A Unified Graph Model for Personalized Query-Oriented Reference Paper Recommendation

Fanqi Meng[1,2], Dehong Gao[1], Wenjie Li[1], Xu Sun[1], Yuexian Hou[2]
[1]The Hong Kong Polytechnic University, [2]Tianjin University, China
mengfanqi928@hotmail.com, {csdgao, cswjli}@comp.polyu.edu.hk,
xusun83@gmail.com, krete1941@hotmail.com

ABSTRACT
With the tremendous amount of research publications, it has become increasingly important to provide a researcher with a rapid and accurate recommendation of a list of reference papers about a research field or topic. In this paper, we propose a unified graph model that can easily incorporate various types of useful information (e.g., content, authorship, citation and collaboration networks etc.) for efficient recommendation. The proposed model not only allows to thoroughly explore how these types of information can be better combined, but also makes personalized query-oriented reference paper recommendation possible, which as far as we know is a new issue that has not been explicitly addressed in the past. The experiments have demonstrated the clear advantages of personalized recommendation over non-personalized recommendation.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval: Information filtering, Retrieval models

General Terms
Algorithms, Experimentation

Keywords
Personalized Reference Paper Recommendation, A Unified Graph-based Recommendation Model

1. INTRODUCTION
With the exponential growth of scientific literature, it has become increasingly difficult for researchers to quickly and efficiently find sufficient publications that can help them to advance their research, especially when they want to step into a new field. This has led to the development of reference paper recommendation, which aims to recommend to researchers a list of references relevant to their information needs.

Reference paper recommendation has always been an open and challenging problem. Some previous work has assumed that researchers were able to provide a full manuscript, or a partial list of references, or even the context of citations as search queries. We argue that all these postulations are not practical. Instead, we introduce a more realistic and novel task. Given the identity of a researcher (if known) and a short query text describing the problem he/she wants to solve or the new methodology he/she wants to learn, a list of relevant references that has better connection to his/her existing knowledge is recommended. We call this task personalized query-oriented reference paper recommendation and address it in this paper. We believe that people have different expectations even when they provide the same query content for recommendation. The advantage of the personalization has been verified in our experiments.

In the literature of reference paper recommendation, we can see three main lines of work. They are collaborative filtering (CF) based, content-based and graph-based approaches. Each of them has their own merits. CF has been successfully used in product/item recommendation. It explored the user product/item interactions and provided recommendation to individuals based on their previously expressed preferences, yet the queries were out of their concern. Early reference recommendation approaches followed this line. Content-based approaches looked into the content of papers in the word and/or topic levels. To uncover the hidden thematic structures embedded in the papers, they modeled the topics with Latent Dirichlet allocation (LDA) and extended the original LDA to incorporate the paper citations etc. Requiring large data and computing resources for parameter estimation is the bottleneck of them. Graph-based approaches on the other hand focused on the citation network connections. Existing graph-based approaches often considered paper recommendation as a citation link predication task and derived their solutions based on the random walk properties. To predict whether a new citation link existed, they have assumed some references were already known. Both content-based and graph-based approaches were query-oriented. Authors' publication behaviors were not fully utilized.

In this paper, we propose a unified graph model to combine the strengths of the three approaches and to address their individual weaknesses. We model the papers, their contents, authors as well as the relationships between them as a multi-layer graph. In this graph, the words and the LDA generated topics are naturally integrated with the network connections like citations and collaborations. Given a query from an individual, a graph-based approach can "walk" on this graph to find out the required publications for recommendation. The advantages of the proposed model are three-fold. First, it can easily incorporate various types of information and thus allows us to thoroughly explore how the various types of information can be better combined for efficient recommendation. Second, by combining topics and citations in a graph-based framework instead of a probabilistic framework, computing efficiency is largely improved and a better performance is achieved. Third, it enables personalized recommendation when the author publication history is properly used.

The contributions of this paper are summarized as follows. (1) We propose a unified graph model based on which different reference paper recommendation approaches have been fairly compared on a series of experiments. (2) We tackle with a novel personalized query-oriented recommendation task based on the proposed model. To the best of our knowledge, this is the first work that explicitly addresses

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

personalization in this field. The details will be introduced in the following sections.

2. RELATED WORK

The early work on reference paper recommendation explored the use of collaborative filtering (CF) techniques based on the rating matrixes created from the paper citation network [8]. It was advanced in [11] by combining the recommendations generated from CF and content-based filtering (CBF). Emphasizing on the citation correlations, [2] proposed a graph-based PageRank-like recommendation approach by performing a biased random walk on the citation network. [13] combined the paper citation graph, the author-paper graph and the paper-venue graph to capture the paper similarities in a latent space. They applied label propagation to learn the labels of candidates by setting some known reference papers as positive labels.

The conventional CF-based and graph-based approaches normally stressed more on the role of network connections. They regarded recommendation as a link prediction problem. In particular, they assumed a partial list of references was already known and then recommended suitable additional references for a target paper. Such an assumption is unrealistic. This problem was soon realized and most recent work defined the task as recommending reference papers for a query that could be the whole or the title/abstract of a target paper. Given an unpublished manuscript as a query, [10] combined several text-based and citation-based features in a weighted linear manner. They reported that neither text-based nor citation-based feature performed well in isolation. The idea of recommendation via feature combination was followed by [12]

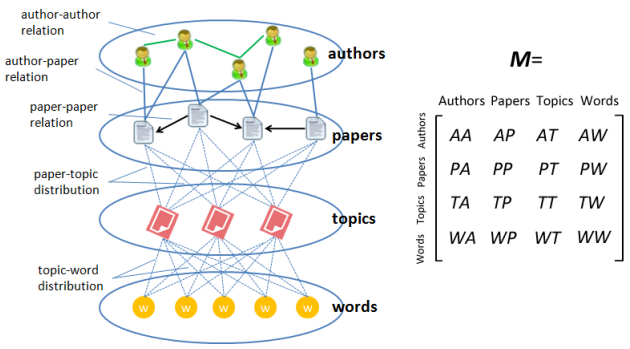
With the success in many text mining applications, Latent Dirichlet allocation (LDA) has drawn a lot of attention to the researchers working on paper recommendation for topic modeling. For example, Link_PLSA_LDA was a latent topic model extended from the basic LDA proposed in [9]. It jointly modeled the topic-level content and citations, and regarded paper recommendation as the citation link prediction task. Later a variation of Link_LDA, called relational topic model (RTM) was introduced by [1] to improve inference and learning efficiency.

There were also some studies looking into the citation context (i.e., the sentences containing or the words surrounding the citations). [4] proposed a probabilistic model to measure the relevance between a query manuscript and a potential cited paper based on the global context (i.e., the title and abstract of a query manuscript) or the local content (i.e., the citation context). [6] employed a translation model to create the relationships between the cited papers and their contexts. Underlying these approaches was the belief that the citation contexts provide more precise descriptions of the cited papers and therefore could capture more precise relevance judgment for linking the query text and the reference paper. It was also reported in [5] that the citation context could help to avoid "topic drift" that the approaches based on citation analysis might cause. Unfortunately, extraction of local citation contexts is not easy for certain reference formats. It may involve named entity recognition and disambiguation and the citation mismatches will inevitably introduce the unexpected noises into the context-based approaches.

3. PERSONALIZED QUERY-ORIENTED REFERENCE PAPER RECOMMENDATION

3.1 Unified Multi-Layer Graph Modeling

The graph-based approaches are capable of formulating the network information. Nevertheless they are not good at capturing the content information. We model the paper content through the links of papers and words (or topics) such that both the network information and the content information can be integrated in a unified multi-layer graph model. The model as illustrated in Figure 1 has four layers corresponding to the four different types' of objects, i.e., authors (A), papers (P), topics (T) and words (W). The connections among them are formulated by the intra-layer links that denotes the relations between the same types of objects and the inter-layer links that relates the objects of different types. Notice that though we only consider paper and author relations in the current work, it is also worth exploring the semantic relations between topics and words (TT and WW) in the future.



(a) Four-Layer Publication Graph (b) Graph Matrix Representation

Figure 1. Multi-Layer Graph Modeling

The design of this model is motivated by the common practice that people do when they search for reference papers. Imagine that you want to find the papers about a topic of your interest. You may consider the papers containing the query words (i.e., keyword matching) or presenting the same or very similar topics as indicated by the query words (i.e., topic matching). You may want to further examine the papers cited by those papers (i.e., citation expansion) or written by the authors of them (i.e., author expansion). You may be also concerned with if those papers have been well recognized (i.e., citations) or if they can enable you a smooth transition from your existing knowledge to a new research filed (i.e., personalization). All these evidences can be explicitly or implicitly incorporated in our graph model (M). Based on the information embedded in this graph, we then apply a random walk based algorithm to recommend the papers to the people who provide the query text.

3.2 LDA based Topic Modeling

A paper can involve multiple topics. With the word content alone, we can hardly find out conceptually related papers that may use different wordings. A topic model like LDA is an effective means that helps us to discover semantic content of papers so as to improve the chance of correct match.

LDA is a generative model. It treats a topic as a probability distribution over a set of words, and characterizes a paper as a probabilistic mixture of the topics. Here, we use the smoothed LDA with symmetric Dirichlet priors introduced in [3]. It uses a Markov chain Monte Carlo algorithm for inference, which is competitive in speed and performance compared with the other parameter estimation algorithms. Given a set of publications P and a set of words W, LDA assumes that there are a set of latent topics T. By applying Gibbs sampling, we construct a Markov chain that converges to the posterior distribution on T and then use the results to infer the two parameter matrixes, i.e. the paper-topic distributions (PT) and the topic-word distributions (TW). With these learned probability distributions, we can use a similar Gibbs sampling process to discover the query topic distribution (q_T).

3.3 Random Walk based Recommendation

For recommendation, we need to measure the query relevance for each publication. To this end, we adopt a Random Walk with Restarts (RWR) framework [7]. Though other random walk based algorithms

are also applicable, RWR has proven to be a powerful tool for computing node proximities on graphs.

The RWR algorithm works like this. Suppose a random walk starts from a starting node. It iteratively transmits to the other linked node with the probability that is proportional to the edge weight between them. Also at each step, it has a restarting probability α to return to the starting node. In this work, we regard the query as the starting points. Let $r^{(t)}$ denote a column vector where $r_i^{(t)}$ is the probability that the random walk at step t visits the node i, and q=[q_A, 0, q_T, q_W] denote a query column, where q_W and q_T are the vectors of query words and the topic distribution of query words. If the author information is considered in q, the values corresponding to the known authors are set to 1 in q_A. Then, RWR can be expressed as:

$$r^{(t+1)} = (1-\alpha)\, M\, r^{(t)} + \alpha\, q \quad (1)$$

where M is the transition probability matrix described in Section 3.1. Currently, the relations between authors and words/topics (AW and AT) are not involved. $r^{(0)}$ represents the initial distribution of the nodes on the graph. We set $r^{(0)}$ equal to q. q and M are both column normalized so that the unique stationary for each node can be obtained by repeating the iteration of Equation (1) until r converges. Then, we sort the paper nodes in descending order of their stationary probabilities and recommend them to the query in that order.

4. EXPERIMENTEAL RESULTS

The experiments are conducted on the ACL Anthology Network (AAN) [1] dataset. After removing the papers missing titles and abstracts, we have 13,885 papers published from 1965 to 2012. Each paper is then preprocessed by (a) extracting its abstract and title; (b) removing the words which consist of 2 characters or less [2]; (c) removing stop words; and (d) stemming the remaining words with porter stemmer. To reduce the noise, we also remove the words appearing less than ten times in the dataset. This yields a set of 13,885 papers with 3,854 unique words. We use 12,762 papers published before 2012 to train the LDA model and construct the relation matrix (PP, AP and AA), and evaluate the recommendations made from those 12,762 papers for the left 1,123 papers published in 2012. For evaluation purpose, we assume a query to be the title and abstract of a paper.

Table 1. Statistics of the AAN dataset

	Papers	Authors	Cited Papers	Citations
1965 – 2011	12,762	9,799	8,544	68,475
2012	1,123	1,557[3]	3,692	10,437

Following the common practice, we use papers' reference lists as ground truth for evaluation, and compare the performance of different recommendation approaches according to recall@N, where N=25, 50, 75, 100 and Mean Average Precision (MAP) which counts ranks of correct recommendations. In all the experiments involving the LDA generated topics, we set the topic number to be 300, which is the best topic number observed in [3]. We have conducted quite a number of experiments. Due to the page limitation, we report only the main findings below.

4.1 Evaluation of Query-Oriented Recommendation

In the first set of experiments, we ignore the author information in the query, and put our focus on the role of paper content, paper citation network and author collaboration network. By this, we mean the recommendation is made simply based on the query text, regardless who submits the query. Table 2 compares the results of the following 4 approaches supported by our model.

- **PW**: Based on the 2-layer model, it uses words in papers and paper citations for recommendation. q =[0, q_W].
- **PT**: Also based on the 2-layer model, but it uses topics in papers and paper citations for recommendation. q=[0, q_T].
- **PTW**: Based on the 3-layer model, it uses words, topics in papers and paper citations for recommendation. q=[0, q_T, q_W].
- **APTW**: Based on the full 4-layer model, the information about papers, citations and authors are all involved for recommendation. q=[0, 0, q_T, q_W].

Table 2. Query-Oriented Recommendation

	MAP	recall@25	recall@50	recall@75	recall@100
PW	0.104486	0.216397	0.303301	0.360773	0.404898
PT	0.079774	0.191252	0.277447	0.339160	0.386088
PTW	**0.109516**	**0.226994**	**0.317538**	**0.375525**	**0.425142**
APTW	0.107532	0.223543	0.309884	0.364998	0.412479

Conclusions and Findings: (1) The word information is more important than the topic information. The topic information helps to improve content-based recommendation on the top of the word information. (2) We examine the difference between the top-50 recommendations made by PW and PT. The overlap of them is about 2/3 of each and PW performs much better in top-2 recommendation. For the first recommendation, the accuracy of PW is about 75% more than that of PT. (3) For non-personalized recommendation, adding the author information into the model seems to have negative impact when it is not concerned in the query. The authorship and co-authorship information embedded in the 4-layer model somehow introduces the unnecessary noise.

In addition, we also run Link_PLSA_LDA[4] on our dataset. Both PTW and Link_PLSA_LDA use the same types of information. While learning Link_PLSA_LDA parameters is about seven times longer than learning LDA parameters, the overall recall of PTW is more than 8.97% above that of Link_PLSA_LDA.

4.2 Personalized vs. Non-Personalized Query-Oriented Recommendation

As what we have declared, given the same query text, we would expect personalized recommendation provide more tailor-made suitable recommendation to the individuals than non-personalized recommendation. It is of great interest to us to look at whether this is true. Here, we compare the results of the following 3 approaches where the author information is utilized in the query. q_1 and q_2 denote a non-personalized and a personalized query, respectively. Notice that when the person who sends the query for recommendation has no previous publication record, our approaches will be reduced to non-personalized summarization for him/her which is simply based on the query text.

- **APW**: Based on the 3-layer model, it uses authors, words in papers and paper citations for recommendation. We compare two queries q_1= [0, 0, q_W] and q_2=[q_A, 0, q_W]
- **APT**: Also based on the 3-layer model, it uses authors, topics in papers and paper citations for recommendation. We compare two queries q_1=[0, 0, q_T] and q=[q_A, 0, q_T].
- **APTW**: Based on the full 4-layer model, it uses all the information about papers, citations and authors for recommendation. We compare two queries, q_1=[0, 0, q_T, q_W] and q_2=[q_A, 0, q_T, q_W].

[1] We download the latest 2012 release from http://clair.eecs.umich.edu/aan/index.php.

[2] When AAN papers are converted from PDF to TEX format, some words are partitioned due to the use of special characters or line feeds. Removing very short character strings can filter out most meaningful "words".

[3] Among them, 1,152 authors have published papers before 2012.

[4] The source code is downloaded from http://gibbslda.sourceforge.net.

Table 3. Personalized vs. Non-Personalized Recommendation

	MAP	recall@25	recall@50	recall@75	recall@100
APW, q_1	0.101420	0.211272	0.293490	0.345080	0.385892
APW, q_2	0.127324	0.243617	0.334266	0.400545	0.455547
APT, q_1	0.079651	0.190381	0.280053	0.340653	0.384428
APT, q_2	0.111473	0.222835	0.333599	0.410765	0.462984
APTW, q_1	0.107532	0.223543	0.309884	0.364998	0.412479
APTW, q_2	0.127207	0.246542	0.350315	0.419907	0.476467
AP, q=[q_A, 0]	0.093919	0.179113	0.241673	0.283644	0.315316

Conclusions and Findings: (1) Personalized recommendation is clearly superior to non-personalized recommendation. In average, it is about 16.20% increase, and APT is increased even more significantly than APW and APTW. (2) To understand what makes such changes, we compare correct recommendations with regard to q_1 and q_2. The latter is able to find out more papers published by coauthors or the papers closer to authors' previous research fields. (3) Using the author citation network rather than the author collaboration network gives quite similar performance. That is why we only present the results with the co-author network in the paper.

Traditional CF-based approaches are personalized but not query-oriented. For comparison purpose, we also experiment with the AP approach which is approximated to CF-based recommendation. AP is based on the 2-layer model that concerns authors' past publications for recommendation. It involves the paper citation and author collaboration network information but ignores the content of papers. For our current task, AP is clearly not a promising approach, as shown in the last row of Table 3. This demonstrates the essential role of content. Though the network-based information helps discover more implicitly related papers, it also introduces more noise and may cause topic drift as noted previously. The content and network information are complementary.

4.3 Further Analysis of the Roles of Paper and Author Networks

To further understand the roles of paper and author networks, we conduct the following additional experiments by removing the paper networks or the author networks from the personalized APTW approach.

Table 4. The Roles of Paper and Author Networks

	MAP	recall@25	recall@50	recall@75	recall@100
APTW	0.127207	0.246542	0.350315	0.419907	0.476467
APTW (-PP)	0.111169	0.224042	0.303270	0.360204	0.401038
APTW (-AA)	0.131114	0.250390	0.354909	0.418431	0.477889

Conclusions and Findings: While the content information is absolutely important, the graph model also benefits a lot from the additional network information. Looking at the results presented in Table 4, paper citations clearly add values to the content based information. The gap between the approach with the citation network and the approaches without the citation network is 15.23% of the latter. In contrast, it appears that the author-paper relations are sufficient and author-author relations don't work very well on our dataset. Actually, *AA* can be deduced from *AP*.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a personalized query-oriented reference paper recommendation task and develop a multi-layer graph model as a solution to tackle with this task. We conduct a series of experiments to evaluate the different roles of the content and network information. We arrive at the conclusions that personalized recommendation performs significantly better than non-personalized recommendation, and the citation network plays more important role than the co-author network. During the experiments, we also notice how challenging it is if we expect to correctly recommend 50% of the whole actual references in top-25 recommendation from a large pool of publications. We would like to examine if paper and author tailored candidate filtering can alleviate the problem. We also want to explore alternative ways to topic discovery.

Acknowledgments

The work described in this paper was supported by the internal grants from the Hong Kong Polytechnic University (Account Numbers: 4-ZZD5 and G-U904)

6. REFERENCES

[1] Chang, J. and Blei, D. 2009. Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*.

[2] Gori, M. and Pucci, A. 2006. Research paper recommender systems: a random-walk based approach. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Pages 778-781.

[3] Griffiths, T. and Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, Pages 5228-5235.

[4] He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, L. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, Pages 421-430.

[5] Huang, S., Xue, G., Zhang, B., Chen, Z., Yu, Y. and Ma, W. 2004. TSSP: a reinforcement algorithm to find related papers. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, Pages 117-123.

[6] Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, L. and Rokach, L. 2012. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Pages 1910-1914.

[7] Konstas, I., Stathopoulos, V. and Jose, J. 2009. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 195-202.

[8] McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J. and Riedl, J. 2002. On the recommending of citations for research papers In *Proceedings of the 2002 ACM conference on Computer Supported Cooperative Work*, Pages 116-125.

[9] Nallapati, R., Ahmed, A., Xing, P. and Cohen, W. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pages 542-550.

[10] Strohman, T., Croft, W. and Jensen, D. 2007. Recommending citations for academic papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pages 705-706.

[11] Torres, R., McNee, S., Abel, M., Konstan, J. and Riedl, J. 2004. Enhancing digital libraries with TechLens+. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Pages 228-236.

[12] Wu, H., Hua Y., Li B. and Pei, Y. 2012. Enhancing citation recommendation with various evidences. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, Pages 1160-1165.

[13] Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B., Zha, H. and Giles, L. 2008. Learning multiple graphs for document recommendations. In *Proceedings of the 17th International Conference on World Wide Web*, Pages 141-150.