Citation Prediction in Heterogeneous Bibliographic Networks

Xiao Yu Quanquan Gu Mianwei Zhou Jiawei Han
University of Illinois at Urbana-Champaign
{xiaoyu1, qgu3, zhou18, hanj}@illinois.edu

Abstract

To reveal information hiding in link space of bibliographical networks, link analysis has been studied from different perspectives in recent years. In this paper, we address a novel problem namely citation prediction, that is: given information about authors, topics, target publication venues as well as time of certain research paper, finding and predicting the citation relationship between a query paper and a set of previous papers. Considering the gigantic size of relevant papers, the loosely connected citation network structure as well as the highly skewed citation relation distribution, citation prediction is more challenging than other link prediction problems which have been studied before. By building a meta-path based prediction model on a topic discriminative search space, we here propose a two-phase citation probability learning approach, in order to predict citation relationship effectively and efficiently. Experiments are performed on real-world dataset with comprehensive measurements, which demonstrate that our framework has substantial advantages over commonly used link prediction approaches in predicting citation relations in bibliographical networks.

1 Introduction

Searching for related scientific literatures (a.k.a, lit. search), is the first and essential step for nearly all scientific research disciplines. Researchers want to find highly related publications in terms of research fields and topics, so that they can learn from related work, compare with previous methods as well as develop new research ideas. However, with the rapid development of science and engineering, a gigantic number of research papers are published each year on various research topics and areas. It is impossible for researchers to follow or read all publications in his/her research fields. Hence a system which could help scientific researchers organize relevant publications is in high demand. Such a system should be able to retrieve high quality publications given research topics, and also measure the relevance between existing publications and researcher's current work.

Google Scholar, PubMed and other key-word-based literature search tools allow users to query publications based on key-word and properties associated with the target papers, e.g., author information, time period of publication, etc. They also provide related articles by measuring document similarity between papers. Although these systems find relevant papers and make lit. search easier than before, key-word-based approach still returns thousands or millions of relevant papers. For example, Google Scholar returns more than 2 million papers with the query "link prediction", and more than 5 million results with the query "citation". Researchers can easily be drown by this huge amount of relevant papers returned by key-word-based lit. search systems. Instead of going through a large number of papers which match query key-word, researchers prefer to only review a relatively smaller number of publications closely related to their research topics, of high quality and also closely related to their research community, so that they can use as references or citations directly. To meet this requirement, we here study citation prediction problem on bibliographic information network. Aiming to help researchers find highly related publications effectively and efficiently, we propose a new citation prediction model and use this model to answer citation queries.

Citation prediction aims at revealing the citation relationship on bibliographic network. Yet it is different from other link prediction problems in this network. For example, citation relationship is directed, while friends recommendation or co-authorship prediction methods are predicting undirected relations. Links among co-authors in publication datasets tend to form communities and have high structural locality, because authors tend to collaborate with researchers within their own research group or with researchers they collaborated before. However, high quality and relevant papers can be from anywhere. Due to the evolution of on-line library systems, scientific researchers can easily get access to nearly all digitized publications. They can find and review any relevant previous paper on bibliographic network, which causes a relatively sparse yet even distribu-

tion in citation link space. Traditional link prediction methods [10] [5] commonly rely on locality assumption, which makes these methods ineffective for citation prediction.

On the other hand, citation prediction methods should be able to measure document similarity and capture topic distribution in bibliographic information network. However, solely relying on topic modeling methods is not sufficient for citation prediction either. Although the number of previous papers is tremendous, research topics are comparably limited. Hundreds or even thousands of papers could share a same topic, which makes topic similarity a very weak evidence in terms of citation relationship inference. Additionally, many critical features which might be more related to citation prediction cannot be represented by topic similarity either. For example, if one paper is written by a well-known researcher in the field, the probability of this paper getting cited by a future publication is higher than a paper by a new researcher. Similarly, ranking of the publication venue (conference/journal) and reputation of a research group all affect citation probability. Furthermore, researchers tend to cite papers of their own, papers within their research groups, or papers of their peers for different reasons. All these heuristics are hiding among bibliographic network structure and none of which can be represented using topic similarity.

In this paper, we study how to predict citation relationship in bibliographic information network effectively and efficiently. We propose a novel two-step approach, attempt to capture both topic and document similarities as well as hidden network structures that are sensitive to citation relationship, and use this approach to setup a citation query processing system. Given author information, target publication venues and certain text description, e.g., title and abstract, of a query paper, our citation query system searches papers in a publication network and returns a list of relevant papers, ranked by the probability of being cited by the query paper. In order to answer citation queries fast and accurate, we propose a two-step approach. First, we build discriminative term buckets, which can capture document and topic similarities without breaking possible citation relations, and put papers into different buckets, which reduces search space for both model learning and citation query answering. Second, we set up a meta path-based feature space to interpret hidden network information in bibliographic dataset, and define citation probability with meta path-based features. With the help of discriminative topic buckets and meta path-based feature space, it is now possible to learn a citation prediction model and use this model to answer citation queries.

The major contributions of this paper are summarized as follows.

- We propose a new problem of citation prediction in a bibliographic network, and analyze the differences between this problem and the related work, e.g., traditional link prediction solutions.
- We propose a new data structure namely discriminative term buckets in order to capture both document similarity and potential citation relationship, and compare this method with traditional topic modeling approaches.
- We propose to use a meta path-based feature space to interpret structural information in citation prediction, and define citation probability within the scope of meta path-based feature space.
- Experiments on real dataset show that we can predict citation relationship with high accuracy and efficiency compared with the state-of-the-art link prediction methods.

In the rest of the paper, we first introduce the background and preliminaries about bibliographic information network in Section 2. We next discuss discriminative term bucket data structure and present how to construct term buckets efficiently in Section 3. Meta path-based feature space building and citation prediction model learning are described in Sections 4 and 5. Experiments and results are presented followed by related work and conclusions.

2 Background and Preliminaries

A citation prediction problem is defined on bibliographical dataset, which can be formatted into a heterogeneous information network. In this section, we briefly introduce some concepts related to information network and the citation prediction problem.

A heterogeneous information network is a directed graph, which contains multiple types of entities and/or links. In order to study meta path-based feature space and discuss citation prediction model, we first introduce the definitions of information network and network schema, which are defined in [13] and [12].

Definition 2.1. (Information Network) *An information network is defined as a directed graph $G = (V, E)$ with an entity type mapping function $\phi : V \to \mathcal{A}$ and a link type mapping function $\psi : E \to \mathcal{R}$, where each entity $v \in V$ belongs to one particular entity type $\phi(v) \in \mathcal{A}$, and each link $e \in E$ belongs to a particular relation type $\psi(e) \in \mathcal{R}$.*

When the types of entities $|\mathcal{A}| > 1$ and also the types of relations $|\mathcal{R}| > 1$, the network is called

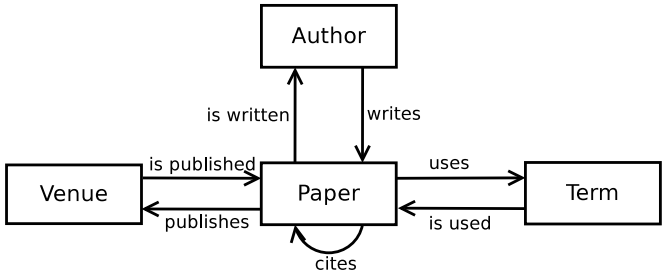


Figure 1: DBLP network schema

Table 1: List of Notations

Notation	Description
$\mathcal{A}$, $\mathcal{R}$	types of entities and relations
$\mathbf{F}$	meta path-based feature space
$\mathbf{P}$, $\mathbf{M}$	meta paths and measures
$\mathbf{T}$	training dataset
$\mathbf{D}$	document collection
$\mathbf{B}$	term bucket set
θ_f	weight for meta path-based feature f
CDM	citation discriminative measure

heterogeneous information network; otherwise it is a *homogeneous information network*.

In this information network definition, we specify both the network structure and types for entities and links. Also, one can notice that types for two entities associated to one link can be different. Without loss of generality, we denote the relation associated to a link as $R(\phi(v), \phi(v')) = \psi(e)$, where v and v' are the two entities associated with link e. We use $dom(R) = \phi(v)$ to denote the domain of R, and $range(R) = \phi(v')$ as the range. We use R^{-1} to denote the inverse relation of R, so $dom(R) = range(R^{-1})$ and $range(R) = dom(R^{-1})$.

Definition 2.2. (Network Schema) *The network schema is a meta template of a heterogeneous network $G = (V, E)$ with the entity type mapping $\phi : V \to \mathcal{A}$ and the link mapping $\psi : E \to \mathcal{R}$, which is a directed graph defined over entity types $\mathcal{A}$, with edges as relations from $\mathcal{R}$, denoted as $T_G = (\mathcal{A}, \mathcal{R})$.*

The definition of network schema is similar to the ER (Entity-Relationship) model in database systems. It serves as a template for a concrete network, and defines the rules of how entities exist and how links should be created. An example of heterogeneous information network and the related network schema can be found as follows.

Example 2.1. *DBLP*[1] *(Digital Bibliography & Library Project) is a computer science bibliographic dataset, which can be described as a heterogeneous information network. It contains four different types of entities (papers, venues, authors and terms). Links exist between papers and authors, papers and venues, papers and terms as well as within paper entities, representing citation relations.*

In this study, we are going to use a subset of DBLP information network, which exactly follows the network schema presented in Figure 1. Besides the citation relationship, defined as a directed meta path from the node "Paper" to itself, which is the relationship we attempt to predict, we also have links between "Term" and "Paper", "Author" and "Paper", as well as publication "Venue" and "Paper", all of which are observable during the citation prediction process. Notations used in definitions as well as the rest part of the paper can be found in Table 1.

3 Discriminative Term Bucketing

For most of citation relations, the prerequisite is a positive topic and/or term correlation between these two papers, i.e., these two papers belong to the same research area, share similar research topics or try to solve a similar problem. Naturally, the first step of our citation prediction framework is to catch such topic or term correlation and be capable of measuring document similarity in the DBLP information network. In order to achieve this requirement, topic modeling methods, which apply and estimate statistical models document collections, toward unfolding hidden topics, are intuitive and popular solutions. By comparing topic distributions of paper pairs, one can calculate document correlation easily. However, topic modeling methods might not be suitable for citation prediction for two reasons. First, topic granularity is hard to guess on an unknown document collection. Second, topics might not be citation relationship discriminative. For instance, one topic might have high weights towards words like "database", "query" and "index", which makes this topic too broad. Two papers belonging to this topic does not imply a potential citation relationship.

Aiming to capture citation discriminative term correlation and measure document similarity in the DBLP information network, we propose a novel method named discriminative term bucketing (we refer this method as term bucketing in the rest of the paper). Very similar to the input and output formats of topic modeling methods, given certain document collection $\mathbf{D}$, term bucketing generates a number of buckets (similar to topics) which contain a set of terms, and also papers in the

[1] http://www.informatik.uni-trier.de/~ley/db/

collection can be distributed into different term buckets. What's more, one paper can belong to multiple buckets. Within each bucket, papers have a positive document similarity and also the existence of citation relationship probability is also higher than papers which do not share buckets.

Discriminative term bucketing contains three steps. First, we identify discriminative terms using a link space discriminative measure, which identifies potential citation relationship, and then by treating discriminative terms as bucket seeds (one seed per bucket), and applying term expansion technique, we collect more terms for each bucket. Finally, we can distribute the entire paper dataset into different buckets and finish the building of term buckets.

In order to measure the ability of identifying citation relations for each term, we first define citation discriminative measure as follows. By generating term paper inverted index in $\mathbf{D}$, for each term t, we collect all papers with t in them, and denote this paper set as P_t. By treating each paper in P_t as a node, all the possible edges within P_t form the complete link space for P_t, denoted by G_t. Consider citation relations within G_t as positive and the rest links as negative, we define the positive-negative ratio as the citation discriminative measure (CDM) for terms, as in Equation 3.1.

$$\forall t \in \mathbf{T}, CDM_t = \frac{count(G_t, +1) + 1}{count(G_t, -1) + 1} \tag{3.1}$$

where $\mathbf{T}$ is the training document collection for citation prediction, G_t is a $|P_t| \times |P_t|$ matrix which contains link labels in the complete link space and $count(G, label)$ counts the number of element in G that equals to $label$.

After calculating CDM for each term, we can pick up terms with CDM higher than a pre-defined threshold and use these terms as discriminative bucket seeds. And each seed term t defines a discriminative term bucket B_t. One should notice that, CDM is calculated on the training dataset, and utilizes label information in order to pick up terms with sufficient citation information. However, in order to reduce search space in both model training and query answering processes, we need to categorize all papers in both training and testing datasets and put them into corresponding term buckets. Training and testing datasets are independent so that term distribution over these two sets might be different, and it is possible that, discriminative terms generated in the training set might not even exist in the testing set. So if term buckets only contain terms in the training dataset, categorizing testing papers will be difficult. In order to propagate citation discriminative information to the testing set, we use term expansion technique to find more terms in $\mathbf{D}$ (from both training and testing datasets) for each bucket.

Term expansion is used to expand the discriminative term bucket by introducing more terms which have a high mutual information with the bucket seed into each bucket. The mutual information of two terms can be used to measure the mutual dependence. The heuristic is, seed terms are terms with citation information, if other terms are highly dependent on seed terms, they should contain citation information as well. For a specific discriminative term bucket B_{t0} which contains bucket seed t_0, we iterate all terms in document set $\mathbf{D}$, and calculate mutual information between each term t and t_0 using Equation 3.2.

$$\forall t \in \mathbf{D}, I(t_0, t) = Pr(t_0, t) log(\frac{Pr(t_0, t)}{Pr(t_0)Pr(t)}) \tag{3.2}$$

where $\mathbf{D}$ is the entire document collection, $Pr(t_0, t)$ is the probability of both t_0 and t appear in one document, and this probability can be estimated using the number of documents which contains both t_0 and t and also the total number of documents in the collection. Similarly, $Pr(t_0)$ and $Pr(t)$ are the marginal probability density functions of terms t_0 and t respectively.

For each discriminative term bucket, we select terms which have high mutual information score with the term seed using a pre-defined threshold MI. By adding these terms into term buckets, now each term bucket has multiple discriminative terms. And based on the term distribution, we can categorize both training and testing papers into different buckets by checking whether a certain paper contains one or more terms in a bucket. What's more, using this categorization method, one paper can be assigned to more than one term buckets as well. Within each bucket, papers share terms as well as topic information so that they have high document similarity, and also papers have a higher probability of been cited by other papers in the same bucket. Term bucketing partitions the entire paper dataset into different buckets, and our citation prediction framework will only search within relevant buckets while learning prediction model and answering citation queries, which reduces search space, and based on our experiments, this approach improves both accuracy and efficiency.

One should notice that, MI is a very important parameter in citation prediction framework. If MI is too low, the number of terms in each bucket will increase severely, which will increase search time complexity. If MI is too high, the number of terms in each bucket will decrease exponentially, which means the ability of capture potential citation relations will be decreased dramatically and the overall prediction performance will be effected. We will discussion this issue with one experiment in Section 6.

4 Meta Path-Based Feature Space Building

After constructing discriminative term buckets and categorizing both training and testing papers into corresponding buckets, we introduced both document similarity as well as citation information into our framework. Papers within the same bucket or share a number of buckets usually share similar research topics and also have a higher probability to be cited by each other compared with other paper pairs. Although term bucket structure helps reduce search space, the number of citation relations within each bucket is still very limited, we need structural features to capture more citation information within term buckets as well as a robust statistical model to improve the prediction accuracy. In this section, we discuss how to define meta path-based features and how to build a comprehensive feature space to define structural similarity between papers in the DBLP information network in order to predict citation relations, and we will introduce citation prediction model learning in the next section.

4.1 Meta Path In a heterogeneous information network schema, two entity types can be connected via different paths [2], which usually carry different semantic meanings. In order to distinguish from path instances in a concrete network, we use meta path definition from [13] in a network schema as follows.

DEFINITION 4.1. (META PATH) *A meta path* $P = A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} ... \xrightarrow{R_l} A_l$ *is a path defined on the graph of network schema* $T_G = (\mathcal{A}, \mathcal{R})$*, which defines a new composite relation* $R_1R_2...R_l$ *between type* A_0 *and* A_l*, where* $A_i \in \mathcal{A}$ *and* $R_i \in \mathcal{R}$ *for* $i = 0, \ldots, l$*,* $A_0 = dom(R_1)$*,* $A_l = range(R_l)$ *and* $A_i = range(R_i) = dom(R_{i+1})$ *for* $i = 1, \ldots, l-1$*.*

Notice that a meta path represents a new composite relation over A_0 and A_l, and we denote $dom(\mathcal{P}) = A_0$ and $range(\mathcal{P}) = A_l$.

Different meta paths can capture different semantics in a heterogeneous information network. Consider DBLP network schema in Figure 1, multiple meta paths can be defined from paper type to paper type, and some examples can be found as follows:

$$P_1 : paper \xrightarrow{PublishedIn} venue \xrightarrow{PublishedIn^{-1}} paper$$

$$P_2 : paper \xrightarrow{Contains} term \xrightarrow{Contains^{-1}} paper$$

Meta path P_1 can capture the relationship of publishing in the same venue for two papers. Intuitively, if two papers have the relationship defined by P_1, it means these two papers belong to the same research area. Very similarly, P_2 defines the term similarity between two papers. One can notice that, different meta paths can capture different relationship information hiding in the network, which can be extremely helpful in order to capture structural information between two papers. In the rest of this paper, we omit the relation name and use the abbreviation of node type to represent a meta path. P_1 will be represented as $P - V - P$, and P_2 will be written as $P - T - P$.

Consider a meta path $P = A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} ... \xrightarrow{R_l} A_l$, if path $p = (a_0a_1...a_l)$ in the corresponding information network follows meta path P, i.e., for each node a_i in path p, we have $\phi(a_i) = A_i$, we call p a *path instance* of meta path P. Similarly, we define $dom(p) = a_0$ and $range(p) = a_l$. A path instance carries relationship information follows the relation defined by its meta path in the network.

4.2 Meta Path-Based Measures The other component of a meta path-based feature is the measure associated with meta paths. Various measures can be defined and implemented in order to measure the similarity or proximity between the query entities and potential result entities given the same meta path. Some of the meta path-based measures can be found as follows.

- Count: the number of path instances between the query entity q and potential result entity r:

$$count(q, r) = |p : dom(p) = q, range(p) = r, p \in P|$$

- Personalized PageRank score [2]. Personalized PageRank can be viewed as converged random walk with restart score which follows certain meta path.
- Random Walk score [10]. Random Walk score is the random walk score along certain meta path with a pre-defined step length.
- PathSim score [13]. PathSim is a newly proposed similarity measure that captures the semantics of similarity between peers, which is a normalized version of count of path instances between entities following the given meta path.

A comparison between these measures can be found in Table 2. Notice that some measures are valid only on a certain type of meta paths. For example, PathSim are only valid for symmetric meta path, which guarantees that the similarity between two entities is symmetric.

4.3 Meta Path-Based Feature Space With meta path and meta path-based measure defined above, a

[2] Similarly, one entity type can be connected via loops.

Table 2: Path-Based Measure Comparison

Name	Symmetric?	Range
Count	Yes	$[0, \infty)$
Personalized PageRank	No	$[0, 1]$
SimRank	Yes	$[0, 1]$
PathSim	Yes	$[0, 1]$

meta path-based feature space $\mathbf{F}$ can be represented as a Cartesian product of the two sets:

$$\mathbf{F} = \mathbf{P} \times \mathbf{M} \tag{4.3}$$

where $\mathbf{P}$ is the set of possible meta paths and $\mathbf{M}$ is the set of possible meta path-based measures. One one hand, the meta path used in a feature represents the relation we are interested in between the entities; on the other hand, different measures can be defined on the same meta path, showing different aspects of quantities of the relations. A combination of the two represents a unique angle of similarity measure between two entities.

In a small heterogeneous network with a simple schema, in order to generate a comprehensive feature space, one can enumerate all meta paths with a length constraint but it is impossible to permute all meta paths in general. And also, it is not necessary to generate all meta paths since some paths does not carry sufficient semantic meanings as others.

For instance, in order to measure the similarity between two authors in the DBLP dataset, two possible meta paths can be generated as follows:

$$P_1 : A \to P \to V \to P \to A$$

$$P_2 : A \to P \to V \to P \to T \to P \to A$$

By measuring similarity along P_1, one can identify authors who published in similar conferences, which can be used to indicate authors who have similar research interests or focus on the same research area. Although P_2 is more complicated than P_1 and also can be used to measure similarity between authors, the semantic meaning of this meta path is not clear, so it is hard to utilize this meta path in any concrete scenarios.

In this paper, rather than spending much time to calculate meta path-based features along meaningless meta paths, we select a subset of meta paths with clear semantic meanings and use this subset to finally learn our meta path-based ranking models for each intention.

5 Learning Citation Prediction Model

After building discriminative term buckets and meta path-based feature space, we can now define and learn citation prediction model within the scope of term buckets using meta path-based features as representation of structural information hiding in an information network. With the learned probability and reduced search space by applying term buckets on potential paper candidates, our framework can answer citation prediction queries efficiently with a high accuracy.

5.1 Citation Probability As discussed in Section 1, citation probability should be a combination of both document similarity and structural information in a publication network. By utilizing discriminative term buckets and meta path-based features, we can define citation probability as follows.

$$Pr(label = 1|p^{(1)}, p^{(2)}; \boldsymbol{\theta}) = \frac{e^z}{e^z + 1} \tag{5.4}$$

where $z = \Sigma_{f_i \in F\prime} \theta_i \cdot f_i$. $Pr(label = 1|p^{(1)}, p^{(2)}; \boldsymbol{\theta})$ is the probability that paper $p^{(1)}$ cites paper $p^{(2)}$. In the definition of citation probability, F' is the feature space defined on the DBLP heterogeneous information network in order to capture citation related information, which is defined as $F' = F \cup F_0$, F is the meta path-based feature space defined in equation 4.3 and F_0 is a set of numerical features for target paper candidates. θ_i is a normalized weight value for feature f_i which indicates which feature is more important for citation prediction. $\boldsymbol{\theta}$ and F' form a linear citation prediction model.

F_0 contains numerical attributes for target paper candidates only. In order to generate a more comprehensive feature space, we add non-meta path-based numerical features including average H-Index [6] as well as publication venue ranking [14] on target paper side. These two numerical features help boost prediction accuracy as supplements to meta path-based feature space. H-Index measures both productivity and impact of the published work of a scientist or scholar, and publication venue ranking measures the reputation and the quality of the research paper indirectly. Both measures should be positively correlated with citation relationship. If the authors of one paper have a higher average H-Index and this paper is published in a highly ranked conference, the probability of this paper being cited by the query paper would be high as well.

In order to learn the citation prediction model, we generate training dataset which contains positive and negative examples of citation relations. However, the DBLP information network is extremely large and the citation relations are very sparse and limited. So search on the entire paper network can be time consuming and ineffective. Randomly generated negative examples can be arbitrary and contain very little representative

information. In order to collect high quality training dataset, we use discriminative term buckets as a filter to first reduce the size of information network, and only generate positive and negative training examples within term buckets. In this way, both positive and negative paper pairs have high document similarity and also high probability of citation relationship. Learning models on such training dataset can help prediction model capture trivial and detailed information hence improve the performance of citation prediction model.

We define the training dataset as follows.

$$\mathcal{T} = \{(p_i^{(1)}, p_i^{(2)}, label) | \exists B_t, p_i^{(1)} \in B_t, p_i^{(2)} \in B_t\} \tag{5.5}$$

where $p^{(1)}$ and $p^{(2)}$ are papers on information network, and they should both belong to at least one discriminative term bucket.

In order to learn citation prediction model, we use logistic regression with L_2 regularization to estimate the optimal $\boldsymbol{\theta}$ given a training dataset $\mathcal{T}$.

$$\hat{\boldsymbol{\theta}} = argmin_{\boldsymbol{\theta}} \Sigma_{i=1}^{n} -\log Pr(label | p_i^{(1)}, p_i^{(2)}; \boldsymbol{\theta}) + \mu \Sigma_{j=0}^{d} \theta_j^2 \tag{5.6}$$

With this objective function defined in Equation 5.6, weights in the citation probability can be easily estimated with a number of optimization methods. We use standard MLE (Maximum Likelihood Estimation) in our experiments to derive $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood of all the training pairs.

5.2 Citation Prediction Model After learning the citation probability defined in Equation 5.4, we can now define a citation prediction model, and use this model to prediction possible citation relations or answer citation queries. The citation prediction model is defined in Equation 5.7.

$$cs = log(1 + cbn(p^{(1)}, p^{(2)})) \cdot Pr(label = 1 | p^{(1)}, p^{(2)}; \hat{\boldsymbol{\theta}}) \tag{5.7}$$

where cs is short for citation score, and $cbn(p^{(1)}, p^{(2)})$ defines the number of common discriminative term buckets shared by papers $p^{(1)}$ and $p^{(2)}$.

Given a citation query paper p^*, we first look up p^* in discriminative term bucket set $\mathcal{B}$, and find all buckets containing paper p^*. We define this subset as B_{p*}. And then, we collection all papers belong to term bucket subset B_{p*}, and denote this paper subset as $P(B_{p*})$. For each paper p in $P(B_{p*})$, we calculate $cs(p*, p)$, and assign this score to each paper p. By ranking paper set $P(B_{p*})$, our citation prediction framework generates a ranked list of papers as the answer to query p^*.

One should notice that, we only calculate cs between p^* and papers which at least share one term bucket with p^*, which is only a subset of the entire paper dataset. In the next section, we conduct a set of experiments and compare our citation prediction framework with the state-of-the-art link prediction methods, which shows that our approach can find citation relations accurately and efficiently.

6 Experiments

In this section, we apply our citation prediction approach along with two state-of-the-art link prediction methods on a DBLP citation dataset generated by Tang et al. [15]. We compare our methods with link prediction methods under different scenarios using a set of experiments.

6.1 Dataset and Methods Setup The original DBLP dataset does not contain citation relations. Tang et al. extracted citation information from other sources and generated a DBLP citation dataset. We use the citation information in this dataset as training examples as well as ground truth to verify the output of different methods. Instead of using the entire dataset, we generated a subset which contains 464 publication venues, 29,615 papers and 215,502 citation relations. Papers in this subset focus on one of the four areas: data mining, database, information retrieval and artificial intelligence. Due to the high coherence of these research areas, citation relation distribution in this subset is very scattered, which makes the task difficult and challenging. We convert this subset into a heterogeneous information network. This information network contains author, paper, publication venues, term as entities (publication year as attribute), as well as paper author relations, paper venue relations, paper term relations as well as citation relations as links, all together 83,625 entities and 682,725 links.

In this study, in order to comprehensively describe different relationships between paper entities in the DBLP heterogeneous information network, we utilize seven different meta paths between paper entities, which are $P-A-P$, $(P-A-P)^2$, $(P-A-P)^3$, $P-C-P$, $P-T-P$, $(P-T-P)^2$ and $(P-T-P)^3$. We use two meta path compatible measures, which are newly proposed PathSim [13] measure and also random walk measure. By combining seven meta paths and two meta path-based measures together, we get total 14 different meta path-based features. As mentioned in the previous section, we also use two numerical features in our feature space in order to have certain bias towards target paper candidates, which are average author H-Index and publication venue ranking score. In discriminative term bucket building step, in order to capture most term information, we use 0 as citation discriminative measure threshold (Equation 3.1) and 0.0003 as mutual

information threshold during term expansion (Equation 3.2). With this setting, discriminative term buckets can capture nearly 90% of citation relations, i.e., for each citation relation in 90% total relations, the two papers which defined this relation can be found in at least one term bucket. What's more, search space can be reduced by 40% if our citation prediction framework only searches within term buckets.

The state-of-the-art link prediction methods which we compare with our framework in these experiments are personalized PageRank (denoted as *pp* in figures and tables) [2] [18], and path-constrained random walk [9] (denoted as rw or refered as trained random walk in figures and tables). Personalized PageRank is a very popular similarity or proximity measuring method which has been widely applied on different problems including link prediction, friend recommendation, as well as network clustering. As an unsupervised method, personalized PageRank simulates information passing along links between entities, and estimates similarity by calculating reachability from query node to the other nodes in the network. Similar to our approach, path-constrained random walk method is a supervised method, which first calculates random walk similarity along different paths, and then assigns different weights to different paths by learning using user-provided examples. We also use random walk features along different paths as part of our meta path-based feature space. In order to have a fair competition, we use the same set of random walk features in both our approach and path-constrained random walk, and also we use the same training dataset and learning method for both approaches as well.

6.2 Measure as Classification Problem We first compare our method with path-constrained random walk by modeling citation prediction as a classification problem. Considering the sparseness of citation relations on the entire search space, we first generate a biased sample on $\langle paper, paper \rangle$ search space as defined in the previous section. In the biased sampled data, we have 45% positive labels, i.e., paper pairs which actually define citation relations, and 55% negative labels, i.e., paper pairs which do not possess citation relation. In order to measure the prediction accuracy, we use five fold cross-validation to assess the quality of each method. Since instances in sampled dataset are presented in link format, and labels are also associated with links, fold partition during cross-validation is performed in link space as well. We use logistic regression with L_2 regularization for both methods, and the average precision on training and testing can be found as follows.

Based on Table 3, our method outperforms path-constrained random walk in both training set and testing set when both approaches are trained using the same dataset with the same learning method. Compared with path-constrained random walk approach, our method has a larger and more comprehensive hybrid feature space, which contains both meta path-based features as well as numerical features on target paper candidates. By generating features from a uniform meta path-based feature space, our approach is capable of capturing more information from the sampled dataset and improves the average precision in training folds by 4% and increases average precision in testing fold by 8.4% compared with path-constrained random walk.

Table 3: Performance using Classification measures

Methods Precision	Training	Testing
Trained Random Walk	0.7168	0.6691
Our Method	**0.7555**	**0.7533**

6.3 Measure as Query Problem Modeling citation prediction as classification problem in fact simplifies the problem itself. The first step of learning classification models using both methods, is to generate biased samples on link space, which makes positive and negative examples comparable in both training and testing processes. By doing so, we manually reduced search space, and only searched a very limited number of paper candidates to make judgment, hence we can achieve such high precisions for both methods. However, if we model citation prediction as query problem, i.e., given a paper with author(s), target publication venue(s), abstract as well as publication time stamp, one approach should return a list of previous publications ranked by the probability of being cited by the query paper, the citation prediction problem becomes citation query processing, which is a much more difficult problem than classification, simply because now the search space for possible citation paper candidates becomes all the papers in the DBLP information network. In this subsection, we test our approach along with the two link prediction problems by experimenting citation prediction as query processing problem.

Path-constrained random walk and our approach are both supervised, so training datasets need to be generated first before query processing. In order to perform a fair competition with other methods, we use a different strategy to sample training data compared with the training dataset sampling method we used in the previous subsection. Instead of partitioning link space as we did for classification measurement, we randomly partition paper node space in the DBLP information network into five folds, use four folds as training set, and the rest one as testing. The reason

Table 4: Performance as Query Processing on DBLP Network

Methods	Group 1			Group 2			Group 3		
	prec@10	prec@20	recall@50	prec@10	prec@20	recall@50	prec@10	prec@20	recall@50
trained rw	0.2000	0.1250	0.1483	0.0857	0.1000	0.1314	0.2167	0.1750	0.1467
pp	0.1833	0.1333	0.1567	0.1143	0.1071	0.1529	0.2000	0.1667	0.1567
w/o bucket	**0.2333**	0.1333	0.2000	0.1429	0.1143	0.1643	**0.3000**	**0.2000**	0.1717
bucket	**0.2333**	**0.1417**	**0.2533**	**0.1714**	**0.1214**	**0.1771**	0.2833	**0.2000**	**0.1867**

we partition paper nodes instead of citation relations for query processing training is that we want to make sure that, during query processing, all test query nodes are new to all the ranking models, since we are testing both supervised and unsupervised methods. Also, since citation relations are directed, while we are searching for citation relations for an unseen paper query during testing, we should search the entire network instead of only within test set, which means, if one approach can find a paper in training dataset, which is cited by the citation query paper, this counts as a hit.

In order to deal with the large search space, in our approach, we first build discriminative term buckets using training dataset, and then add test papers into the buckets by applying term expansion technique. While generating training dataset, we only focus on positive links and negative links within the same term bucket, since the search space for our approach is within term buckets. While answering queries, very similarly, we only search the term buckets which contain the query paper instead of searching the entire DBLP information network. We use the biased random sampling technique to generate training dataset for the path-constrained random walk approach, and during query answering, the ranking model learned by path-constrained random walk searches the entire network for possible citation relations. Personalized PageRank does not require training, so this approach simply calculate similarity score between the query paper and all other papers using the DBLP network structure, i.e., along paper-venue links, paper-term links as well as paper-author links only, and return papers with the highest similarity as the query results. What's more, to demonstrate the power of discriminative term bucketing, we add another competitor method, which uses the same feature space as our method, but searches the entire paper set in the DBLP information network. To distinguish these two methods, we call our method meta path-based citation prediction framework with discriminative term bucketing (denoted as bucket in tables and figures), and we refer the new competitor method as meta path-based citation prediction framework without term bucketing (denoted as w/o bucket in tables and figures).

We randomly pick 19 query papers from the testing set, and divide them into three groups based on the number of citation relations associated with them. Group 1 contains 6 papers which cite less than 20 previous papers each, group 2 contains 7 papers whose reference size is between 20 to 30, and group 3 has 6 papers, each of which cites more than 30 papers. The query processing performance results can be found in Table 4 and Figures 2(a), 2(b) and 2(c).

We use three query processing measures to evaluate the performance of each method, which are precision at top 10 query results, precision at top 20 query results and recall at top 50 query results, denoted as prec@10, prec@20 and recall@50, respectively. Based on these measurements, one can notice that, our methods can find more citation relations than link prediction methods in general. For example, our methods improve recall@50 by 10% in query group 1 compared with link prediction methods, and also increase prec@10 by 7−8% in query group 3. Discriminative term bucketing technique helps our method reduce search space by around 40% on average, and as we can see in Figures 2(a), 2(b) and 2(c), by eliminating irrelevant citation candidates, meta path-based prediction model with bucketing outperforms the one searches the entire publication network. Another interesting observation is, personalized PageRank gives a relatively better performance than path-constrained random walk method. The reason is, in path-constrained random walk training process, we only use short meta paths (length up to 3), so path-constrained random walk model is only able to reach its neighbors which are three steps away from the queries nodes, while personalized PageRank can reach all possible papers on the network since the calculation does not stop until similarity vector converges. This actually proves our observation in Section 1, which is citation relations does not have high locality as other links, and cited papers can be from anywhere on the DBLP information network.

We also use precision-recall plot to demostrate a more comprehensive comparison of these four different methods in Figure 2(d). From which we can conclude that, meta path-based prediction model with bucketing gives a good performance overall, the precision of which can achieve almost 70% when the recall is low (e.g.,

when we only need top-1 or 2 results). While at the same recall level, link prediction methods can only achieve precision level around 30%. However, the meta path-based precision model without bucketing outperforms our method when the recall is around 5%, which suggests that although our discriminative term bucketing method is very effective in terms of reducing search space, eliminating irrelevant papers and maintaining potential citation relations in buckets, since in this method, we only search citation paper candidates within the same buckets as query paper, we lose the chance of finding citation relations which do not have high document similarity correlation with the query paper (remind that bucketing can only capture 90% citation relations). From precision recall plot, we can also conclude that personalized PageRank is better than path-constrained random walk when searching on the entire paper network because personalized PageRank can reach paper candidates that are far away from the query paper.

6.4 Parameter Turning and Time Complexity

As discussed in Section 3, in meta path-based citation prediction framework, the precision boundary is highly determined by the quality of discriminative term bucket building step. And mutual information threshold MI is the parameter which controls the size of entire term bucket set. The lower this threshold is, the more terms and related papers can be introduced into bucket set, and the more citation relations will be captured in term buckets as well, i.e., most citation relations belong to at least one term bucket. If this threshold is too low, the effect of search space reduction will disappear, which will increase query processing time since now our method needs to search a large link space. On the other hand, if the mutual information threshold MI is too high, the number of potential citation relations and citation information which are supposed to be captured by term bucket techniques will disappear. This can lead to a very low query precision. We here study the relationship between mutual information threshold MI and the number of citation relations which can be captured in term bucket technique. From Figure 2(e), we can see that, with the increase of mutual information threshold, the number of citation relationship we can capture in term buckets decreases exponentially, and search space will be shrinking quickly as well. In our experiment, by parameter tuning, we choose 0.0003 as our mutual information threshold. Using this setting, discriminative term buckets can capture around 90% of citation relations and reduce search space by around 40%, which makes a good balance of precision upper bound and search space reduction.

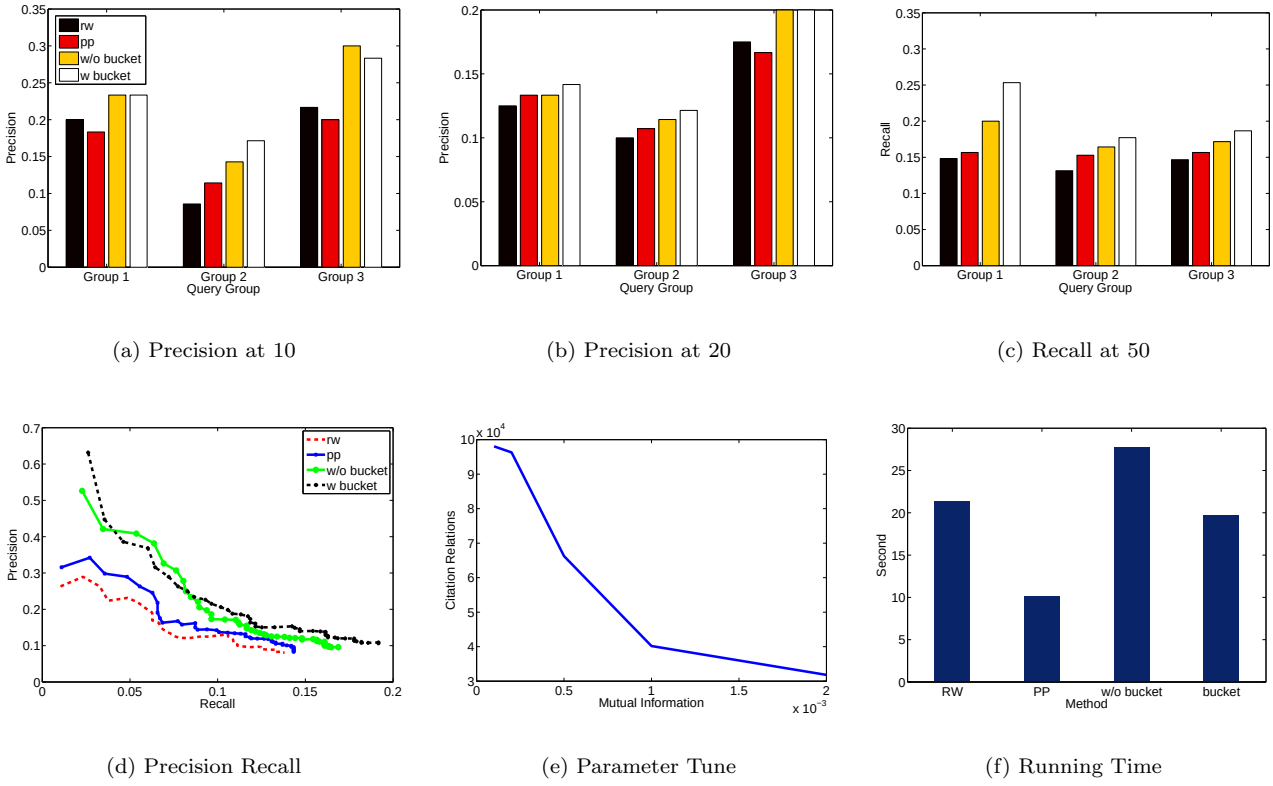
We also studied query processing time using four different methods as well. In order to answer citation queries efficiently, we need to first train ranking models for supervised methods and calculate associated features off-line. On-line query processing time is related to a couple of factors, the first one is the number of features in the prediction model and the second is the size of the search space. We recorded the average query processing time for all four methods, and the result can be found in Figure 2(f). From the plot we can see that, personalized PageRank is the most efficient method, and it takes less than 10 seconds to search through the entire network and generate top citation relations given a citation query, the reason why this method is so fast is because it only uses one feature, which is the personalized PageRank score calculated in the entire network. In both meta path-based prediction methods, we have all together 16 features to process for each citation candidate. However, with discriminative term bucketing technique applying to our ranking model, we do not need to go through the entire DBLP information network anymore, so the query processing time is only around 20 seconds compared with 28 seconds, which is the query processing time of meta path-based prediction model without bucketing, and also this approach is faster than path-constrained random walk as well.

7 Related Work

Citation prediction has been rarely studied in the literature.

The pioneering work about citation prediction is [11]. The authors studied the problem of citation number prediction, which estimates the number of citation for each paper in each time epoch. They used time series analysis technique for citation count prediction. Recently, [17] studied similar problem using machine learning technique. They trained a classifier based on several features of fundamental characteristics for those papers which are highly cited and predicted the popularity degree of each paper. In contrast, the problem we studied in this paper is much more challenging than simple citation number prediction. Rather than simply estimating the count of citations for each paper, we attempted to predict citation relationship. In other words, we aim at telling people which papers would cite which papers, or which papers would be cited by which papers. In fact, citation number prediction can be seen as a by-product of the proposed method in this paper, because as long as we predict the citation relationship, it is very straightforward to count the number of citations in each time epoch.

For search functions in networks, the ranking function defined on networks is the essential component to



(a) Precision at 10 (b) Precision at 20 (c) Recall at 50

(d) Precision Recall (e) Parameter Tune (f) Running Time

Figure 2: Performance: (a) Precision@10; (b) Precision@20; (c) Recall@50; (d) Precision Recall Curve; (e) Parameter Tuning; and (f) Running time.

provide high quality answers. SimRank [7] is a well-known similarity function defined on networks, by which the most similar objects for the given object can be returned. However, due to the fact that it is a global graph measure, the high computational cost usually prevents it from implementing in real search systems. P-PageRank (personalized PageRank) [8, 16, 2] evaluates the probability of a given object to all the objects in the network, and is usually used as a ranking function in network queries. Other extensions, such as ObjectRank [1], try to assign different weights to different relation types in a heterogeneous network, to achieve better results. However, these ranking functions are fixed given a network, and users are not able to intervene to the search by showing their preferences. Most recently, [13] proposed a meta path-based framework in defining the similarity measures between two objects of the same type, which is proved to be more effective than SimRank [7]. However, they have not addressed the learning issue for different query tasks, not to mention the intention understanding issue. In order to build the intention model for queries, we have systematically defined a meta path-based feature space following their work. Each ranking function under each intention then can be built as a linear combination of these different features.

[12] studied coauthor prediction, which aims at predicting the coauthor relationship in the future. However, the problem setting of coauthor prediction is much simple than citation prediction. Coauthor prediction is a short term prediction problem. An author who published paper in 2011 is unlikely to become a coauthor of another author who published paper in 1960. In contrast, citation prediction is a long term prediction. It is reasonable for a paper published in 2011 to cite another paper which is published in 1960. On the other hand, coauthor relationship has strong propagation property. For example, if author A and author B are coauthors, author B and author C are coauthors, then the probability of author A and author C being coauthors is high. So when we do coauthor prediction, given two authors, if there is no shared coauthor for these 2 authors, the probability that they will become coauthor is low. However, it is not true in citation prediction. If paper A cited paper C, but paper B did not cite paper C, the probability of paper B cite paper A is still very uncer-

tain. In a word, citation prediction is more challenging than coauthor prediction.

8 Conclusions and Future Work

In this paper, we propose the problem of citation prediction in the DBLP heterogeneous information network. We proposed a novel two-step approach in order to answer citation prediction queries effective and efficiently. By building discriminative term buckets, our approach first eliminates irrelevant paper candidates, and reduces search space. Then we define a hybrid feature space in order to fully capture citation sensitive structural information, which includes both meta path-based features as well as numerical paper attributes. By learning citation probability model using meta path-based features within the reduced search space, we can define a citation prediction model using both citation probability and the number of common term buckets shared between query paper and citation candidate.

After citation prediction model training process, given a query paper as input, our framework first puts query paper into one or more term buckets, and then generates a set of citation paper candidates by merging papers from related buckets. Citation score is calculated and assigned to each citation candidate, and those candidates with high citation scores will be returned as citation query results.

Empirical study shows that our approach can find citation relations with much higher accuracy compared with traditional link prediction methods. Also, by comparing our method with a similar meta path-based citation prediction approach without bucketing technique, we demonstrate the power of discriminative term bucketing technique, which can reduce search space and improve prediction precision at the same time.

Interesting future work includes, citation prediction study on different information networks, e.g., predicting retweet relations on twitter, exploring new feature selection method [3] [4] instead of mutual information, and also meta path-based feature space index technique in order to further improve query processing efficiency.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. The work was supported in part by NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

References

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: authority-based keyword search in databases. In *VLDB'04*, pages 564–575, 2004.

[2] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *WWW'07*, pages 571–580, 2007.

[3] Q. Gu and J. Han. Towards feature selection in network. In *CIKM*, pages 1175–1184, 2011.

[4] Q. Gu, Z. Li, and J. Han. Generalized fisher score for feature selection. In *UAI*, pages 266–273, 2011.

[5] Q. Gu, J. Zhou, and C. H. Q. Ding. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *SDM*, pages 199–210, 2010.

[6] J. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569, 2005.

[7] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD*, pages 538–543, 2002.

[8] G. Jeh and J. Widom. Scaling personalized web search. In *WWW'03*, pages 271–279, 2003.

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

[10] R. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD*, pages 243–252, 2010.

[11] J. N. Manjunatha, K. R. Sivaramakrishnan, R. K. Pandey, and M. N. Murty. Citation prediction using time series approach kdd cup 2003 (task 1). *SIGKDD Explorations*, 5(2):152–153, 2003.

[12] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks. In *Proceedings of 2011 Int. Conf. on Advances in Social Network Analysis and Mining*. IEEE, 2011.

[13] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[14] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, pages 797–806, 2009.

[15] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.

[16] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM*, pages 613–622, 2006.

[17] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *CIKM*, 2011.

[18] X. Yu, A. Pan, L. A. Tang, Z. Li, and J. Han. Geo-friends recommendation in gps-based cyber-physical social network. In *ASONAM*, pages 361–368. IEEE Computer Society, 2011.