Full-Text based Context-Rich Heterogeneous Network Mining Approach 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

Liangcai Gao
Institute of Computer Science and Technology
Peking University
Beijing, China, 100871
glc@pku.edu.cn

ABSTRACT

Citation relationship between scientific publications has been successfully used for scholarly bibliometrics, information retrieval and data mining tasks, and citation-based recommendation algorithms are well documented. While previous studies investigated citation relations from various viewpoints, most of them share the same assumption that, if $paper_1$ cites $paper_2$ (or $author_1$ cites $author_2$), they are connected, regardless of citation importance, sentiment, reason, topic, or motivation. However, this assumption is oversimplified. In this study, we employ an innovative "context-rich heterogeneous network" approach, which paves a new way for citation recommendation task. In the network, we characterize 1) the importance of citation relationships between citing and cited papers, and 2) the topical citation motivation. Unlike earlier studies, the citation information, in this paper, is characterized by citation textual contexts extracted from the full-text citing paper. We also propose algorithm to cope with the situation when large portion of full-text missing information exists in the bibliographic repository. Evaluation results show that, context-rich heterogeneous network can significantly enhance the citation recommendation performance.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Measurement

Keywords

Citation Recommendation, Full-text Citation Analysis, Meta-Path, Heterogeneous Information Network

1. INTRODUCTION

The volume of scientific publications has increased dramatically in the past couple of decades, which accelerates research and education. However, sheer volume of scholarly publications also challenges classical systems and methods to retrieve and access scientific resources, and it is impossible for a researcher or student to absorb all the new information available in the scientific repository. In order to solve this problem, researchers have proposed a number different solutions to effectively recommend high-quality resources to users. Along with classical textual information retrieval and recommendation approaches, link information such as citation relationships between papers have been proven to be quite effective to enhance the recommendation performance. For instance, PageRank or random walk-based graph mining algorithms [1–3] were well documented for scientific recommendation.

In most previous studies [4], while various methods were used to characterize the citation relationship, however, the basic assumption was oversimplified: all that matters is whether $paper_1$ cites $paper_2$ (or $author_1$ cites $author_2$), regardless of importance, sentiment, reason, topic, or motivation. In practice, the publications that are with full-text information, comparing with simple scholar metadata, convey significant in-depth knowledge for citations. With the help of full-text, it is feasible to locate the citation contexts (of cited paper) for every citation. Note that, one paper may cite some paper multiple times, and the corresponding citation contexts along with citation motivations can be different. We propose to use supervised topic modeling algorithm (labelled LDA [5]) to assign keyword labels and topic probability to each paper and citation. In this way, each citation can be modeled as a node which links to a set of keywords. A heterogeneous information network thus can be constructed by modeling citation in a more in-depth way.

We then use heterogeneous information network mining approach to solve the paper recommendation task. More specifically, the citation relationship between citing and cited publications are repre-

sented by very different kinds of meta-paths, i.e., paper to paper via citation path, and paper to paper via topic motivated citation nodes. Evaluation result shows that full-text (citation frequency + topical citation motivation) empowered heterogeneous graph, in most cases, can significantly enhance the citation recommendation performance, and topical citation motivation based meta-path is helpful.

The main contribution of this paper is threefold. Firstly, we propose innovative heterogeneous scholarly graph to host these two types of knowledge. In this paper, we create three different types of heterogeneous graphs for the same corpus. The difference is focusing on how to characterize the citation relationship between publications. Type I graph, like other classical studies, is based on metadata information. The citation relationship between two papers can be denoted as $P_i \xrightarrow{c} P_j$, which means if paper P_i cites paper P_j, and citation is one directed link from P_i to P_j. For citation frequency hypothesis, we create type II graph, which conveys the citation count, such like $P_i \overset{c}{\underset{c}{\rightrightarrows}} P_j$. Compared to type I, multiple edges could exist between P_i and P_j, and the random walk probability between the paper pair, $RW(P_i^{(1)} \rightsquigarrow P_j^{(l+1)})$, is closely related to the number of edges between them. For type III, we take the topical citation motivation probability into consideration. Unlike type I and II, each citation is represented as a node, i.e., C_{ij_k}, rather than an edge on the heterogeneous graph. Note that, there can be several different citation nodes between two papers, which denotes the citation count in type II. The structure of citation part can be illustrated as $P_i \xrightarrow{c} C_{ij_1} \xrightarrow{c} P_j$ with $C_{ij_1} \xrightarrow{m} K$ and $C_{ij_2} \xrightarrow{m} K$, which shows that each citation C_{ij_k}, from paper P_i to P_j, can be motivated by several keyword labeled topics K, i.e., $C_{ij_k} \xrightarrow{m} K$ (probability that the citation C_{ij_k} is motivated by topic K, inferred from citation context). The random walk between P_i to P_j, then, will be defined not only by the tour between them, but also by the topical path restrictions. For instance, if we know users information need is focusing on topic K_{z_t}, the citation path restricted by K_{z_t} (citation motivated by K_{z_t}) can be more important than others restricted (motivated) by irrelevant topics. The publications with full-text provide us a wealth of information to enrich the topology of heterogeneous graph.

Secondly, while the citation topic motivation is available on the heterogeneous graph, we propose new citation recommendation by leveraging citation topic motivation. Unlike earlier studies, we used citation topic motivation information to calculate the citation (node) prior, and calculate the citation recommendation probability by using a tailored random walk algorithm. Meanwhile, we also propose meta-path based pseudo relevance feedback algorithm for citation recommendation. Unlike prior textual feedback targeting on updating user initial query, we employed graphical feedback via different meta-paths on the heterogeneous graph.

Thirdly, while full-text citation analysis can be useful for citation recommendation, unfortunately, full-text publications are not always readily available in the scientific repository. For instance, a lot of papers published before 2000 do not have full-text information in ACM digital library, and encoding or OCR barriers challenge the full-text data quality (i.e., we cannot locate the citation context in the citing paper). Without full-text citing paper along with citation context, we cannot use above method to infer the citation topic motivation. In this paper, we propose a new method to cope with full-text missing problem. We also used a real-world corpus, where citations with context only account for 16.5% of all the citations, to validate the new method. Evaluation results show that the new graph and recommendation method outperforms classical method and graph even only a portion of publications possess full-text.

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 DEFINITION

In this section, we introduce preliminary knowledge on heterogeneous information network and meta-path, as well as the problem we we are investigating for this study.

2.1 Preliminaries and Definition

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 [6], we define an information network 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.*

Given a complex heterogeneous information network, it is necessary to provide its meta level (*i.e.*, schema-level) description for better understanding the object types and link types in the network. Therefore, we propose the concept of network schema to describe the meta structure of a network. The network schema of a heterogeneous information network specifies type constraints on the sets of objects and relationships between the objects. 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 [6], 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 keywords. In this paper, we extend

the meta-path to restricted meta-path, which can further select the path instances following some constraints that we are most keen on.

2.2 Problem Definition

In this paper, we propose to solve the citation recommendation problem, i.e., recommending citations for a given paper. The required input is a piece of text that briefly summarizes the research work, i.e., paper abstract or research idea description. The optional input is a list of scientific keywords. The output is a list of ranked papers that could potentially be cited given user's input. For instance, for papers from ACM DL, author input could be the paper abstract and paper keywords, and the output is the reference list.

3. RESEARCH METHODS

In this section, we introduce our novel methodology for citation recommendation by using context-rich heterogeneous network mining approach, which includes the following components: to characterize citation topic motivation by using full-text information (3.1), to construct context-rich heterogeneous graph by using citation motivation (3.2), to infer the citation motivation when full-text citing papers are not available (3.3), and to rank the candidate citations on the heterogeneous graph for the given keyword query(3.4).

3.1 Citation Motivation Modeling by Full-Text

Full-text publication alone with the citation context analysis has been used for a number of tasks to cope with the limitations of statistical citation relationship extracted from the scholarly metadata. For this study, based on our earlier studies [4, 7], 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) [5] algorithm, which assigns multiple labels to a citation-based context. In our case, we treat each keyword as a possible topic label. The citation topical motivation is captured by $P(Z_{k_i}|C_j)$, where Z_{k_i} is the topic labeled by keyword k_i provided in citing or cited paper, and C_j is a citation relationship represented by citation context (left and right 300 words, which have been proved useful in [4, 7]) in the citing paper. The citation-topic distribution $P(Z_{k_i}|C_j)$ is considered as the weight between citation node C_j to the keyword node k_i, i.e., $C_j \xrightarrow{P(Z_{k_i}|C_j)} K_i$. Note that, one citation could be motivated by multiple topics, due to the multiple labels generated by LLDA [5]. In this study, we assume that each (author-assigned) scientific keyword is a topic label and that each scientific publication is a mixture of its author-assigned topics (keywords). As a result, both topic labels and topic numbers (the total number of keywords in the metadata repository) are given. The labeled LDA algorithm was used in training the labeled topic model. Unlike the LDA method, LLDA is a supervised topic modeling algorithm that assumes the availability of topic labels (keywords) and the characterization of each topic by a multinomial distribution β_{key_i} over all vocabulary words. More detailed citation topic motivation inference algorithm can be found in [4].

3.2 Heterogeneous Network Construction

In this section, we are going to describe how to create heterogeneous information networks when various types of citation relationships are available. As Figure 1 shows, we can generate three types of heterogeneous information networks by using different types of citation relationships. The main differences are summarized in Table 1.

For Type I and II, citation relationship is modeled as edges between citing and cited paper nodes, and, for Type III, citation is modeled as nodes, which is linked to paper nodes and motivated by keyword nodes, $C \xrightarrow{m} K$. Type III, comparing with the other two types, is much more informative.

Table 1: Citation Relations in Different Types of Heterogeneous Networks

Type	Description	Topological Rep.
I	Citation relation based on metadata	$P_i \xrightarrow{c} P_j$
II	Citation relation with count (extracted from full-text citing paper)	$P_i \overset{c}{\underset{c}{\rightrightarrows}} P_j$
III	Citation Context as a Node	$P_i \xrightarrow{c} C_{ij_1} \xrightarrow{c} P_j$, $C_{ij_1} \xrightarrow{m} K$, $C_{ij_2} \xrightarrow{m} K$, $C_{ij_2} \xleftarrow{c} C_{ij_1}$

3.2.1 Type I: Citation Relation Based on Metadata

Unlike most previous research, we adopt heterogeneous network rather than homogeneous network to present the relationships among the scholarly entities, which will be able to host various ranking hypothesis based on different meta-paths. The graph structure is shown as Figure 1.

Based on the metadata of scholarly publications, the main entities include paper, keyword(topic), author and venue. The relationships can be defined as follows:

Table 2: Relations in the Type I Heterogeneous Graph

Relation	Description
$P \xrightarrow{wr} A$	Paper is written by an author
$P \xrightarrow{p} V$	Paper is published at venue
$A \xrightarrow{co} A$	Co-author relationship
$\mathbf{P \xrightarrow{c} P}$	**Paper cites another paper**
$P \xrightarrow{r} K$	Paper is relevant to keyword(topic) K

For any node on the graph, the sum of the same type of outgoing links equals 1. For instance, the weight of the link from paper p_i to author a_j is defined as $w(p_i \xrightarrow{wr} a_j) = \frac{1}{d(p_i \xrightarrow{wr} A)}$, where $d(p_i \xrightarrow{wr} A)$ is the total number of authors of paper p_i. Similarly, we defined the weights of edges in $P \xrightarrow{c} P$. Since the citation relation is based on metadata only, if paper P_i cites P_j, there is only one link between them. $w(p_i \xrightarrow{c} p_j) = \frac{1}{d(p_i \xrightarrow{c} P)}$, where $d(p_i \xrightarrow{c} P)$ is the total number of reference papers of paper p_i. $w(a_i \xrightarrow{co} a_j) = \frac{d(a_i \xrightarrow{co} a_j)}{d(a_i \xrightarrow{co} A)}$ denotes the weight of the link from author a_i to author a_j, where $d(a_i \xrightarrow{co} a_j)$ represents how many papers a_i has collaborated with a_j. As one paper can only submit to one venue, one citation only points to one cited paper, $w(p \xrightarrow{p} v) = 1$.

3.2.2 Type II: Citation Relation with Count

In Type I, the weight of a citation link from one paper to its reference is the same for all citation links between the same paper and its other references, which equals to $\frac{1}{d(p_i \xrightarrow{c} P)}$. However, if paper p_i cites paper p_j multiple times, we can conjecture that p_j is very important to p_i. Hence, the citation frequency is of great value [8] and we need take it into account. The structure of type II is shown as Figure 1. Once we get the citing paper content, we are able to capture the frequency of each citation relation and generate

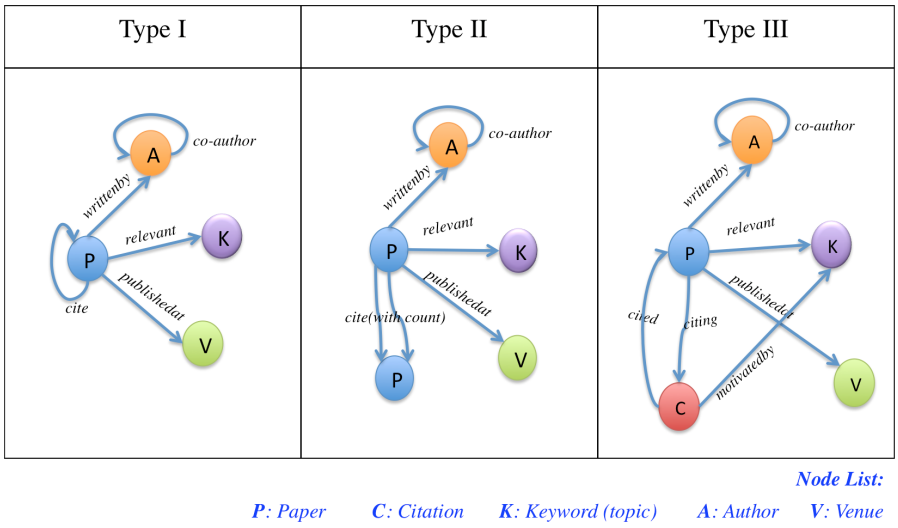

Figure 1: Heterogeneous networks generated from bibliographic data with different citation modeling.

multiple links for it. The weight of each citation link is calculated as: $w(p_i \xrightarrow{c} p_j) = \frac{d(p_i \xrightarrow{c} p_j)}{d(p_i \xrightarrow{c} P)}$.

While in both Type I and II, the weight of $p_i \xrightarrow{r} k_j$ is the LLDA probability of topic k_j given the content of p_i, $P(Z_{k_j}|p_i)$ 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, which has been used in [9].

3.2.3 *Type III: Citation Context as a Node*

In Type III, as Figure 1 shows, the citation is not an edge but a node on the graph. Depending on the citation context window (from citing paper), we can find that each citation conveys different information and focuses on several different topics. So there is possibility for us to characterize the topic motivation distribution for each citation. As aforementioned, we calculate the citation topic motivation results by using LLDA algorithm, $w(c_j \xrightarrow{m} k_i) = P(Z_{k_i}|c_j)$, 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 a citation with the citation context in the citing paper. The relations in Type III are listed in Table 3. There is no direct link from paper to paper, but they are connected through citation node. Each citation on the graph will be related to some topics with weights. Since the citation node only directs to one cited paper, the weight of cited link $w(c \xrightarrow{c} p)$ always equals 1.

Table 3: Relations in the Type III Heterogeneous Graph

Relation	Description
$P \xrightarrow{wr} A$	Paper is written by an author
$P \xrightarrow{p} V$	Paper is 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 is relevant to keyword(topic) K, $P(Z_{k_j}|paper_i)$

3.3 Citation Motivation Estimation

In the real-world scientific repository, it is difficult to access or extract full-text for every paper (especially for those old papers). Language, encoding, format, and OCR barriers bring challenges to characterize topical motivation for all the citation contexts. For Type III graph, when full-text data (citation context) is missing, we need to compromise the method for citation topic motivation inference. In this study, we will verify the usefulness of the proposed new graph in a real-world scientific corpus, where full-text citations are partially missing. The Algorithm 1 depicts the citation topic motivation process with/without full-text of citing paper.

Algorithm 1 Construct Citation Relationship with Topic Motivation (with Imperfect Data)

```
Open repository R and read all the scientific papers
loop For each Citing Paper P_i ∈ R
    if P_i has full-text data then
        Extract citation text context for C_ij, P_i cite P_j;
        Infer citation topic distribution for C_ij, P(Z_k|C_ij), via LLDA inference;
    else
        loop For Each Paper P_j cited by P_i
            Infer citation topic distribution for P(Z_k|C_ij) = 1/(|Z_Pi|+|Z_Pj|);
        end loop
    end if
    Create Edges P_i -c-> C_ij and C_ij -c-> P_j
    Create Edge C_j -P(Z_ki|C_j)-> K_i ▷ Citation motivation link
end loop
```

For each citing paper in the scientific, we need to distinguish two different scenarios in terms of the full-text availability. When we have citing paper full-text, as Algorithm 1 shows, we can estimate the citation topical motivation by using LLDA inference. While full-text of citing paper is missing, we assume all the topics in the citing paper and cited paper (Z_{k_i} and Z_{k_j}) have the equal opportunity to motivate the citation existence, and, then, the citation topical motivation is $P(Z_k|C_{ij}) = \frac{1}{|Z_{P_i}|+|Z_{P_j}|}$. Note that, while the full-text data is missing, we also lose the citation count information.

In that, without full-text, there is only one tour $P_i \xrightarrow{c} C_{ij} \xrightarrow{c} P_j$ between citing and cited paper pair.

The premise of this method is that we use supervised topic modeling to characterize the citation motivation, where keyword (metadata) is the topic label. If we use unsupervised approach, i.e., LSI or LDA, citation topic characterization is always depend on the citation context (text observation), and we can hardly estimate the citation topic information without full-text citing paper.

3.4 Meta-Path-Based Recommendation

We now model the citation recommendation task as a ranking problem in the constructed heterogeneous networks.

As introduced in Section 2, Meta-path defines how nodes are connected in a heterogeneous network, which provides a pattern for lots of path instances of the same kind. There can be many kinds of meta-paths on scholarly heterogeneous network. As we depicted before, for example, $P \xrightarrow{wr} A \xleftarrow{wr} P$ is a simple meta-path on the heterogeneous information networks, denoting all the papers (second P) published by the target paper's (first P) authors (A). Note that the edge direction is ignored in this paper. Once a meta-path is specified, a meta-path-based ranking function is defined so that relevant papers determined by the ranking function can be recommended. In this paper, we adopt a random walk based algorithm to calculate the ranking score of each candidate.

In this study, given user initial textual information need, we first retrieve a number of papers by using text search (language model) to obtain seed paper nodes P^* on the graph. We then use those seed nodes to rank candidate cited papers $P^?$ through the specified different meta-paths. In particular, in order to confine the meta-path into the path instances with certain constraints, we propose to use restricted meta-path in the ranking. For example, we may choose papers in the seeds only to be considered as the first node in the meta-path $P \xrightarrow{wr} A \xleftarrow{wr} P$.

Formally, a restricted meta-path can be represented as:

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

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, denoted as P^*, and type A_{l+1} is the type of nodes to be queried, denoted as $P^?$. For example, $\begin{matrix} P^* \xrightarrow{c} C \xrightarrow{c} P^? \\ \qquad\quad \searrow^{m} K^* \end{matrix}$ is a restricted meta-path from paper type P to paper type P via citation relationship on Type III graph. The constraints are associated with the target papers (P) and citations C on the meta-path, meaning we only consider target papers that are in the seeded paper set and citation nodes that link to the seeded keywords. Note that, formally, the constraint on citation nodes can be represented as $\sigma_{c|\exists k \in K^* \text{ such that } k \to c}$.

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 $\sigma_{S_{l+1}}\dot{A}_{l+1}$ (e.g., the candidate cited papers $P^?$) and objects in $\sigma_{S_1}(\dot{A}_1)$ (e.g., the seed papers P^*):

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

where t is a tour from $a_i^{(1)}$ to $a_j^{(l+1)}$ following the specified restricted meta-path $\mathcal{P}$, and $RW(t)$ is the random walk probability of the tour t. Suppose $t = (a_{i1}^{(1)}, a_{i2}^{(2)}, \dots, a_{il+1}^{(l+1)})$, 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 $\dot{A}_{j+1}$.

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 citation type on Type III graph, a relevance score is also added to these citations 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, on Type III graph, the paper to paper random walk probability via citation node also depends on the citation prior, which is defined by the random walk probability from $C \xrightarrow{m} K^*$, which indicates the citation motivated by an important topic K^* is more important, and the cited paper on this meta-path has the higher chance to be recommended to user.

All meta-paths investigated in this study are listed in Table 4. As the initial seed papers are identified by using text search methods, all these ranking features can also be used as pseudo relevance feedback ranking features. While classical feedback methods optimize the user initial query with feedback documents, we employ meta-path based feedback function on the heterogeneous graph.

4. LITERATURE REVIEW

In this section, we will review previous studies focusing on (1) full-text citation analysis, (2) citation recommendation and citation context, and (3) meta-path-based heterogeneous network mining.

4.1 Full-Text Citation Analysis

In most existing citation network analysis, complex citation behavior is reduced to a simple edge, namely, node A cites node B. The implicit assumption is that A is giving credit to, or acknowledging, B. Previous theoretical studies show that in-depth citation analysis beyond citation metadata is critical for the future retrieval and bibliometrics research. For instance, The Citation Typing Ontology (CiTO) [10] is ontology-based metadata model for citations. CiTO captures the intent of citations and allows authors to categorize the reasons for their citations by providing a taxonomy: confirm, correct, credit, critique, disagree with, discuss, extend, or obtain background from a study, which can hardly be achieved, at large scale, with only scholarly metadata. Similar studies, i.e., The Bibliographic Ontology (BIBO) [11] and the Dublin Core Metadata Element Set [12], also addressed this problem.

Some more recent studies investigated the citation relations by investigating the full-text citing papers. Ding et al. [13], for instance, treated the contributions of all citations (for a citing paper) unequally, and they find some citations appear multiply in a text and others appear only once. They applied easy text search to a relatively large dataset (866 information science articles) to demonstrate the differential contributions made by references. They investigated the placement of citations across the different sections of a journal article and identified highly cited works using two different counting methods (CountOne and CountX). CountX indicates the citation frequency extracted from the citing paper, and they found CountX provides important information. Similarly, [14] investigated how to estimate the strength value (importance) of citation. It found that the simply using citation-count might be not enough, and the importance of each citation is of great use.

Based on these studies, we proposed citation context mining via

Table 4: All the meta-path ranking features used in this study

Meta-path	Hypothesis	
	Type I & Type II	Type III
$[P_{citing} \rightsquigarrow P_{cited}]^?$	Relevant paper's cited papers can be relevant	Relevant paper's cited papers can be relevant, if the citation is motivated by an important topic
$P^* \overset{wr}{\rightarrow} A \overset{wr}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	Relevant paper's author's paper can be relevant, if the candidate paper cited relevant paper	Relevant paper's author's paper can be relevant, if the candidate paper cited relevant paper and the citation is motivated by an important topic
$P^* \overset{p}{\rightarrow} V \overset{p}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	Paper can be relevant if it is published at the same venue as the relevant paper, and the candidate paper cited relevant paper	Paper can be relevant if it is published at the same venue as the relevant paper, the candidate paper cited relevant paper, and the citation is motivated by an important topic

$[P_{citing} \rightsquigarrow P_{cited}]^?$ **represented by:**

Type I: $P \overset{c}{\rightarrow} P^?$ (One path between paper pair)

Type II: $P \overset{c}{\rightarrow} P^?$ (Multiple paths between paper pair)

Type III: $P \overset{c}{\rightarrow} C \overset{c}{\rightarrow} P^?$, $C \overset{m}{\searrow} K^*$

citation topic motivation modeling [4, 7], which investigated the paper ranking problem by using supervised topic modeling. Unlike earlier studies, we used the LDA based inference to estimate the citation motivation. Evaluation based on citation counts shows citation motivation is a useful indicator for ranking.

Unlike all those studies, in this research, we investigate the citation recommendation problem via heterogeneous information network mining approach, which naturally models the scholar information into a network and provides systematic methodology to mine knowledge from such network.

4.2 Citation Recommendation and Context

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. [15], 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. [1] 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 [1], 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. Similarly, [16] integrated linkage weighting calculated from a citation graph into the content-based probabilistic weighting model to facilitate the publication retrieval. The linkage weighting model based on link frequency can substantially and stably improve the retrieval performances. Unsupervised topic modeling is also used for citation analysis [17], 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 [18] 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 [3] 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. However, the citation relationship, on this study, is based on simple reference metadata.

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 [19] 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 [20] 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. [21] 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, along with 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, Robertson, and Teufel [22] and Bernstam et al. [23] have found that citation context can provide important in-

formation 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 [24] used dynamic influence model to characterize scholar impact without using citation information. These studies motivated us to use the 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 pseudo relevance feedback perspective. Meanwhile, we investigate the deep knowledge on the novel graph by utilizing restricted meta-path plus citation motivation modeling. As another critical difference, we cope with imperfect scientific repository, where full-text publications (with citation context) only account for a small proportion of the corpus.

4.3 Meta-Path on Heterogeneous Graph

The concept of meta-path was first proposed in [6], 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 [6], relationship prediction [25,26], user-guided clustering [27], and recommendation [28,29]. 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. Our analysis and conclusions are presented in the next section.

5.1 Data and Network Construction

We used 248,893 publications (as candidate citation collection) from computer science discipline 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 this corpus, we extracted 28,013 publications' full-text (accounting for 11.26% of all the publications), and all other papers have metadata, i.e., titles, abstracts, and paper keywords.

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]*" (ACM style citations) as citations from the content of a publication. Each citation extracted from the publication text was associated with a reference (cited paper ID). A total of 168,554 citation contexts were extracted from the full-text publications by using regular expression, which come from unique 93,398 references. Note that, some references may have been cited more than once in the citing papers.

For the later citation recommendation evaluation, we also use a test collection with 274 papers. The selected papers meet the following conditions: (1) the selected papers are exclusive from the 248,893 publication candidate citation collection; (2) each selected paper has more than 15 citations from the candidate citation collection, and (3) each each paper's abstract has at least 150 words. The paper's abstract is used as a working context to represent a user's information need, and we recommend citations from the candidate citation collection.

5.2 LLDA Topic Model Training

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*".

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 3,910 topics (keywords). These topics were used to infer the publication and citation topic distribution.

5.3 Graph Construction

[240K+ Graph Construction]: For all the papers in the experiment corpus, we construct three heterogeneous graphs for all 240K+ papers by using the methods presented in section 3. All the graphs share the same number of paper, author, venue, and keyword nodes. The difference lies in the citation relationships. More detailed statistics is presented in Table 5. In the Type III graph, LLDA inferred the citations with topic motivation account for 16.5% of all the citations, and we estimated the topic motivation of the rest citations (without full-text citation context) by using citing and cited paper keyword metadata (as presented in Algorithm 1).

[40K+ Graph Construction]: We construct another three graphs by using 28,013 full-text publications along with their cited papers. The total number of papers on the graph reaches 41,370. This smaller experiment collection is a subset of the 240K+ corpus. All the 168,554 citations on type III graph have the full-text LLDA inferred topic motivation. These graphs simulate a perfect scenario, where almost all the citations on the graph have the full-text information. More detailed statistics can be found in Table 5.

Table 5: Graph statistics between two scenarios

Node/Edge	Number	
	240K+	40K+
Paper	248,893	41,370
Author	479,270	63,323
Venue	601	369
Keyword	3,910	3,910
Citation (in type III)	1,018,816	168,554
$P \xrightarrow{c} P$ (in type I)	752,767	93,398
$P \xrightarrow{c} P$ (in type II)	1,018,816	168,554
$P \xrightarrow{wr} A$	631,221	105,992
$P \xrightarrow{p} V$	246,765	41,013
$A \xrightarrow{co} A$	1,507,822	239,744
$P \xrightarrow{r} K$	794,483	587,252
$C \xrightarrow{m} K$	15,165,877	2,855,628

5.4 Experiment Result

The goal of this study is to compare different methods to characterize citation relationships on the heterogeneous graphs. Type I uses classical scholarly metadata, and citing and cited paper is connected by one edge. Whereas Type II graph employs full-text citing publication to extract multiple edges between citing and cited pa-

pers. Unlike Type I & II graphs, Type III offers citation topic motivation, and citation is characterized as a node connected to keyword (topic) nodes.

In this section, we will compare the graphical citation recommendation performance by using three types of graphs constructed for both 40K+ and 240K+ collections. For all the 274 testing papers, we will first retrieve the relevant candidate cited papers by using language model with Dirichlet smoothing. The retrieved candidate papers are treated as the "seed paper nodes" on the graph. In Table 6 and 7, we compared three types of graphs for 40K and 240K corpus. MAP and NDCG performance is reported with seed paper number $|P^*| = 10$.

In Table 6 and 7, we used t-test to verify: (1) if Type II recommendation performance is significantly better than Type I, and (2) if Type III is significantly better than Type II. It is clear, from citation recommendation perspective, in most cases, Type III performance is significantly better ($p < 0.001$) than Type II and Type II outperforms Type I (for both 40K+ and 240K+). In the last two columns, we report the percentage of improvement (Type III outperform Type II and Type III outperform Type I). Averagely, for 40K+ graph (perfect scenario, all the citations have LLDA motivation), Type III outperforms Type II **48.50%** and Type III outperforms Type I **57.78%**. For 240K+ group (imperfect scenario, citation LLDA motivation partially available), averagely, Type III outperforms Type II **34.24%** and Type III outperforms Type I **44.27%**. Based on these numbers, we found when large percentage of full-text citation topic motivation is available, the citation recommendation performance can be improved. However, when citation topic motivation (extracted from full-text citing papers) is only partially available, citation motivation estimation (from citing and cited papers' keywords) is an effective means to enhance the recommendation performance. We assume, if the percentage of the full-text citing papers increases, the recommendation performance will be enhanced.

Note that, the goal of this study is not to compare different meta-paths' or other recommendation methods' performance. Instead, we are more focusing on the performance of each heterogeneous graph and its represented citation relation characterization method. In the future work, we will need to integrate different meta-path, as ranking (or pseudo relevance feedback) feature, by using learning-to-rank method. We will address this problem in the future work section.

In this experiment, we also tested citation recommendation performance for different paper seed numbers, $|P^*|$, with the range from 3 to 60. Figure 2 depicts the citation recommendation performance for meta-path $[P_{citing} \rightsquigarrow P_{cited}]^?$ for 240K+ and 40K+ corpuses, which verifies our earlier finding that Type III is significantly better than other two method with different paper seed numbers (from 3 to 60).

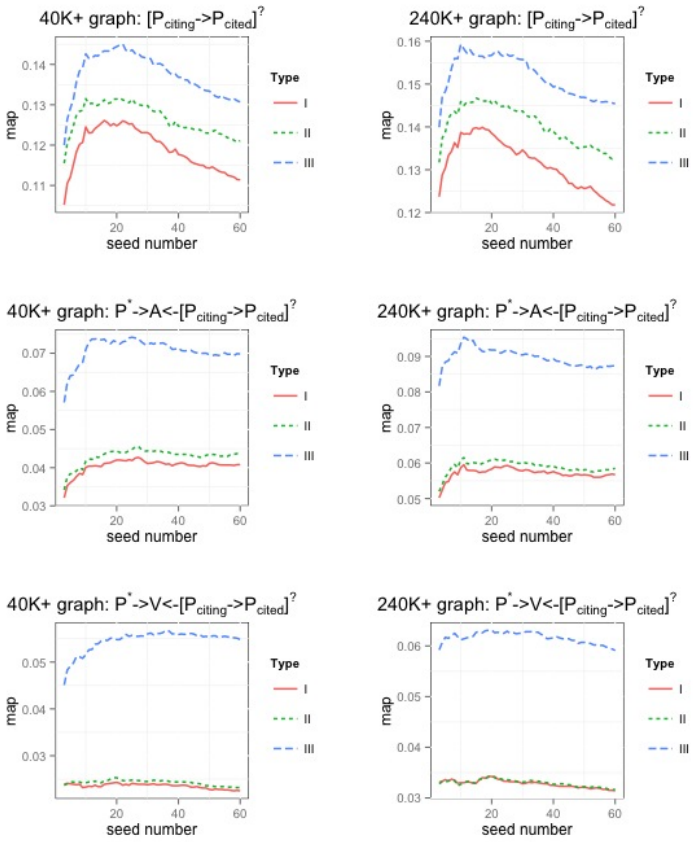

Figure 2: MAP comparison for 40K+ and 240K+ corpuses

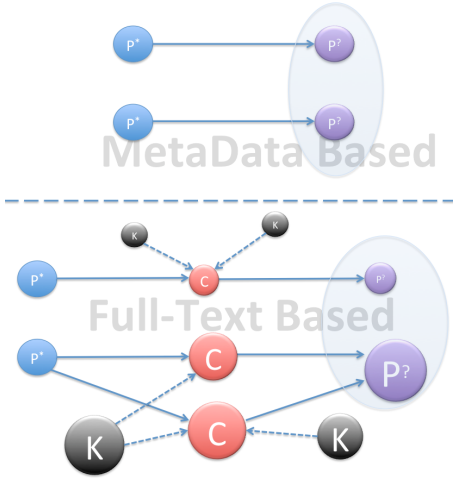

Figure 3: Heterogeneous graph generated via metadata.

6. ANALYSIS AND CONCLUSION

In this study, we propose a novel heterogeneous information network approach (Type III) to host innovative citation information: *citation count* and *citation topic motivation*, and use restricted meta-path-based ranking function to recommend citations to the queries. Evaluation shows that the new approach significantly enhanced the citation recommendation performance. Figure 3 visualizes the reasons why new graph outperforms other baseline methods.

For metadata-based graph, seed paper's credit is equally distributed to all the cited papers, ignoring all the citation path importance (Figure 3 upper part). Full-text based graph, on the contrary, hosts the citation count and citation topic motivation information as additional nodes and edge weights. Then, the random

Table 6: Citation Recommendation Performance Comparison for 40K+ Corpus ($|P^*| = 10$)

40K		Type I	Type II	Type III	III outperform II	III outperform I
$[P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.1245	0.1315 ***	0.1426 **	8.44%	14.54%
	map@10	0.4319	0.4535 *	0.4776	5.31%	10.58%
	map@50	0.3065	0.3249 **	0.3474 *	6.93%	13.34%
	ndcg	0.2088	0.2193 **	0.2283 *	4.10%	9.34%
	ndcg@10	0.1350	0.1548 ***	0.1714 **	10.72%	26.96%
	ndcg@50	0.1930	0.2062 ***	0.2183 **	5.87%	13.11%
$P^* \overset{wr}{\rightarrow} A \overset{wr}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.0401	0.0417	0.0710 ***	70.26%	77.06%
	map@10	0.1874	0.1926	0.2851 ***	48.03%	52.13%
	map@50	0.1478	0.1540	0.2155 ***	39.94%	45.81%
	ndcg	0.1168	0.1219 ***	0.1502 ***	23.22%	28.60%
	ndcg@10	0.0356	0.0426 **	0.0819 ***	92.25%	130.06%
	ndcg@50	0.0662	0.0706 *	0.1138 ***	61.19%	71.90%
$P^* \overset{p}{\rightarrow} V \overset{p}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.0233	0.0242 **	0.0516 ***	113.22%	121.46%
	map@10	0.1260	0.1249	0.2053 ***	64.37%	62.94%
	map@50	0.1025	0.1088 *	0.1613 ***	48.25%	57.37%
	ndcg	0.1525	0.1549	0.1920 ***	23.95%	25.90%
	ndcg@10	0.0218	0.0231 *	0.0524 ***	126.84%	140.37%
	ndcg@50	0.0345	0.0374 **	0.0823 ***	120.05%	138.55%
$p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***						

Table 7: Citation Recommendation Performance Comparison for 240K+ Corpus ($|P^*| = 10$)

240K		Type I	Type II	Type III	III outperform II	III outperform I
$[P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.1387	0.1459 ***	0.1593 **	9.18%	14.85%
	map@10	0.4600	0.4869 *	0.4891	0.45%	6.33%
	map@50	0.3279	0.3429 *	0.3530	2.95%	7.65%
	ndcg	0.2294	0.2428 ***	0.2532 *	4.28%	10.37%
	ndcg@10	0.1427	0.1623 ***	0.1791 **	10.35%	25.51%
	ndcg@50	0.1985	0.2157 ***	0.2306 **	6.91%	16.17%
$P^* \overset{wr}{\rightarrow} A \overset{wr}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.0587	0.0609 **	0.0936 ***	53.69%	59.45%
	map@10	0.2583	0.2734 *	0.3323 ***	21.54%	28.65%
	map@50	0.1915	0.2046 **	0.2535 ***	23.90%	32.38%
	ndcg	0.1707	0.1758 ***	0.2052 ***	16.72%	20.21%
	ndcg@10	0.0586	0.0634 **	0.0974 ***	53.63%	66.21%
	ndcg@50	0.0915	0.0978 ***	0.1398 ***	42.94%	52.79%
$P^* \overset{p}{\rightarrow} V \overset{p}{\leftarrow} [P_{citing} \rightsquigarrow P_{cited}]^?$	map	0.0330	0.0327	0.0612 ***	87.16%	85.45%
	map@10	0.1485	0.1422	0.2081 ***	46.34%	40.13%
	map@50	0.1254	0.1192	0.1636 ***	37.25%	30.46%
	ndcg	0.1931	0.1943 *	0.2356 ***	21.26%	22.01%
	ndcg@10	0.0306	0.0303	0.0575 ***	89.77%	87.91%
	ndcg@50	0.0318	0.0491	0.0923 ***	87.98%	190.25%
$p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***						

walk probability $RW(P_i^{(1)} \rightsquigarrow P_j^{(l+1)})$ is closely related to the number of tours between the citing and cited papers, and the citation node prior between them. As Figure 3 lower part shows, if citation is motivated by important topics, the cited paper's ranking score can be much higher than other cited ones (which is motived by unimportant or less important topics). As another contribution, we find full-text citation analysis is still very useful when only partially full-text data is available in the scientific repository, which can be useful for real world scholar retrieval and recommendation systems. In the experiment, we proposed a compromised solution to estimate the citation topic motivation by using the citing and cited papers' topics. Evaluation shows this method can be very helpful.

Meanwhile, as we used supervised topic modeling algorithm to characterize the citation motivation, the keyword labelled topics are more accurate than unsupervised topic modeling. For instance, the total number of topic is given, and the citing and cited paper's topic labels are available, which is the premise of citation topic motivation estimation without full-text.

7. FUTURE WORK

The most interesting finding of this paper is to find a more accurate way to characterize citation relations on the heterogeneous graph. In the future, we will need to combine different graph based ranking functions (meta-paths) by using learning-to-rank methods. A similar method FeedbackBoost, proposed by Lv, Zhai, and Chen [30], 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 feedback based ranking.

As another direction, we will try to employ more sophisticated graph-based ranking methods to enhance the recommendation performance. Meanwhile, different recommendation tasks, i.e., venue recommendation, author recommendation, and topic recommendation, will be tested.

8. REFERENCES

[1] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 421–430.

[2] Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation recommendation without author supervision," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 755–764.

[3] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine learning*, vol. 81, no. 1, pp. 53–67, 2010.

[4] 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, vol. 64, no. 9, pp. 1852–1863, 2013.

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

[6] 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, pp. 992–1003.

[7] X. Liu, J. Zhang, and C. Guo, "Full-text citation analysis: enhancing bibliometric and scientific publication ranking," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1975–1979.

[8] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, 2013.

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

[10] D. Shotton, "Cito, the citation typing ontology, and its use for annotation of reference lists and visualization of citation networks," in *Bio-Ontologies 2009 Special Interest Group meeting at ISMB*, 2009.

[11] M. Dabrowski, M. Synak, and S. R. Kruk, "Bibliographic ontology," in *Semantic digital libraries*. Springer, 2009, pp. 103–122.

[12] D. C. M. Initiative *et al.*, "Dublin core metadata element set, version 1.1," 2008.

[13] Y. Ding, X. Liu, C. Guo, and B. Cronin, "The distribution of references across texts: Some implications for citation analysis," *Journal of Informetrics*, vol. 7, no. 3, pp. 583–592, 2013.

[14] X. Wan and F. Liu, "Are all literature citations equally important? automatic citation strength estimation and its applications," *Journal of the Association for Information Science and Technology*, 2014.

[15] 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*. Springer, 2008, pp. 83–92.

[16] X. Yin, J. X. Huang, and Z. Li, "Mining and modeling linkage information from citation context for improving biomedical literature retrieval," *Information Processing and Management*, vol. 47, no. 1, pp. 53–67, 2011.

[17] 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*. Springer, 2012, pp. 213–227.

[18] X. Shi, J. Leskovec, and D. A. McFarland, "Citing for high impact," in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 2010, pp. 49–58.

[19] 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*, vol. 2007, no. 06, p. P06010, 2007.

[20] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 233–240.

[21] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550.

[22] A. Ritchie, S. Robertson, and S. Teufel, "Comparing citation contexts for information retrieval," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 213–222.

[23] 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*, vol. 13, no. 1, pp. 96–105, 2006.

[24] S. Gerrish and D. M. Blei, "A language-based approach to measuring scholarly impact," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 375–382.

[25] 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, pp. 121–128.

[26] 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.

[27] 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, pp. 1348–1356.

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

[29] 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)*, New York, 2014, pp. 283–292.

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