers vary in different topics. As a result, discovering influencers in different topics is of great value in the citation network.

While most works for topic-sensitive influential node discovery in networks aim at identifying important users in social networks[10, 18], little attention is paid to citation networks. Among limited amount of related works, Wang et al.[17] use topic model to generate citation networks. They introduce the notion of "bag of citation" and consider a topic as a mixture of documents. Then, they learn the topic-document distribution from the citation network. The distribution describes the importance of documents in a specific topic. Lu et al.[6] extend the method by considering additional factors that influence the importance of papers, such as authorship and published venues. The model proposed by Lu et al.[6] discovers the important papers for different topics, authors and venues.

In our model, different from [6, 17], we use the topics extracted from textual information. In this way, we can make full use of the rich information in text.

[1] `https://dl.acm.org/`

[2] `http://dblp.uni-trier.de/`

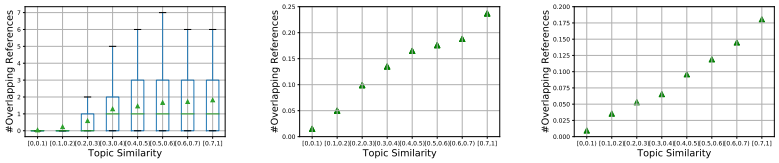
Another difference lies in the generation of networks. In [6, 17], the reference for a document is determined by sampling cited documents according to the topic-document distribution. In this case, the same document could appear more than once in the reference. Our model, however, overcomes this problem by modeling the probability of whether one paper cites another.

3 Empirical Observation

One important direction of discovering important papers in the citation network is to figure out how papers cite other papers. In this section, we adopt empirical observations on the academic network to discover the factors that affect papers to connect with each other by citations. In general, we mainly cope with two important questions. How topics of a document affect the way it cites? Which kind of documents are frequently cited?

We observe three academic datasets: AAN, DBLP and ACM, the detailed description is shown in Section 5.1. For each paper, we extract the topic from the text with LDA[1]. According to the topic diversity of papers in datasets, we set the number of topics to 10 for AAN and 100 for DBLP and ACM.

First, we analyze how topics affect the way documents cite. To do this, for papers published last year in each dataset, we select two papers with more than 10 references and calculate the cosine similarity of their topic distributions. The higher the value is, the more similar the articles are. We repeat this process among all paper pairs of the last-year publication in each dataset. Then, we divide the pairs into 8 different parts according to the topic similarity. Finally we analyze the relation between the size of the overlapping references and their topic similarity within the document pairs.



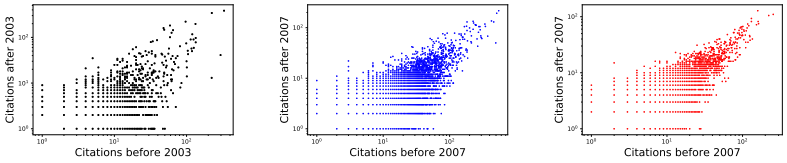
(a) AAN (b) DBLP (c) ACM

Fig. 1. The relation between topic similarity and overlapping references

Fig 1(a), Fig 1(b) and Fig 1(c) respectively show the result of dataset AAN, DBLP, and ACM. For DBLP and ACM, due to the large topic diversity, only a tiny portion of paper pairs have overlapping references. Thus, we only plot the average number of overlapping references for these two datasets. In these figures, the number of overlapping references grows when the similarity of document pairs increases. The result shows that papers with similar topic distribution are likely to cite similar papers.

Second, we analyze, in the citation network, what documents are likely to be cited. For early publications, e.g. papers published in the first three years, we construct two sets based on a timestamp, e.g. the penultimate year. The first set contains the citations before this timestamp, the second set includes citations after this timestamp. For example, for an early publication in ACM dataset, the

first set contains citations before 2007 and the second set includes citations in or after 2007. We then compare the size of these two sets for each early publication.



(a) AAN (b) DBLP (c) ACM

Fig. 2. The relation between exiting citations and incremental citations

Fig.2 shows the relation of the size of two citation sets. It shows that the number of citations in later years is positively correlated to that in early years. In other words, papers with high citations are likely to be selected as the papers to refer. As the number of citations is positively related to the importance of a paper, papers with higher importance are more frequently to be cited.

4 Proposed Model

4.1 Generation of Citation Network

Based on previous observations, we can conclude that whether papers are linked in citation networks is related to the topic of the citing paper and the importance of the cited paper. The former is the intrinsic characteristic of the article and the latter depends on the network structure.

In order to precisely represent these two factors that affect papers to cite, we introduce two new parameters, **N** and **M**, to represent respectively the topic distribution and the importance of a document.

N is the parameter which represents the topical distribution of citing papers. It can be labelled manually or extracted from the text. N_u is a column vector which describes the topic representation of document u and N_{ui} represents how likely the topic i could describe the document u. In topic modeling, N_u is the document-topic distribution in the document u and N_{ui} is the proportion of topic i in this document.

M is the parameter which represents the importance of papers in different topics. It describes how likely a paper would be cited. M_v is a column vector which represents the importance of document v in different topics. $M_{vj} = 0$ indicates that the document v is not important in topic j. The larger this parameter is, the more important the paper is in the topic.

We then present MTID(Model for Topic-sensitive Influential paper Discovery), a probabilistic model which generates the citation network and models the importance of papers in different topics simultaneously. In the generating process, MTID follows the idea presented in [19] and models the citation network with Poisson distribution. Suppose that in a citation network, a non-negative random variable X_{uv} represents the latent connection strength for the pair of papers (u, v). We define that paper u cites paper v if and only if $X_{uv}>0$. Now we consider the case of a single topic. We define X^i_{uv} as the random variable of latent connection strength in topic i for the pair of papers (u, v), this random

variable follows the Poisson distribution with the parameter $N_{ui} \cdot M_{vi}$. Then the total connection strength X_{uv} is the sum of X_{uv}^i, with the additivity of the Poisson random variable:

$$X_{uv} = \sum_{i=1}^{K} X_{uv}^i \qquad X_{uv}^i \sim Poission(N_{ui} \cdot M_{vi})$$

The total connection strength X_{uv} follows the Poisson distribution with the parameter $\sum_{i=1}^{K} N_{ui} \cdot M_{vi}$, where K denotes the number of topics. The probability $P(u \to v)$ is defined as follows:

$$P(X_{uv} > 0) = 1 - P(X_{uv} = 0) = 1 - \exp\left(-\sum_{i=1}^{K} N_{ui} \cdot M_{vi}\right) = 1 - \exp\left(-M_v^{\mathsf{T}} N_u\right)$$

Finally, MTID learns the importance matrix $\mathbf{M}$ and maximizes the log likelihood of the observed network G. The problem could be formalized as follows:

$$\hat{\mathbf{M}} = \arg\max(L(\mathbf{M})) \tag{1}$$

where the nonnegative matrix $\mathbf{M} \in \mathbb{R}^{K \times N}$ and K, N denote the number of topics and nodes, respectively. The log likelihood can be written as below:

$$L(\mathbf{M}) = \sum_{v} \left\{ \sum_{u \in R_v} \log(1 - \exp(-M_v^{\mathsf{T}} N_u)) - \sum_{u \notin R_v, u \in C_v} M_v^{\mathsf{T}} N_u \right\} \tag{2}$$

R_v is a set of papers that cite paper v and C_v is a set of papers that are published later than paper v.

4.2 Parameter Learning

We solve the optimization problem defined in Eq.1 through block coordinate gradient ascent. At each iteration, we update the importance vector M_v for each paper v with M_u for all other papers $u \neq v$ fixed. To update the importance parameter M_v for paper v, we solve the following subproblem:

$$\hat{M}_v = \arg\max(L(M_v)) \tag{3}$$

where $L(M_v)$ is the part of $L(M)$ defined in Eq.2 that involves M_v, i.e.,

$$L(M_v) = \sum_{u \in R_v} \log(1 - \exp(-M_v^{\mathsf{T}} N_u)) - \sum_{u \notin R_v, v \in C_v} M_v^{\mathsf{T}} N_u \tag{4}$$

Noticing M_v is a non-negative vector, this subproblem can be further solved by projected gradient ascent.

$$M_{v_{new}} \leftarrow \max(0, M_{v_{old}} + \alpha_{M_v}(\nabla L(M_v))$$

α_{M_v} is the step size computed by backtracking line search[2], and the gradient is:

$$\frac{\mathrm{d}L(M_v)}{\mathrm{d}M_v} = \sum_{u \in R_v} N_u \frac{\exp(-M_v^{\mathsf{T}} N_u)}{1 - \exp(-M_v^{\mathsf{T}} N_u)} - \sum_{u \notin R_v, u \in C_v} N_u \tag{5}$$

During the iterations, only the calculation of the first term in Eq.5 is required and

the second term is a constant given a paper v. This constant can be computed in $O(In_{degree}(v))$ time according to Eq.6 and cached during the training process.

$$\sum_{u \notin R_v, v \in C_v} N_u = \sum_{u \in C_v} N_u - \sum_{u \in R_v} N_u \tag{6}$$

Thus, the computation of Eq.5 requires $O(In_{degree}(v))$ time. As the real-world citation networks are extremely sparse ($In_{degree}(v) \ll N$), we can update M_v for each iteration in near-constant time.

4.3 Initialization and Parallelization

One advantage of our model is that the optimization problem shown in Eq.4 is concave. In this case, parameters will converge to the same result with different initializations, thus we could randomly initiate the vector M_v for each paper v. Another advantage is that our approach also allows for parallelization, which further increases the scalability of MTID. When updating M_v for each paper v, we observe that each subproblem is separable. That is, updating the value of M_v for a specific node v does not affect the updates of M_u for all other nodes u. Consequencely, parallelization does not affect the final result of the model. The implementation is available in `https://github.com/hxin18/mtid`.

5 Experiment

5.1 Dataset

We evaluate our model with three citation networks, AAN, DBLP and ACM. We use LDA[1] to extract topics from the text of documents. We note that, at the same time, topics extracted by other topic modeling methods are also compatible with our model and tend to have similar results.

ACL Anthology Network ACL Anthology Network (AAN)[11] is a dataset which includes papers about *Natural Language Processing*. AAN includes 19,435 papers published from 1980 to 2013 with full text and reference. For the text, we remove invalid tokens and stop words. As AAN only contains papers in one scientific field, topics detected in AAN are more specific and some of them are quite similar. As a result, we set the number of topics to 10.

DataBase systems and Logic Programming DataBase systems and Logic Programming(DBLP)[15] is a dataset on computer science journals and proceedings. From the dataset, we extract 298,840 papers published from 1996 to 2007 with abstract and reference to build the training set. For the textual information, we remove the invalid tokens and stop words. DBLP contains papers in all sub-fields of computer science. As a result, we set the number of topics to 100.

Association for Computing Machinery Association for Computing Machinery(ACM)[15] is an online dataset on computer science journals and proceedings. From the dataset, we extract 413,373 papers published from 2003 to 2007 with abstract and reference to build the training set. For the text, we remove the invalid tokens and stop words. ACM contains papers in all sub-fields of computer science, as a result, we set the number of topics to 100.

5.2 Citation Prediction

In this section, we evaluate MTID by predicting citations for new documents. The whole dataset is divided into a training set and a test set. For the test set of each dataset, we include the late publications with at least 10 references. The details are shown in the following table.

Dataset	AAN	DBLP	ACM
Size of Training Set	14305	298840	413373
Size of Test Set	2137	16490	16318

We fit our model with the training set and predict citations for the papers in the test set.

Procedure For a new query document with topic distribution N_{new}, the MTID recommends citations among existing papers based on the importance parameter **M**. In details, we compute the citation strength of each existing paper to the query document, then we rank the papers according to the strength and recommend them based on the ranking. The strength is defined as

$$S_d = M_d^{\mathsf{T}} N_{new}$$

Baselines We utilize random selection and three topic-based citation prediction methods as baselines:

- **Random**: For each query in test set, we randomly recommend papers to cite. The result of this method is the average of 10 measures.
- **TopicSim**: TopicSim is to compare the topic similarity between queries and the cited papers. For each query, it returns the papers with the most similar topic distribution. The topic distribution of documents is measured by LDA.
- **Link-PLSA-LDA**:Link-PLSA-LDA[9] is a mixed membership method that models both text and citation. In the citation prediction, the cited papers are ranked in terms of the conditional probability of citations associated with the topic distribution of query.
- **Topic PageRank** This method considers not only the topic similarity between queries and the cited papers, but also the importance of cited papers in the network. For a query, cited papers are ranked in terms of the multiplication of the weight of cited documents in PageRank and the similarity between cited documents and queries.

Metric We adopt Precision and Recall at number N (P@N and R@N) as the evaluation metrics for citation prediction. R@M is defined as the percentage of correct references that appear in the top-N prediction. P@N is used to quantify whether correct references are ranked top for the query. A higher recall and precision indicate a better result.

Furthermore, it is important that ground-truth references appear earlier in the prediction. Therefore, we adopt Mean Reciprocal Rank (MRR) as a metric. The MRR is defined as $\frac{1}{|S_{test}|}\sum_{d\in S_{test}}\frac{1}{rank(d)}$, where S_{test} denotes the test set and $rank(d)$ denotes the rank of first correct citation for query d.

Result Table 1 shows the result of citation prediction for three datasets. Topic-based methods significantly outperform random selection. Among all topic-based models, TopicSim performs the worst because it only exploits information in text. For other three methods that consider both text and citation, MTID significantly

Dataset	Model	P@10	P@20	R@20	R@50	MRR
AAN	Random	0.000967	0.000896	0.001381	0.003419	0.006369
	TopicSim	0.001872	0.002433	0.003669	0.014069	0.014534
	Link-PLSA-LDA	0.016341	0.012858	0.018686	0.034708	0.059297
	Topic PageRank	0.037406	0.041362	0.062698	0.109224	0.095652
	MTID	**0.141056**	**0.101980**	**0.150841**	**0.22537**	**0.309743**
DBLP	Random	0.000030	0.000034	0.000067	0.000148	0.000371
	TopicSim	0.002197	0.001551	0.002880	0.005391	0.008760
	Link-PLSA-LDA	0.013687	0.010525	0.025611	0.034949	0.058961
	Topic PageRank	0.013621	0.010598	0.019777	0.037148	0.056319
	MTID	**0.040889**	**0.032202**	**0.058351**	**0.101983**	**0.136039**
ACM	Random	0.000014	0.000025	0.000050	0.000091	0.000225
	TopicSim	0.000978	0.000889	0.001958	0.004279	0.005619
	Link-PLSA-LDA	0.014373	0.011093	0.022353	0.039949	0.053244
	Topic PageRank	0.008274	0.006572	0.013942	0.025045	0.039397
	MTID	**0.022046**	**0.017035**	**0.035423**	**0.060776**	**0.085698**

Table 1. Result of citation prediction

outperforms other two methods. We can also notice that performance on AAN is much better than other two datasets. It is because that DBLP and ACM are large networks with wider range of topics. This makes the prediction more difficult. The result proves the effectiveness of MTID in citation prediction.

5.3 Finding Influential Papers

In this section, we adopt MTID in discovering influential papers of different topics in the citation network. To do this, for each topic, we rank the papers according to the importance in this topic, which can be reflected by the parameter **M** in our model.

Table 2, 3 and 4 display five most important papers of three topics selected in each dataset, the keywords of topic are displayed in the left of the table. For each topic i, we rank the papers according to the value of M_i. In general, the importance of papers in the citation network is positively related to the number of citations. However, there are some exceptions in our result.

Topic	**M**	**Paper Title**	Year	#Citation
model feature data training set use learning using word result	0.123106	A Maximum Entropy Approach To Natural Language Processing	1996	390
	0.117956	Discriminative Training Methods For Hidden Markov Models: Theory And Experiments With Perceptron Algorithms	2002	351
	0.098413	Word Representations: A Simple and General Method for Semi-Supervised Learning	2010	133
	0.081476	Building A Large Annotated Corpus Of English: The Penn Treebank	1993	1008
	0.081431	A Maximum Entropy Model For Part-Of-Speech Tagging	1996	281

Table 2. Important papers for Topic 5 of AAN

In Table 2, most papers describe *Machine Learning* for *Natural Language Processing*, while the fourth important paper *Building A Large Annotated Corpus Of English: The Penn Treebank*[8] is about parsing and contains little information about the *Machine Learning*. However, it is considered as an influential paper in *Machine Learning* and cited by papers in this topic. The reason is that [8] serves as a standard dataset for papers in *Machine Learning*. For example, [7] uses "gold standard" to describe [8]. In this case, [8] plays a role as an important dataset in the field of *Machine Learning*.

Table 3 presents the important papers in the topic of *Wireless Sensor Network*, the first three papers focus exactly on this field. However, *Chord: A scalable*

Topic	M	Paper Title	Year	#Citation
network networks sensor wireless routing detection nodes mobile protocol performance	0.436223	Directed diffusion: a scalable and robust communication paradigm for sensor networks	2000	450
	0.417814	Ad-hoc On-Demand Distance Vector Routing.	1999	456
	0.407508	System Architecture Directions for Networked Sensors	2000	351
	0.365696	Chord: A scalable peer-to-peer lookup service for internet applications	2001	703
	0.353928	GPSR: greedy perimeter stateless routing for wireless networks	2000	361

Table 3. Important papers for Topic 1 of DBLP

peer-to-peer lookup service for internet applications[13], which ranks the fourth in this topic, is about *Content Distributed Network*. The reason is that *Content Distributed Network* and *Wireless Sensor Network* are two highly correlated topics. Methods utilized by papers in *Content Distributed Network* are frequently referenced by papers in *Wireless Sensor Network*. In this case, important papers in *Content Distributed Network*, such as [13], are also considered as important references in the topic of *Wireless Sensor Network*.

Topic	M	Paper Title	Year	#Citation
data query database queries mining search databases efficient processing time	0.149107	Aurora: a new model and architecture for data stream management	2003	114
	0.134329	Compressed full-text indexes	2007	20
	0.114798	Issues in data stream management	2003	80
	0.099354	Load shedding in a data stream manager	2003	72
	0.096368	What's hot and what's not: tracking most frequent items dynamically	2003	68

Table 4. Important papers for Topic 38 of ACM

In Table 4, paper *Compressed full-text indexes* ranks the second with only 20 citations. Recalling that the ACM dataset contains papers published from year 2003 to 2007, we can know that paper *Compressed full-text indexes* gains 20 citations in less than one year. In the academic network, papers like *Compressed full-text indexes* are considered as rising stars. Thus, the paper *Compressed full-text indexes* should be recognized as an important paper in *Data Management*.

The examples above prove that MTID not only recommends papers with high citations but also discovers important references such as the papers about standard datasets and raising stars. This propriety improves the performance of our model in academic recommendation. Here, we use *Microsoft Academic Graph*(MAG)[3] to construct a demo of an academic recommender system. MAG dataset contains over 100 million scientific papers with title, references, publish time, and a hierachy of "Field of Study" (FoS) ranging from Level 0 to Level 3. From the dataset, we extract 92,992 papers with Level 1 FoS under *Computer Science* such as *Computer Vision*, *Data mining*, etc. The FoS at Level 1 is considered as ground-truth topics of papers and a paper could belong to one or several topics.

Fig.3 displays our implementation of model on a recommender system. Totally, we recommend papers in 9 topics. For each topic, we have 10 paper recommendations and papers are ranked according to M-score, i.e the parameter **M**

[3] `http://acemap.sjtu.edu.cn/acenap/datasets`

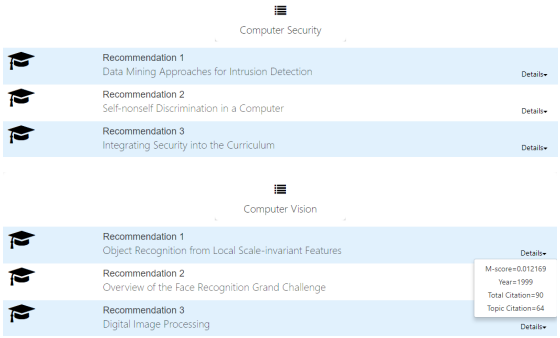


Fig. 3. Demo of academic recommendation in Computer Security and Computer Vision

in our model. Fig.3 gives a snapshot of the recommendations in topic *Computer Security* and *Computer Vision*. On the right of each paper, there is a button called *Details* which reveals the comprehensive statistics about the paper. Apart from *M-score*, we also utilize *Total Citation* and *Topic Citation* to present the importance of papers. The *Total Citation* refers to the number of citations in the whole citation network and *Topic Citation* refers to the citations among documents in the same topic.

In fact, there are 6,975 papers in the topic of *Computer Vision*, which means that the recommended paper *Object Recognition from Local Scale-invariant Features* is cited by nearly 1% of all papers in the same topic. This proportion is high enough, considering the sparsity of the citation networks.

Note that the example is not a special case in our recommendations, more details are availabe in `https://hxin18.github.io/mtid_demo/`

6 Conclusion

In this paper, we study the problem of topic-sensitive influential paper discovery in citation networks. We study how papers cite other papers by observing three standard citation networks and find that the citations are related to the topic of citing papers and the importance of cited papers. Based on the observations, we bring in two parameters to represent the topic and the importance of documents. Combining these two parameters, we propose MTID, a generative model to generate citation networks and learn the importance of papers in different topics from the data. Extensive experiments show that MTID significantly outperforms other topic-based methods in citation prediction. Furthermore, we demonstrate that MTID not only identifies papers with high citations, but also succeeds in discovering other important papers in different topics, including papers about standard datasets and the rising stars. Taking advantage of this property, we apply our model to an academic recommender system.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)

2. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
3. Chang, J., Blei, D.M.: Relational topic models for document networks. In: International conference on artificial intelligence and statistics. pp. 81–88 (2009)
4. He, J., Huang, Y., Liu, C., Shen, J., Jia, Y., Wang, X.: Text network exploration via heterogeneous web of topics. In: Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. pp. 99–106. IEEE (2016)
5. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, L.: Detecting topic evolution in scientific literature: How can citations help? In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 957–966. ACM (2009)
6. Lu, Z., Mamoulis, N., Cheung, D.W.: A collective topic model for milestone paper discovery. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 1019–1022. ACM (2014)
7. Madhyastha, P.S., Carreras Pérez, X., Quattoni, A.: Learning task-specific bilexical embeddings. In: Proceedings of the 25th International Conference on Computational Linguistics. pp. 161–171 (2014)
8. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational linguistics 19(2), 313–330 (1993)
9. Nallapati, R.M., 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. pp. 542–550. ACM (2008)
10. Pal, A., Counts, S.: Identifying topical authorities in microblogs. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 45–54. ACM (2011)
11. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The acl anthology network corpus. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. pp. 54–61. Association for Computational Linguistics (2009)
12. Sayyadi, H., Getoor, L.: Futurerank: Ranking scientific articles by predicting their future pagerank. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 533–544. SIAM (2009)
13. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. ACM SIGCOMM Computer Communication Review 31(4), 149–160 (2001)
14. Sun, Y., Han, J., Gao, J., Yu, Y.: itopicmodel: Information network-integrated topic modeling. In: Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. pp. 493–502. IEEE (2009)
15. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: KDD'08. pp. 990–998 (2008)
16. Walker, D., Xie, H., Yan, K.K., Maslov, S.: Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment 2007(06), P06010 (2007)
17. Wang, X., Zhai, C., Roth, D.: 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. pp. 1115–1123. ACM (2013)
18. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM (2010)
19. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 587–596. ACM (2013)