Meta-Path-Based Ranking with Pseudo Relevance Feedback on Heterogeneous Graph for Citation Recommendation

Xiaozhong Liu
School of Informatics and Computing
Indiana University Bloomington
Bloomington, IN, USA, 47405
liu237@indiana.edu

Yingying Yu
College of Transportation Management
Dalian Maritime University
Dalian, China, 116026
uee870927@126.com

Chun Guo
School of Informatics and Computing
Indiana University Bloomington
Bloomington, IN, USA, 47405
chunguo@indiana.edu

Yizhou Sun
College of Computer and Information Science
Northeastern University
Boston, MA, USA, 02115
yzsun@ccs.neu.edu

ABSTRACT

The sheer volume of scholarly publications available online significantly challenges how scholars retrieve the new information available and locate the candidate reference papers. While classical text retrieval and pseudo relevance feedback (PRF) algorithms can assist scholars in accessing needed publications, in this study, we propose an innovative publication ranking method with PRF by leveraging a number of meta-paths on the heterogeneous bibliographic graph. Different meta-paths on the graph address different ranking hypotheses, whereas the pseudo-relevant papers (from the retrieval results) are used as the seed nodes on the graph. Meanwhile, unlike prior studies, we propose "restricted meta-path" facilitated by a new context-rich heterogeneous network extracted from full-text publication content along with citation context. By using learning-to-rank, we integrate 18 different meta-path-based ranking features to derive the final ranking scores for candidate cited papers. Experimental results with ACM full-text corpus show that meta-path-based ranking with PRF on the new graph significantly ($p < 0.0001$) outperforms text retrieval algorithms with text-based or PageRank-based PRF.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

1. INTRODUCTION

In the past few decades, the volume of scholarly publications has increased dramatically, which has had a significant effect on how scholars perceive, retrieve, and consume publications. While rapid access to digital publications can accelerate research, some challenges should be addressed. As domain knowledge in most disciplines expands at a frenetic pace, researchers need academic retrieval systems to efficiently locate the scientific publications they are looking for as the candidate citations.

However, characterizing high-quality scientific information needs (given a textual query) can be more complex and challenging than for other domains. Sometimes, textual queries cannot adequately represent what scholars are looking for, especially when researchers venturing into unexplored academic realms where they feel ill-prepared. For instance, an information retrieval study may cite social network analysis and parallel computing papers, which may not be explicitly addressed by a textual query. Meanwhile, the complex relationships among scientific topics, papers, authors, and venues can be important to characterize the scholar information needs.

Recently, some studies have shown that heterogeneous bibliographic networks can be constructed by utilizing multiple types of links from the scientific repository. It has been demonstrated that by using the heterogeneous link information in network, mining functions, such as similarity search, ranking, clustering, and classification can be significantly enhanced. However, to the best of our knowledge, few prior studies have addressed the "ad-hoc" academic search or citation recommendation problem from a pseudo relevance feedback (PRF) perspective. How to utilize the heterogeneous graph-based PRF is not trivial.

Take the text query "*relevance feedback with language model*" as an example, classical search or feedback algorithms are able to find the papers such as "*Model-based feedback in the language modeling approach to information retrieval*". Feedback based on heterogeneous graph, however, provides different result set, i.e., "*Latent Concept Expansion Using Markov Random Fields*", which comes from the complex relations and paths on the graph and not necessarily similar to the initial query. From PRF's viewpoint, given the target query, we first retrieve a number of top ranked papers (as seed papers), then we can locate important authors, citations, topics, and venues (nodes) on the graph as well as find other important

papers via the paths to those important nodes. For instance, given the aforementioned query, system can find "*Latent Concept Expansion Using Markov Random Fields*", because it is related to "*Croft, W. B*" (important author node), "*SIGIR*" (important venue node), "*Markov Random Fields*" (important topic node), and "*A cluster-based resampling method for pseudo-relevance feedback*" (pseudo relevant paper node) via different kinds of paths. In other words, the heterogeneous graph based PRF conceptualizes various kinds of paths on the graph as the ranking functions (or features) and different paths can "vote" for the recommended citations through a learning model. Comparing with text based search and PRF, heterogeneous graph based PRF tells more *global scholarly information*, which is be very important for citation recommendation tasks.

The contribution of this paper is twofold. First, we propose an innovative ranking method with PRF by employing a number of meta-paths and learning to rank on the heterogeneous graph for citation recommendation task. We use meta-path plus random walk as the PRF ranking functions (features), which can prioritize important publications on the graph that is based on a number of seed nodes. The seed nodes include (1) the publication seed nodes (top ranked papers in the retrieved result, a.k.a., pseudo relevant papers, from the retrieved results) and (2) the keyword seed nodes (from user queries). For example, by using a simple meta-path $P^* \xrightarrow{w} A \xleftarrow{w} P^?$, where P^* are the seed relevant paper nodes in the graph, $P^?$ are the candidate cited papers, and $\xrightarrow{w}$ are the "written by" relationship between *paper* (P) and *author* (A), we rank the candidate cited papers based on the likelihood that the paper is written by a (pseudo) relevant paper's authors . Similarly, we can propose a number of meta-path based feedback ranking functions by addressing other ranking hypothesis. For instance:

- $P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{c} C \xrightarrow{c} P^*$: Relevant paper's author's paper can be relevant, if the candidate paper cited relevant paper.
- $P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$ (with $C \xrightarrow{m} K^*$): Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic and the candidate paper contributes to an important topic.

These two exemplar meta-paths address different feedback ranking hypotheses. All the 18 meta-path PRF ranking features investigated in this study will be introduced in the method section. We utilize learning-to-rank to integrate innovative PRF features for citation recommendation.

Second, we propose a new concept "restricted meta-path", which enables in-depth knowledge mining on the heterogeneous bibliographic networks by allowing restrictions on the node set. For instance , we propose a restricted meta-path $P^* \xrightarrow{c} C \xrightarrow{c} P^?$ (with $C \xrightarrow{m} K^*$), which indicates relevant papers' cited papers can be relevant, if the citation is motivated by important topics K^*. For restricted meta-path, some node type on the path are restricted by some seed nodes, i.e. $C \xrightarrow{m} K^*$ (citation should be motivated by the relevant keyword topic K^*), which indicates that the restricted nodes on the main path should be related to the target type of seed nodes. In order to achieve this goal, a "context-rich heterogeneous graph" is constructed by using full-text publication data along with citation motivation modeling, rather than simple scholar metadata. There are some major differences between previous heterogeneous graph and the one we used for this study. As Figure 1 shows, citation is a node (C) on the graph, instead of an edge, and citation is connected to the keyword nodes ($C \xrightarrow{m} K$), which indicates the citation topical motivation probability inferred from full-text citation context. Meanwhile, each keyword topic (K) is extracted by labeled LDA [23], and each topic is contributed by a number of papers ($K \xrightarrow{con} P$), authors ($K \xrightarrow{con} A$), and venues ($K \xrightarrow{con} V$).

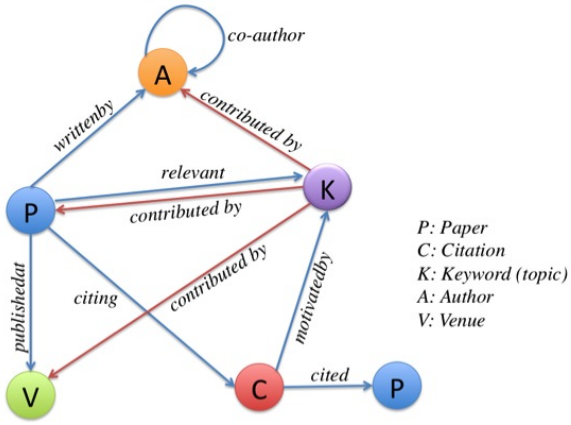

Figure 1: Context-rich heterogeneous graph generated via full-text publications.

Various meta-path-based PRF ranking features along with text search algorithms are integrated via learning-to-rank. By using meta-path-based ranking, citation recommendation PRF does not merely depend on "**term expansion**", instead, "**complex topology based expansion**" on the heterogeneous graph could be used to enhance the ranking performance. Experimental results on ACM full-text corpus show that meta-path-based PRF significantly ($p < 0.0001$) outperforms text and PageRank based PRF for citation recommendation task.

In the remainder of this paper, we: 1) introduce the preliminaries and problem definition, 2) propose our novel method for constructing a context-rich heterogeneous graph and meta-path-based pseudo relevance feedback for citation recommendation, 3) review relevant literature and methodology for citation recommendation, bibliometric analysis, and heterogeneous graph mining, 4) describe the experiment setting and evaluation results, and 5) discuss the findings and limitations of the study and identify subsequent research steps.

2. PRELIMINARIES AND PROBLEM DEFINITION

In this section, we introduce preliminary knowledge on heterogeneous information networks, as well as the pseudo feedback-based techniques in information retrieval.

2.1 Heterogeneous Information Networks and Meta-Path

An information network represents an abstraction of the real world, focusing on the *objects* and the *interactions* between the objects. It turns out that this level of abstraction has great power not only in representing and storing the essential information about the real-world, but also in providing a useful tool to mine knowledge from it, by exploring the power of links. Formally, following the work [32], an information network can be defined as follows.

DEFINITION 1. ***(Information network)*** *An* information network *is defined as a directed graph $G = (\mathcal{V}, \mathcal{E})$ with an object type mapping function $\tau : \mathcal{V} \rightarrow \mathcal{A}$ and a link type mapping function $\phi : \mathcal{E} \rightarrow \mathcal{R}$, where each object $v \in \mathcal{V}$ belongs to one particular*

object type $\tau(v) \in \mathcal{A}$, each link $e \in \mathcal{E}$ belongs to a particular relation $\phi(e) \in \mathcal{R}$, and if two links belong to the same relation type, the two links share the same starting object type as well as the ending object type.

When there are more than one type of node or link in the information network, it is called *heterogeneous information network*.

Network schema is used to specify type constraints on the sets of objects and relationships between the objects of a heterogeneous information network. These constraints make a heterogeneous information network semi-structured, guiding the exploration of the semantics of the network. An information network following a network schema is then called a *network instance* of the network schema. For example, Fig. 1 denotes a heterogeneous information network schema studied in this paper.

In heterogeneous information networks, objects can be connected via different types of relationships. In [32], Sun proposed to use meta-path to systematically capture the relation type between two object types, which is formally defined as follows.

DEFINITION 2. ***(Meta-path)*** *A* meta-path $\mathcal{P}$ *is a path defined on the graph of network schema $T_G = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $\dot{A}_1 \xrightarrow{R_1} \dot{A}_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} \dot{A}_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between types $\dot{A}_1$ and $\dot{A}_{l+1}$, where $\circ$ denotes the composition operator on relations.*

For example, $P - K - P$ denotes a meta-path between papers who connect together due to shared keyword(s). In this paper, we further extend the classical meta-path to restricted meta-path, which can allow more sophisticated path instances selection by putting constraints on some node type along the path (i.e., citation motivated by a restricted set of topics). Also, notation wise, we use P^* to denote that the paper objects are only restricted to the seed set P^*, and use $P^?$ to denote that paper objects are the candidate object type to be expected in the results. More detailed explanations of restricted meta-path will be introduced in Section 3.3.

2.2 Problem Definition

The goal of this research is to enhance the citation recommendation performance based on a piece of text query and a number of user provided keywords.

- **Required Input:** A piece of text to briefly summarize the research work, i.e., paper abstract or research idea description.
- **Optional Input:** Scientific keywords.
- **Output:** A list of ranked papers could potentially be cited given user's input.

For instance, for papers from ACM DL, the input query could be paper abstract and paper keywords, and the output is the recommended reference list.

3. RESEARCH METHODS

In this section, we discuss our innovative method in details, which includes: to construct context-rich heterogeneous graph by using full-text publication data (3.1), to rank the publications via meta-path based ranking function with PRF (3.2 and 3.3), and to combine different meta-path based PRF features with learning-to-rank (3.4).

3.1 Context-rich Heterogeneous Network Construction

In most previous studies, while various methods were used to characterize the citation network, the basic assumption was quite simple: either one paper cites another, or one author cites another, regardless of sentiment, reason, topic, or motivation. For instance, the credits from citing paper are usually assumed evenly distributed to the cited papers. However, intuitively, this is not true (i.e., for a citing paper, some cited papers are more important than the others). Such information loss limits the retrieval or recommendation performance [15].

Full-text publication alone with the citation context analysis has been used for a number of tasks to cope with this limitation. For this study, based on [16, 17], we extract citations in the full-text publication data by using regular expression. Meanwhile, by using the text window before and after each target citation, we inferred the citation topical motivation by using Labeled LDA (LLDA) algorithm [23], $P(Z_{k_i}|c_j)$, where Z_{k_i} is the topic, from citing or cited paper, labeled by keyword k_i provided by paper author, and c_j is the context around citation c_j (left and right 300 words) in the citing paper. More detailed citation topic motivation inference algorithm can be found in [17].

Based on this information, we constructed a novel heterogeneous graph (depicted in Figure 1), and the relations are described as follows. A similar graph is constructed in [15].

Table 1: Relations in the constructed heterogeneous graph

Relation	Description
$P \xrightarrow{w} A$	Paper written by an author
$P \xrightarrow{p} V$	Paper published at venue
$A \xrightarrow{co} A$	Co-author relationship
$P \xrightarrow{c} C$	Paper citing a citation
$C \xrightarrow{c} P$	Citation cited a paper
$C \xrightarrow{m} K$	Citation is motivated by keyword (topic) K, $P(Z_{k_j}\|citation_i)$
$P \xrightarrow{r} K$	Paper relevant to keyword(topic) K, $P(Z_{k_j}\|paper_i)$
$K \xrightarrow{con} P$	Keyword (topic) is contributed by paper
$K \xrightarrow{con} A$	Keyword (topic) is contributed by author
$K \xrightarrow{con} V$	Keyword (topic) is contributed by venue

For any vertex on the graph, the sum of the same type of outgoing links equals 1. For instance, the weight of the link from $paper_i$ to $author_j$ is defined as $w(p_i \xrightarrow{w} a_j) = \frac{1}{d(p_i \xrightarrow{w} A)}$, where $d(p_i \xrightarrow{w} A)$ is the total number of authors of paper p_i. Similarly, we defined the weights of edges in $A \xrightarrow{co} A$ and $P \xrightarrow{c} C$. As one paper can only be submitted to one venue, and one citation only points to one cited paper, $w(p \xrightarrow{p} v) = 1$ and $w(c \xrightarrow{c} p) = 1$. The weight of $p_i \xrightarrow{r} k_j$ is the LLDA probability of topic k_j given the content of p_i, $P(Z_{k_j}|paper_i)$ and k_j is the keyword provided by paper p_i. One limitation of this approach, however, is that a large number of publications in the corpus do not have keyword metadata. In order to solve this problem, we used greedy matching to generate pseudo-keywords for each paper, as used in [9].

In order to estimate the contribution of each paper, venue and author to a topic, we calculated the paper, venue, and author importance given a topic K by using PageRank with Prior algorithm [37]. The normalized topical PageRank authority score is used for the weights of $K \xrightarrow{con} P$, $K \xrightarrow{con} A$, and $K \xrightarrow{con} V$. For this step, we used classical homogeneous graphs, where, on each graph, the vertex is a paper, author or venue. The citation rela-

tionship between the vertexes is utilized to calculate the PageRank authority scores. Each vertex is also characterized by a topic prior vector, i.e., for paper graph, paper topical prior distribution is $P(Z_{k_i}|paper)$, while for author graph, author topical prior distribution is $\sum P(Z_{k_j}|paper_i)$ and $paper_i$ is published by the target author. The result of PageRank (with prior) is the topic based paper, author or venue topic authority vector, i.e., for $paper_i$, the result is a paper authority vector $Authority(paper_i|Z_{k_j})$, the authority score of $paper_i$ given a topic Z_{k_j}.

The weight of $K \overset{con}{\rightarrow} P$, $K \overset{con}{\rightarrow} A$, and $K \overset{con}{\rightarrow} V$ is the normalized topic authority score, which characterize the importance (or contribution) of each paper, author and venue given a keyword. Note that $topic_j$ is contributed by $paper_i$ ($K_j \overset{con}{\rightarrow} P_i$) doesn't necessarily mean $paper_i$ is relevant to $topic_j$ ($P_i \overset{r}{\rightarrow} K_j$). For example, some "natural language processing" papers can be important for "information retrieval" topic.

3.2 Pseudo Feedback Generation

Based on previous studies in feedback, we could employ two kinds of seed nodes for meta-path based feedback ranking: explicit relevant keyword nodes from user initial query (i.e., author provided paper keywords), and pseudo relevant feedback paper nodes from top ranked papers. For this method, a key parameter should be optimized for pseudo relevance feedback - the number of seed paper nodes ($fbDocs$). When we utilize only a few paper seed nodes (a few top ranked papers), the paper seed nodes are more likely to be relevant for the initial query, and the meta-path like $P^* \overset{p}{\rightarrow} V \overset{p}{\leftarrow} P^?$ could more likely find significant venues for feedback. However, small number of $fbDocs$ can be biased for some other meta-path, like $P^* \overset{w}{\rightarrow} A \overset{w}{\leftarrow} P^?$. For instance, if too few paper nodes are used, the number of selected authors is also small, and the feedback ranking result could be biased to those top ranked papers' authors. For this reason, we hypothesize that the optimized paper seed numbers are different for different meta-paths. We will validate this hypothesis in the evaluation part by comparing the optimized $fbDocs$ for different meta-paths.

3.3 Restricted Meta-Path-Based Ranking with Feedback

Unlike most existing PRF methods trying to update the initial query with feedback, the feedback generated using the above approach gives us seed nodes in the network that are most relevant to the query. We then propose to use meta-path based ranking functions via heterogeneous graph to find most relevant papers to these seeds, i.e., paper seeds P^* and the keyword seeds K^*. The ranking score of a candidate cited paper, $P^?$, given a meta-path is the random walk probabilities starting from seed node(s).

3.3.1 Restricted Meta-Path

In the existing meta-path techniques, all nodes from the specified node type and all path instances following the meta-path are considered. However, in our citation prediction task, many constraints need to be put on the objects and thus the propagation from the seed nodes to the candidates following the selected meta-path can be further confined. For instance, the meta-path $P^* \overset{w}{\rightarrow} A \overset{w}{\leftarrow} P^?$ indicates that all papers published by the same authors who have written those seed papers should be considered as candidates. Also, all authors on the meta-path are not necessarily important or relevant to user's initial information need. Meta-path with restriction $P^* \overset{w}{\rightarrow} A \overset{w}{\leftarrow} P^?$ ($A \overset{con}{\nwarrow} K^*$) denotes relevant papers' author can be relevant **only if the author contributes to the relevant keyword topic**. The restricted author, venue, and citation nodes on the meta-path could enhance the accuracy of PRF ranking function.

Thus in this paper, we propose to use restricted meta-paths to confine our interested path instances. A restricted meta-path can be represented as:

$$\sigma_{S_1}(\dot{A}_1) \xrightarrow{R_1} \sigma_{S_2}(\dot{A}_2) \xrightarrow{R_2} \ldots \xrightarrow{R_l} \sigma_{S_{l+1}}(\dot{A}_{l+1})$$

where $\sigma_{S_i}(\dot{A}_i)$ is a selection operator and means only objects in $\dot{A}_i$ that satisfies predicate S_i will be considered. In our case, type $\dot{A}_1$ is the type with seeds, and type $\dot{A}_{l+1}$ is the type of nodes to be queried. For example, $P^* \overset{w}{\rightarrow} A \overset{w}{\leftarrow} P^?$ ($A \overset{con}{\nwarrow} K^*$) is a restricted meta-path from paper type P to paper type P via authors. The constraints are associated with the first paper type P and the author type A in the meta-path. Formally, the first constraint can be represented as $\sigma_{p \in P^*}(P)$ and the second constraint can be represented as $\sigma_{a|\exists k \in K^* \text{such that } k \rightarrow a}(A)$.

In order to quantify the ranking score of candidates relevant to the seeds following the meta-path, a random walk based measure is proposed to compute the relevance between objects in $\sigma_{S_{l+1}} A_{l+1}$ (e.g., the candidate cited papers $P^?$) and objects in $\sigma_{S_1}(A_1)$ (e.g., the seed papers P^*):

$$s(a_i^{(1)}, a_j^{(l+1)}) = \sum_{t = a_i^{(1)} \rightsquigarrow a_j^{(l+1)}} RW(t)$$

where t is a tour from $a_i^{(1)}$ to $a_j^{(l+1)}$ following the specified restricted meta-path, and $RW(t)$ is the random walk probability of the tour t. Suppose $t = (a_{i1}^{(1)}, a_{i2}^{(2)}, \ldots, a_{il+1}^{(l+1)})$, the random walk probability is then $RW(t) = \prod_j \frac{w(a_{ij}^{(j)}, a_{i,j+1}^{(j+1)})}{d(a_{ij}^{(j)})}$, where $d(a_{ij}^{(j)})$ is the restricted weighted degree of node $a_{ij}^{(j)}$ to all the qualified nodes in type A_{j+1}.

In many cases, we also need to add the node prior probability to the random walk function. For example, when the keyword restrictions are added to author type, a relevance score is also added to these authors as defined in the equation, which can be considered as a meta-path dependent prior probability of these nodes. In this case, the above random walk probability of a tour t is then defined as: $RW(t) = \prod_j \frac{w(a_{ij}^{(j)}, a_{i,j+1}^{(j+1)}) p(a_{i(j+1)}^{(j+1)})}{d(a_{ij}^{(j)})}$, where $p(a_{i(j+1)}^{(j+1)})$ is the prior probability of the node. For example, given meta-path $P^* \overset{w}{\rightarrow} A \overset{w}{\leftarrow} P^?$ ($A \overset{con}{\nwarrow} K^*$), as author node A is restricted by topic node K with contribution edge, the prior of author (given the meta-path) is defined by random walk probability from $A \overset{con}{\leftarrow\!\!-} K^*$, which indicates the papers on paths with more important authors (who made more contribution to the seed topics) are more likely to be cited.

The distance from a set of seed objects to a candidate result node can be defined as $s(Q, a_j^{(l+1)}) = \sum_{a_i^{(1)} \in Q} s(a_i^{(1)}, a_j^{(l+1)})$.

3.3.2 Combined Restricted Meta-Path

We can also consider two or multiple parallel meta-paths leading to the same type of query nodes, for example, $\sigma_{S_1}(A_1) \xrightarrow{R_1} \sigma_{S_2}(A_2) \xrightarrow{R_2} \ldots \xrightarrow{R_l} \sigma_{S_{l+1}}(A_{l+1})$ and $\sigma_{S'_1}(A'_1) \xrightarrow{R'_1} \sigma_{S'_2}(A'_2) \xrightarrow{R'_2} \ldots \xrightarrow{R'_l} \sigma_{S'_{t+1}}(A'_{t+1})$, where $S_{l+1} = S'_{t+1}$ and $A_{l+1} = A'_{t+1}$. In this case, we can define similarity from different sets of objects to

a result node from different meta-paths:

$$s(Q_1 \cup Q_2, r) = \alpha s_{MP_1}(Q_1, r) + (1-\alpha) s_{MP_2}(Q_2, r)$$

For instance, $P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{w} A \xleftarrow{con} K^*$ $\underset{m}{\searrow} K^*$ is a combined restricted meta-path, which combines two meta-paths: $P^* \xrightarrow{c} C \xrightarrow{c} P^?$ $\underset{m}{\searrow} K^*$ and $K^* \xrightarrow{con} A \xleftarrow{w} P^?$. So, based on the above formula, the ranking score of candidate cited paper $P^?$ is the linear combination of two random walk scores for both sub-meta-paths. Theoretically, we need to tune parameter α for each meta-path to optimize the weight of each sub-meta-path. For this study, because of the sparse of the training data, we set $\alpha = 0.5$. More sophisticated parameter tuning will be saved for future.

All 18 meta-paths investigated in this study are listed in Table 2. Theoretically, we can propose a very large number of meta-paths for citation recommendation PRF. In this study, however, we use the hypothesis driven approach. All the experimental meta-paths along with the ranking hypotheses are proposed by bibliometrics expert, and those ranking hypotheses can be informative to find candidate citations. As another selection criteria, given the potential online service requirement, the utilized meta-paths should be not too complex, in order to save the computational cost for future.

3.4 Combine Different Ranking Features via Learning to Rank

As we proposed a number of restricted meta-path PRF ranking methods, we need to use learning to rank to statistically combine different ranking features, while avoiding manual parameter tuning. As this study is not focusing on learning to rank, we used a relative simple algorithm, Coordinate Ascent [21], which iteratively optimizes a multivariate objective ranking function, for meta-path PRF feature integration and algorithm evaluation.

As we use paper abstract and author provided keywords as the initial user query, the paper provided references (cited papers) are used as relevant publications. MAP or NDCG can be used as the ranking function training and evaluation metrics. For MAP case, binary judgment is provided for each candidate cited paper (cited or not cited). NDCG estimates the cumulative relevance gain a user receives by examining recommendation results up to a given rank on the list. In this research, we used an importance score, 0-4, as the candidate cited paper importance to calculate NDCG scores. For instance, if an candidate cited paper is not cited by the target testing paper, the importance score is 0, and if a citation is cited 4 or more times in the citing paper, then it is probably very important for the target citing publication, and its importance score is 4.

Table 2: All the meta-path PRF features used in this study

Meta-path	Feedback ranking hypothesis
$P^* \xrightarrow{w} A \xleftarrow{w} P^?$	Relevant paper's author's other papers can be relevant
$P^* \xrightarrow{w} A \xrightarrow{co} A \xleftarrow{w} P^?$	Relevant paper's author's co-author's papers can be relevant
$P^* \xrightarrow{p} V \xleftarrow{p} P^?$	Paper can be relevant if it is published at the same venue as the relevant paper
$P^* \xrightarrow{w} A \xleftarrow{w} P^?$ $\underset{con}{\nwarrow} K^*$	Relevant paper's author's other papers can be relevant if the author contributes to an important topic
$P^* \xrightarrow{c} C \xrightarrow{c} P^?$ $\underset{m}{\searrow} K^*$	Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic
$P^* \xrightarrow{p} V \xleftarrow{p} P^?$ $\underset{con}{\nwarrow} K^*$	Paper can be relevant if it is published at the same venue as the relevant paper, and the venue contributes to an important topic
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xleftarrow{con} K^*$	Relevant paper's author's other papers can be relevant, if the candidate paper contributes to an important topic
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{r} K^*$	Relevant paper's author's other papers can be relevant, if the candidate paper is relevant to an important topic
$P^* \xrightarrow{p} V \xleftarrow{p} P^? \xleftarrow{con} K^*$	Paper can be relevant if it is published at the same venue as the relevant paper, and the candidate paper contributes to an important topic
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{c} C \xrightarrow{c} P^*$	Relevant paper's author's paper can be relevant, if the candidate paper cited relevant paper
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{w} A \xleftarrow{con} K^*$	Relevant paper's cited papers can be relevant, if the candidate paper's author contributes to an important topic
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$	Relevant paper's cited papers can be relevant, if the candidate paper contributes to an important topic
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{r} K^*$	Relevant paper's cited papers can be relevant, if the candidate paper is relevant to an important topic
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{r} K^*$ $\underset{con}{\nwarrow} K^*$	Relevant paper's author's other papers can be relevant, if the candidate paper is relevant to an important topic and the author contributes to an important topic
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xleftarrow{con} K^*$ $\underset{con}{\nwarrow} K^*$	Relevant paper's author's other papers can be relevant, if the candidate paper and the author both contribute to important topics
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$ $\underset{m}{\searrow} K^*$	Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic and the candidate paper contributes to an important topic
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{w} A \xleftarrow{con} K^*$ $\underset{m}{\searrow} K^*$	Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic and the candidate paper's author contributes to an important topic
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{p} V \xleftarrow{con} K^*$ $\underset{m}{\searrow} K^*$	Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic and the venue which the candidate paper published at contributes to an important topic

4. LITERATURE REVIEW

In this section, we will review previous studies focusing on (1) pseudo relevance feedback, (2) academic retrieval and recommendation, and (3) meta-path based recommendation.

4.1 Pseudo Relevance Feedback

Pseudo relevance feedback is an effective re-ranking method to improve the retrieval performance [3, 13, 25, 26, 28, 41]. However, earlier experiments also show that text-based relevance feedback approaches, i.e., Rocchio's query expansion and term reweighing method [26], do not perform well or even harm the ranking performance in some search scenarios [19,35] due to the noisy top-ranked documents. Collins-Thompson et al., [5], for example, used multiple sources of domain knowledge or evidence to enhance the robustness of pseudo feedback by characterizing feedback gain, feedback benefit and feedback risk while minimizing uncertainty in the dataset via risk-aware algorithms.

Recently, there are various efforts to enhance the classical pseudo feedback from feedback document and term optimization perspective. For instance, Lv and Zhai proposed a positional relevance model [19], which extended of the pseudo feedback model to exploit term positions and proximity so as to assign more weights to words closer to query words. For feedback document optimization, more recently, Lee and Croft [14] proposed a pseudo feedback document resampling method, which can skip some documents in the initial high-ranked documents and deterministically construct overlapping clusters as sampling units. A similar document sampling

method is implemented by Sakai et al., [27], which skipped some top-ranked documents based on a clustering criterion. The cluster is generated by the relationship between document and query terms. The sampling algorithm is to select a more varied and novel set of documents for feedback. As another similar approach, [1] and [18] used selective feedback and adaptive feedback to optimize the pseudo feedback performance by selecting "optimized query set" or "optimized amount of query expansion". A pseudo feedback boosting method, FeedbackBoost, proposed by Lv, Zhai, and Chen [20], combined different document weighting and a set of basis feedback algorithms using a loss function defined to directly measure both robustness and effectiveness to improve the overall effectiveness of pseudo feedback. While these methods used different approaches to locate the important feedback documents and terms, they shared one basic assumption: some words in some top-ranked (or selected top ranked) retrieved results can be used to enhance the user initial textual query for better estimation of information need. As noted, this assumption, could be problematic.

Graph based feedback or pseudo feedback is a new ranking assumption, which is rooted in topology-based search. For instance, Dean and Henzinger [6] found the disadvantage of user formulated queries that can hardly characterize information needs, and they utilized connectivity between web pages to recommend related websites to users on the basis of their initial URLs. While graph based feedback is not well studied in the previous researches, Vassilvitskii and Brill [35] investigated the new feedback method by employing hyperlink based web-graph distance for relevance feedback in web searches. The experiment results showed that, for web search, graph based feedback outperforms standard text based relevance feedback methods. Unlike web search, academic retrieval and citation recommendation provide a more complex search scenario, in that, the candidate papers can be interlinked in a heterogeneous graph. While other prior studies simplified this scenario to a homogeneous graph, i.e., paper-cite-paper, two important factors are missing: (1) The complex research objects on the graph, i.e., author, venue, citation, and keyword; and (2) The complex linkage between nodes, i.e., the citation relation between papers, is simplified, when in-depth knowledge, i.e., citation motivation, is overlooked.

4.2 Citation Context and Recommendation

Scientific recommendation is an important research area. This occurs when a scientific publication, venue, or author is recommended to users based on the similarity between the recommended resource and user profiles or samples of text that they are working on. Chandrasekaran et al. [4], for example, present a method of recommending scientific papers of potential interest to users by using the ACM Computing Classification System along with hierarchical concept information from both author profiles and paper content. Based on this work, He et al. [11] proposed a method to recommend global and local citations based on a piece of given text under both context-oblivious and context-aware conditions. In [11], the authors recommend citations to users based on the similarity between a candidate publication's in-link citation contexts and a user's input texts. Unsupervised topic modeling is also used for citation analysis [38], where visible candidate citations, hidden scientific topics, and visible words are represented in different layers. A restricted Boltzmann machine model was used for building the relationship between user input and recommended citation ranking.

Another important approach, scholarly or bibliographic networks, i.e., networks based on citation or co-authorship, have also been used to recommend scientific resources. For instance, Shi, Leskovec, and McFarland [29] developed citation projection graphs by investigating citations among publications that a given paper cites. In this study, the authors investigated high-impact and low-impact citation behavior, where "citation impact" is defined as the number of citations a publication receives normalized by the average number of citations of all other publications published in the same year and same area. More recently, Lao and Cohen [12] used both supervised and unsupervised methods with the Random Walk with Restart (RWR) algorithm for citation, author, and venue recommendation. In this study, a large heterogeneous network (with venue, author, and publication as the vertices, and co-author and citation as the edges) was constructed for the recommendation task. The evaluation results show that supervised RWR can significantly enhance recommendation performance.

As aforementioned, most previous studies in text mining, bibliometrics, and scholar information retrieval/recommendation used citation as a statistical relation between citing and cited papers, while the in-depth knowledge of citation, i.e., topical motivation, is ignored. With further study of citation analysis, increasing numbers of researchers have come to doubt and challenge the reasonableness of assuming that the raw citations reflects an article's influence. For instance, CiteRank [36] is an enhanced ranking algorithm over PageRank, which enables ranking method to estimate the traffic $T_i(\tau_{dir}, \alpha)$ to a given $paper_i$. For this method, a recent paper is more likely to be selected with a probability that is exponentially discounted according to the age of the paper, τ_{dir}. At every step of the path, with probability α the researcher is satisfied/saturated and halts his/her line of inquiry. Dietz et. al,'s Citation Influence Model [7] is another effective method of weighing the importance of a citation relation, which employed citing and cited paper topic distribution and the compatibility-based citation weighting of two topic mixtures is measured by the Jensen-Shannon Divergence. Based on these work, Nallapati et. al. [22] proposed Pairwise-Link-LDA and Link-PLSA-LDA, whose goal is to predict important unseen citation between papers by using topic based graph models.

Unlike those studies, we employed citation context and citation topology to estimate topic based citation motivation, while we assume full-text analysis has to some extent compensated for the weaknesses of citation counts. Moreover, the citation graph with supervised topic analysis is converted to a publication topical prior for language model, which is used to address user textual information need. Ritchie et al., [24] and Bernstam et al., [2] have found that citation context can provide important information for the retrieval task, and that the closeness of a word in the citation context provides stronger semantic information about the cited paper. Meanwhile, Gerrish, and Blei [8] used dynamic influence model to characterize scholar impact without using citation information. Liu et al., [17] motivated us to use the proximity for citation topic inference at the topic level for the recommendation task.

The proposed work differs from previous research in that we use meta-path based candidate cited paper ranking on heterogeneous graph from PRF perspective. Meanwhile, we investigate the deep knowledge on the novel graph by utilizing restricted meta-path plus citation motivation modeling.

4.3 Meta-Path on Heterogeneous Graph

The concept of meta-path was first proposed in [32], which can systematically capture the semantic relation between objects in a heterogeneous information network scenario. Different meta-path-based mining tasks are studied, including similarity search [32], relationship prediction [30, 31], user-guided clustering [33], and recommendation [39, 40]. It turns out that meta-path serves as a very critical feature extraction tool for most of the mining tasks

in heterogeneous information network. In this paper, we propose a novel meta-path-based approach, which is restricted meta-path, to refine the meta-paths that we are interested in. Further, our proposed task is rather different from existing work, which is to use restricted meta-path to re-rank the paper objects in a heterogeneous bibliographic network according to the pseudo feedback nodes and thus provide very accurate citation recommendation for a text-based query.

5. EXPERIMENT

In this section, we describe the experimental setting and results.

5.1 Data and Preprocessing

We used 41,370 publications (as candidate citation collection) from 111 journals and 1,442 conference proceedings or workshops on computer science for the experiment (mainly from the ACM digital library), where full text and citations were extracted from the PDF files. The selected papers were published between 1951 and 2011. From these we extracted 28,013 publication texts (accounting for 67.7% of all the sampled publications), including titles, abstracts, and the full text. For the other publications, we used the title, the abstract, and keyword information from a metadata repository to represent the content of the paper.

We then wrote a list of regular expression rules to extract all the possible citations from paper's full text. For example, the rules could extract " [number]" and " [number, number, number]" as citations from the content of a publication. Each citation extracted from the publication text was associated with a reference (cited paper ID). In a total of 223,810 references ($paper_1$ cites $paper_2$ relations), we successfully identified 94,051 references (from publication full-text), which accounted for 42.0% of all references. Of course, references may have been cited more than once in a citing paper and located in multiple sections.

For graphical PRF, author name disambiguation should not be totally ignored since quite a few meta-paths rely on authors. In this study, we employed the author disambiguation algorithm from [34] to enhance the authorship quality[1], which takes an author's name, affiliation, email, paper title, co-authors, position in the author list as input and matches the author to a canonic author record in the ACM database.

For the later citation recommendation evaluation, we also used a test collection with 274 papers. The selected papers met the following conditions: (1) the selected papers were exclusive from the 41,370 publication candidate citation collection; (2) Each selected paper had more than 15 citations from the candidate citation collection, and (3) Each each paper's abstract had at least 150 words. The paper's abstract was used as a working context to represent a user's information need, and we recommended citations from the candidate citation collection.

5.2 LLDA Topic Model Training and Graph Construction

We sampled 10,000 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. Thus, our LLDA training would have assumed that each paper is a multinomial distribution over a number of topics. During preprocessing we also clustered similar keywords if the edit distance between them was very small, e.g., "k-means" and "k means", or if two keywords shared the same stemmed root, e.g., "web searches" and "web search".

[1]The source code from author disambiguation is available at: https://github.com/SeerLabs/CiteSeerX

If a keyword appeared less than 10 times in the selected publications, we removed it from the training topic space. For publication content we first used tokenization to extract words from the title, abstract, and publication full text. If the word had less than three characters, it was removed. Snowball stemming was then employed to extract the root of the target word. We also removed the most frequently used 100 stemmed words and words that appeared less than three times in the training collection. Finally, we trained an LLDA model with 3,911 topics (keywords). These topics were used to infer the publication and citation topic distribution.

As Figure 1 shows, we constructed a heterogeneous graph, with 41,370 paper nodes, 63,323 author nodes, 369 venue nodes, 3,911 keyword (topic) nodes, and 168,554 citation nodes. Based on the method in section 3.1, we calculated the weighted and normalized the weights on the edges, and used the graph for meta-path based ranking.

5.3 Pseudo Relevance Feedback Experiment Result

[*Meta-path ranking performance comparison*] In the first part of this evaluation, we validated and compared different meta-paths for pseudo relevance feedback ranking. In this part, we also tried to find the optimized seed number for some selected pseudo relevance feedback functions ($fbDocs$). Figure 2 depicts the paper seed number change ($fbDocs$ change from 3 to 100) for ranking MAP performance. Three selected meta-paths and KL-divergency Mixture model [41] are compared. The peak $fbDocs$ number is pointed by an arrow on the Figure. Table 3 compares MAP and NDCG for 18 experimental meta-paths.

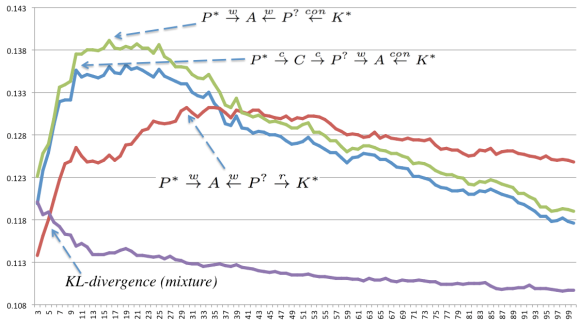

Figure 2: Meta-path and KL-divergency feedback performance (MAP) and seed paper number ($fbDocs$ from 3 to 100).

Based on this result table, we can derive three interesting findings: (1) As Figure 2 shows, different meta-paths have different optimized paper seed node number, which verified our initial hypothesis that different meta-paths need different amounts of information, and they tolerate noisy differently. KL-Divergence feedback performance is lower than meta-path based PRF, and KL-divergence feedback performance consistently decrease when $fbDocs$ increasing; (2) While the performances differ for various meta-paths, we found that, in most cases, complex meta-paths outperform simple ones, i.e., most combined meta-paths outperform the simple ones. Meanwhile, we also found the restrictions (i.e., $A \xleftarrow{con} K^*$ or $C \xrightarrow{m} K^*$), in most cases, are very helpful in enhancing the ranking performance, and (3) Among all the meta-paths, we found that, overall, for different kinds of relationships in the meta-paths, $Citation > Author > Venue$ from ranking performance perspective. This makes sense, in that, citation relation can find the most important papers very closely related to the seed nodes, while venue meta-paths may find too many candidate papers and most of

Table 3: Meta-path feedback comparison

Path	MAP	MAP@5	MAP@10	NDCG	NDCG@5	NDCG@10
$P^* \xrightarrow{p} V \xleftarrow{p} P^?$	0.0112	0.0024	0.0034	0.1110	0.0045	0.0052
$P^* \xrightarrow{p} V \xleftarrow{p} P^?$ $\overset{con}{\nwarrow} K^*$	0.0096	0.0020	0.0029	0.0822	0.0052	0.0072
$P^* \xrightarrow{w} A \xleftarrow{w} P^?$	0.0277	0.0085	0.0129	0.1035	0.0306	0.0394
$P^* \xrightarrow{p} V \xleftarrow{p} P^? \xleftarrow{con} K^*$	0.0405	0.0168	0.0212	0.1450	0.0414	0.0457
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xleftarrow{con} K^*$	0.0406	0.0156	0.0220	0.1145	0.0521	0.0593
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{c} C \xrightarrow{c} P^*$	0.0327	0.0234	0.0300	0.0734	0.0693	0.0748
$P^* \xrightarrow{w} A \xleftarrow{w} P^?$ $\overset{con}{\nwarrow} K^*$	0.0296	0.0105	0.0149	0.1052	0.0335	0.0427
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{r} K^*$	0.0507	0.0252	0.0338	0.1356	0.0865	0.0924
$P^* \xrightarrow{w} A \xrightarrow{co} A \xleftarrow{w} P^?$	0.0436	0.0121	0.0187	0.1672	0.0476	0.0585
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xrightarrow{r} K^*$ $\overset{con}{\nwarrow} K^*$	0.0499	0.0251	0.0326	0.1344	0.0887	0.0888
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$	0.1375	0.0601	0.0821	0.2224	0.1587	0.1623
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{p} V \xleftarrow{con} K^*$ $\overset{m}{\searrow} K^*$	0.0790	0.0421	0.0535	0.1434	0.1117	0.1126
$P^* \xrightarrow{w} A \xleftarrow{w} P^? \xleftarrow{con} K^*$ $\overset{con}{\nwarrow} K^*$	0.0722	0.0417	0.0540	0.1640	0.1335	0.1338
$P^* \xrightarrow{c} C \xrightarrow{c} P^?$ $\overset{m}{\searrow} K^*$	0.1426	0.0622	0.0876	0.2283	0.1611	0.1714
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{w} A \xleftarrow{con} K^*$	0.1356	0.0567	0.0798	0.2222	0.1478	0.1594
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{w} A \xleftarrow{con} K^*$ $\overset{m}{\searrow} K^*$	0.1478	0.0639	0.0913	0.2325	0.1654	0.1785
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xrightarrow{r} K^*$	0.1265	0.0522	0.0723	0.2106	0.1373	0.1511
$P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$ $\overset{m}{\searrow} K^*$	0.1489	0.0672	0.0945	0.2299	0.1724	0.1796

them could be irrelevant to the user information need (i.e., papers published in a relevant venue can be irrelevant). For all 18 meta-paths, $P^* \xrightarrow{c} C \xrightarrow{c} P^? \xleftarrow{con} K^*$ $\overset{m}{\searrow} K^*$ works best. This feedback function indicates that important papers' cited papers could be important, if the citation topical motivation is also important. Meanwhile, the candidate cited paper should make important contribution to the seed topics. Note that, in this meta-path, citation motivation is inferred by citation context, and this information cannot be characterized in classical heterogenous graph, which is generated only by using publication metadata.

[*PRF ranking integration via learning-to-rank*] In this part, we integrated different ranking functions via learning-to-rank. Note that, even Table 3 shows that some meta-path functions are not well performed compared with others, we still use those for ranking integration. Learning-to-rank algorithm will use the ranking feature as long as the feature provide new useful ranking information. For all the PRF methods (text, PageRank, and meta-path based PRF methods), in this experiment, we employed language model with Dirichlet prior smoothing as the initial ranking algorithm for PRF. We employed 10 cross fold validation for learning-to-rank based ranking evaluation.

In this experiment, we employed the following baseline feature groups:

• **T:** text ranking features, including 1. language model (with Jelinek-Mercer smoothing), 2. language model (with Dirichlet prior smoothing), 3. BM25, 4. TFIDF (with Okapi TF).

• **PR:** PageRank ranking (query-independent) with homogeneous paper-paper citation relationship.

• **T-PRF(5):** Text-based pseudo relevance feedback (query expansion) with top 5 feedback papers ($fbDocs = 5$, and feedback term number = 5). We employed 4 ranking features: 1. PRF with cosine similarity. 2. PRF with TFIDF. 3. PRF with KL-divergency Mixture model [41], and 4. PRF with KL-Divergency Minimization Model [41].

• **T-PRF(10):** Text-based pseudo relevance feedback with top 10 feedback papers ($fbDocs = 10$).

• **PR-PRF(5):** PageRank-based pseudo relevance feedback with top 5 feedback papers ($fbDocs = 5$) with personalized PageRank [37] (query-dependent). On the homogeneous paper citation network, the prior probability of $paper_i$ = 0.2 ($P_{prior}(paper_i) = 1/fbDocs$) if $paper_i \in$ seeds, else, paper prior = 0. The PageRank PRF considers three factors: total number of citations (incoming links), citation quality (citing paper authority), and paper relevance score (paper prior).

• **PR-PRF(10):** PageRank-based pseudo relevance feedback with top 10 feedback papers ($fbDocs = 10$) with personalized PageRank. $P_{prior}(paper_i)$ = 0.1, if $paper_i \in$ seeds.

We used two experimental ranking feature groups in this study with meta-path based PRF (18 ranking features listed in Table 2):

• **MP-PRF(5):** meta-path based pseudo feedback with top 5 feedback papers ($fbDocs = 5$). 18 meta-paths listed in Table 2.

• **MP-PRF(10):** meta-path based pseudo feedback with top 10 feedback papers ($fbDocs = 10$).

Experiment result shows that text based PRF (T + T-PRF) cannot outperform text ranking methods (T), which verified previous studies [20] that text based pseudo feedback methods are not robust for some queries and some tasks [10]. Meanwhile, we find citation relationship between papers plus query independent PageRank algorithm (T + PR) on a homogeneous graph can significantly enhance the citation recommendation performance. PageRank based pseudo feedback (T + PR-PRF) via personalized PageRank (query dependent), comparing with other baseline methods, is very effec-

Table 4: Different ranking feature comparison (MAP)

Feature Group	MAP	MAP@5	MAP@10	MAP@50	MAP@100
T	0.1261	0.0595	0.0800	0.1113	0.1188
T + PR	0.1435	0.0665	0.0887	0.1280	0.1364
T + T-PRF(5)	0.1250	0.0575	0.0773	0.1098	0.1174
T + T-PRF(10)	0.1294	0.0617	0.0816	0.1148	0.1220
T + PR-PRF(5)	0.1585	0.0623	0.0996	0.1445	0.1519
T + PR-PRF(10)	0.1550	0.0695	0.0850	0.1407	0.1488
T + MP-PRF(5)	**0.1667**	**0.0826**	**0.1107**	**0.1518**	**0.1599**
T + MP-PRF(10)	**0.1687**	**0.0788**	**0.1086**	**0.1546**	**0.1620**

Table 5: Different ranking feature comparison (NDCG)

Feature Group	NDCG	NDCG@5	NDCG@10	NDCG@50	NDCG@100
T	0.2837	0.1789	0.1858	0.2299	0.2493
T + PR	0.2979	0.1991	0.2029	0.2482	0.2689
T + T-PRF(5)	0.2832	0.1790	0.1848	0.2279	0.2479
T + T-PRF(10)	0.2852	0.1833	0.1879	0.2312	0.2508
T + PR-PRF(5)	0.2973	0.1737	0.2074	0.2540	0.2712
T + PR-PRF(10)	0.2959	0.1882	0.1811	0.2524	0.2716
T + MP-PRF(5)	**0.3153**	**0.2247**	**0.2324**	**0.2696**	**0.2873**
T + MP-PRF(10)	**0.3195**	**0.2329**	**0.2372**	**0.2773**	**0.2928**

tive to further enhance the ranking performance, which is better than text-based PRF. Meta-path based feedback (T + MP-PRF) on heterogeneous graph performs best in all the results, which is significant better than PageRank based PRF (best performed baseline method), with t-test $p < 0.01$. MP-PRF(10), in the result, is better than MP-PRF(5), which indicates a reasonable $fbDocs$ can be helpful to improve the citation recommendation performance.

6. ANALYSIS AND CONCLUSION

In this study we proposed a new ranking method with pseudo relevance feedback by investigating a number of hypothesis driven meta-paths on the scholarly heterogeneous graph. Meanwhile, unlike previous studies, we propose restricted meta-path and combined meta-path facilitated by an innovative context-rich heterogeneous graph generated via full-text publication along with citation context, which enables in-depth heterogeneous graph mining, i.e., citation topical motivation analysis on the meta-path.

Experiment result with ACM full-text data shows that meta-path based PRF is very effective for scholar citation recommendation task compared with text-based and PageRank-based PRF. We assume the main reason is that meta-path based PRF can provide different kinds of novel ranking information from very different perspectives, i.e., from author, venue, citation relation, or citation topic motivation perspectives. For example, in the experiment results, we found author-centric, citation-centric, and venue-centric meta-paths provide very different ranking results, which help learning-to-rank algorithm prioritize important candidate cited papers.

We also found restricted meta-path is efficient for ranking. For instance, we found citation motivation, $C \xrightarrow{m} K^*$, in most cases, can enhance the ranking performance. Similarly, author and venue restrictions ($A \xleftarrow{con} K^*$ and $V \xleftarrow{con} K^*$) can improve the ranking performance. When we investigate the reason in the ranking results, we find restrictions, in most cases, can help to find the most important paths on the graph. Taking paper-author relation as an example, $P \xrightarrow{w} A$ relation identified too many authors on the graph (averagely, one paper has 2.562 authors), and a large number of these authors are not relevant to user information need. By using restriction, $A \xleftarrow{con} K^*$, the new random walk ranking algorithm can identify the most important authors by using the contribution edge between author and topic, which is critical for ranking.

7. LIMITATIONS AND FUTURE WORK

The limitations of this work are twofold. With respect to data, our test corpus came mostly from the ACM DL, from which we cannot access full-text data (and citation context) for all papers. In our experiment we only extracted 67.7% of the papers' full text, and most of those papers were published after 1995 (because old paper PDF files are scanned, we cannot extract text directly from them). These problems need to be addressed in future work.

With respect to PRF experiment, we will use more sophisticated learning to rank method, i.e., Lv et al., [20], to integrate different PRF features. Meanwhile, as experiment result shows that different meta-paths performs differently for different $fbDocs$, we should find a more effective method to tune the $fbDocs$ for each selected meta-path. This parameter tuning process needs additional training.

For future, we will validate the new PRF method in other heterogeneous graphs for other search tasks, i.e., music search, patent search, or web search. As another future work, instead of using PRF, we can expend this work to real user explicit feedback. When the user feedback via search interface is available, we can update the feedback algorithm by investigating the negative feedback ($paper_i$ is irrelevant) or pairwise feedback ($paper_i$ is more relevant than $paper_j$), which may enhance the ranking performance.

8. REFERENCES

[1] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Advances in information retrieval*, pages 127–137. Springer, 2004.

[2] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh. Using citation data to improve retrieval from medline. *Journal of the American Medical Informatics Association*, 13(1):96–105, 2006.

[3] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. *NIST SPECIAL PUBLICATION SP*, pages 69–69, 1995.

[4] K. Chandrasekaran, S. Gauch, P. Lakkaraju, and H. P. Luong. Concept-based document recommendations for citeseer authors. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 83–92. Springer, 2008.

[5] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proc. of the 18th ACM conference on Information and knowledge management*, pages 837–846. ACM, 2009.

[6] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *Computer networks*, 31(11):1467–1479, 1999.

[7] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proc. of the 24th Int. Conf. on Machine learning*, pages 233–240. ACM, 2007.

[8] S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *Proc. of the 27th Int. Conf. on Machine Learning (ICML-10)*, pages 375–382, 2010.

[9] C. Guo, J. Zhang, and X. Liu. Scientific metadata quality enhancement for scholarly publications. In *iConference*, pages 777–780, 2012.

[10] D. Harman and C. Buckley. The nrrc reliable information access (ria) workshop. In *Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 528–529. ACM, 2004.

[11] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *Proc. of the 19th Int. Conf. on World wide web*, pages 421–430. ACM, 2010.

[12] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[13] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM, 2001.

[14] K. S. Lee and W. B. Croft. A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback. *Information Processing & Management*, 49(4):792–806, 2013.

[15] X. Liu, Y. Yu, C. Guo, Y. Sun, and L. Gao. Full-text based context-rich heterogeneous network mining approach for citation recommendation. In *ACM/IEEE Joint Conference on Digital Libraries*, 2014.

[16] X. Liu, J. Zhang, and C. Guo. Full-text citation analysis: enhancing bibliometric and scientific publication ranking. In *Proc. of the 21st ACM Int. Conf. on Information and knowledge management*, pages 1975–1979. ACM, 2012.

[17] X. Liu, J. Zhang, and C. Guo. Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, 64(9):1852–1863, 2013.

[18] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *Proc. of the 18th ACM conference on Information and knowledge management*, pages 255–264. ACM, 2009.

[19] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *Proc. of the 33rd int. ACM SIGIR conf. on Research and development in information retrieval*, pages 579–586. ACM, 2010.

[20] Y. Lv, C. Zhai, and W. Chen. A boosting approach to improving pseudo-relevance feedback. In *Proc. of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174. ACM, 2011.

[21] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[22] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 542–550. ACM, 2008.

[23] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[24] A. Ritchie, S. Robertson, and S. Teufel. Comparing citation contexts for information retrieval. In *Proc. of the 17th ACM conference on Information and knowledge management*, pages 213–222. ACM, 2008.

[25] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3):129–146, 1976.

[26] J. J. Rocchio. Relevance feedback in information retrieval. pages 313–323, 1971.

[27] T. Sakai, T. Manabe, and M. Koyama. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135, 2005.

[28] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, 24(5):355–363, 1997.

[29] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *Proc. of the 10th annual joint conference on Digital libraries*, pages 49–58. ACM, 2010.

[30] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proc. 2011 Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'11)*, Kaohsiung, Taiwan, 2011.

[31] Y. Sun, J. Han, C. C. Aggarwal, and N. Chawla. When will it happen? relationship prediction in heterogeneous information networks. In *Proc. 2012 ACM Int. Conf. on Web Search and Data Mining (WSDM'12)*, Seattle, WA, 2012.

[32] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proc. 2011 Int. Conf. Very Large Data Bases (VLDB'11)*, Seattle, WA, 2011.

[33] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12)*, Beijing, China, 2012.

[34] P. Treeratpituk and C. L. Giles. Disambiguating authors in academic publications using random forests. In *Proc. of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 39–48. ACM, 2009.

[35] S. Vassilvitskii and E. Brill. Using web-graph distance for relevance feedback in web search. In *Proc. of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–153. ACM, 2006.

[36] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[37] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proc. of the ninth ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pages 266–275. ACM, 2003.

[38] H. Xia, J. Li, J. Tang, and M.-F. Moens. Plink-lda: Using link as prior information in topic modeling. In *Database Systems for Advanced Applications*, pages 213–227. Springer, 2012.

[39] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han. Personalized entity recommendation: A heterogeneous information network approach. In *Proc. 2014 ACM Int. Conf. on Web Search and Data Mining (WSDM'14)*, pages 283–292, New York, 2014.

[40] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Recommendation in heterogeneous information networks with implicit user feedback. In *Proc. of 2013 ACM Int. Conf. Series on Recommendation Systems (RecSys'13)*, pages 347–350, Hong Kong, 2013.

[41] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proc. of the tenth Int. Conf. on Information and knowledge management*, pages 403–410. ACM, 2001.