ClusCite: Effective Citation Recommendation by Information Network-Based Clustering

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, Jiawei Han
University of Illinois at Urbana-Champaign, Urbana, IL
{xren7, jliu64, xiaoyu1, khndlwl2, qgu3, lidan, hanj}@illinois.edu

ABSTRACT

Citation recommendation is an interesting but challenging research problem. Most existing studies assume that all papers adopt the same criterion and follow the same behavioral pattern in deciding relevance and authority of a paper. However, in reality, papers have distinct citation behavioral patterns when looking for different references, depending on paper content, authors and target venues. In this study, we investigate the problem in the context of heterogenous bibliographic networks and propose a novel cluster-based citation recommendation framework, called **ClusCite**, which explores the principle that citations tend to be *softly* clustered into *interest groups* based on multiple types of relationships in the network. Therefore, we predict *each* query's citations based on related interest groups, each having its own model for paper authority and relevance. Specifically, we learn group memberships for objects and the significance of relevance features for each interest group, while also propagating relative authority between objects, by solving a joint optimization problem. Experiments on both DBLP and PubMed datasets demonstrate the power of the proposed approach, with 17.68% improvement in Recall@50 and 9.57% growth in MRR over the best performing baseline.

Categories and Subject Descriptors

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

Keywords

Citation Recommendation; Heterogeneous Information Network; Clustering; Citation Behavioral Pattern

1. INTRODUCTION

A research paper needs to cite relevant and important previous work to help readers understand its background, context and innovation. However, the already large, and rapidly growing body of scientific literature makes it hard for anyone to go through and digest all the papers. It is thus desirable to design a system that could automatically generate quality citation recommendations. Traditional literature search engines, such as Google Scholar, can retrieve a list of relevant papers using keyword-based queries. But casting one's rich information needs into a few keywords may not be feasible. Moreover, a user may be looking for papers that are not only relevant to their work, but also important and of high quality. To this end, citation recommendation aims to suggest a small number of publications that can be used as high quality references to satisfy such citation requirements.

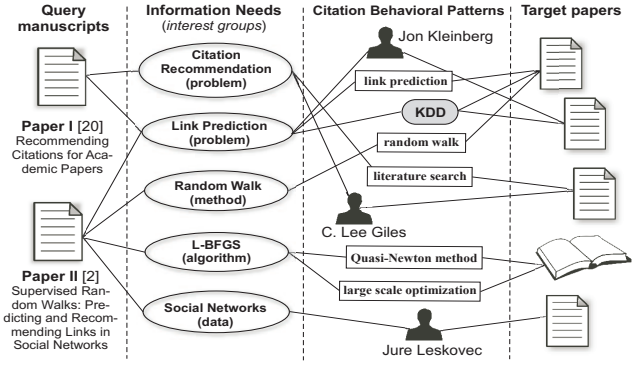


Figure 1: A toy example showing the diverse information needs of two query manuscripts and the corresponding citation behavioral patterns.

There exist some interesting studies on citation recommendation. Context-aware recommendation [10, 12] analyzes each citation's local context to capture its specific information needs. However, local context can be ambiguous or too short a query, causing inaccurate predictions. Topical similarity-based methods [18, 24] find conceptually related papers by taking advantage of latent topic models. But solely relying on topic distributions to measure relevance is insufficient. A large number of papers may share the same topic, making topical similarity weak in indicating importance of a paper. Both methods primarily focus on recommending relevant papers based on content, but ignore critical information related to importance and quality.

Recent studies [16, 20] utilize citation links to derive structural similarity and authority, which serve as good complements to content-based relevance features. With paper text, authors and target venues as queries, one can further generate a rich set of structural features [5, 27] based on multiple types of relations between different entities. However, existing hybrid methods have difficulty in handling the diverse information needs since they impose the same citation behavioral pattern on every query manuscript. Fig. 1 il-

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623630.

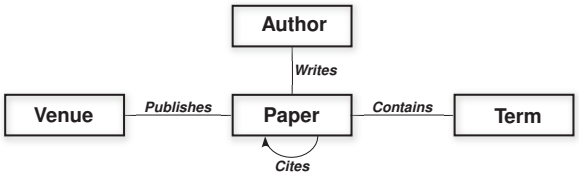


Figure 2: Schema for DBLP bibliographic network.

lustrates the diversity of these behavioral patterns using a toy example. Paper I is on "citation recommendation" and "link prediction", which are studied by a relatively compact group of researchers and venues, and one can find useful papers through related researchers, venues and key terms effectively. On the other hand, for "random walk", relations through authors and venues would be less informative on paper relevance since this method is widely studied by authors working on a variety of topics, and published at venues focusing on a variety of fields. Previous hybrid methods learn and apply the same recommendation model across all queries, ignoring the variations in citation behaviors when seeking quality references. Intuitively, paper citations should be organized into different groups and each group should have its own behavior pattern to identify information of interest.

In this paper, we propose a novel citation recommendation framework to capture citation behaviors for *each* query manuscript, based on both paper relevance and importance. By softly clustering citations into different interest groups, we aim to study the significance of different relevance features for each interest group, and derive paper relative authority within each group. In doing so, the challenge of satisfying diverse information needs behind a paper's citations can be properly tackled by making a paper-specific recommendation according to the query's interest group membership. Meanwhile, integration of paper importance can be accurately accomplished using relative authority. This idea, though interesting, leads to two critical problems: (1) how to discover hidden interest groups for effective citation recommendation, and (2) how to derive behavioral patterns on relevance and authority for each group.

To facilitate our study, a heterogenous bibliographic network, encoding the multiple types of relations between different objects, is constructed (Fig. 2). A rich set of structural features is derived from the network, representing various relation semantics (Table 1) between two papers. We then formulate a joint optimization problem to learn the proposed model such that prediction error along with graph regularization is minimized over known citations, based on the network. Specifically, the optimization problem conducts graph-regularized co-clustering to learn group membership for attribute objects and weights on relevance features for each group. It also propagates relative authority between different objects. An alternative minimization algorithm, called **ClusCite**, is further designed to iterate between co-clustering and authority propagation. Intuitively, feature weights and relative authority can be better learned with high quality interest groups, and in turn they assist in mining higher quality interest groups.

Our experiments on the DBLP and PubMed datasets demonstrate the power of the proposed model. ClusCite achieves 17.68% improvement in Recall@50 and 9.57% growth in MRR over the best baseline on the DBLP dataset. Our performance analysis shows that ClusCite can achieve even better results with richer attribute objects, and our case studies demonstrate the effectiveness of discovered interest groups and object relative authority for citation recommendation.

The rest of the paper is organized as follows. Sec. 2 gives background and the problem definition. Sec. 3 introduces our new model. The learning algorithm and its computational complexity analysis are in Sec. 4. We present and analyze our experimental results in Sec. 5, discuss the related work in Sec. 6, and conclude this study in Sec. 7.

Table 1: Meta paths with different semantics.

Meta path	Semantic meaning of the relation
$P-A-P$	p_i and p_j share same author(s)
$P-T-P$	p_i and p_j contain same term(s)
$P-V-P$	p_i and p_j are in the same venue
$P-T-P \rightarrow P$	p_i share term(s) with the paper(s) that cite p_j
$P-A-P \leftarrow P$	p_i share the same author(s) with the paper(s) cited by p_j

2. BACKGROUND

This section introduces concepts on heterogeneous bibliographic networks and presents the formal problem definition.

A **heterogeneous bibliographic network** [27, 23] is a directed graph G, that consists of multiple types of objects and relationships, derived from a bibliographic dataset.

Suppose there are n papers $\mathcal{P} = \{p_1, \ldots, p_n\}$, $|\mathcal{A}|$ authors $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$, $|\mathcal{V}|$ venues (conferences or journals) $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$, and $|\mathcal{T}|$ terms $\mathcal{T} = \{e_1, \ldots, e_{|\mathcal{T}|}\}$ in the network. Citations between papers form a directed subgraph denoted by an adjacency matrix $\mathbf{Y} \in \mathbb{R}^{n \times n}$ with $Y_{ij} = 1$ if paper p_i cites paper p_j and $Y_{ij} = 0$ otherwise. For relationships between papers and authors, we use an undirected bipartite graph, denoted by a biadjacency matrix $\mathbf{R}^{(\mathcal{A})} \in \mathbb{R}^{n \times |\mathcal{A}|}$, where $R^{(\mathcal{A})}_{ij} = 1$ if paper p_i has the author a_j and $R^{(\mathcal{A})}_{ij} = 0$ otherwise. Similarly, the relationships between papers and venues can also be represented by a biadjacency matrix $\mathbf{R}^{(\mathcal{V})} \in \mathbb{R}^{n \times |\mathcal{V}|}$, $R^{(\mathcal{V})}_{ij} = 1$ if paper p_i is published in the venue v_j and $R^{(\mathcal{V})}_{ij} = 0$ otherwise. We extract a set of term objects $\mathcal{T}$ from the paper's free-text and further construct an undirected bipartite subgraph between these terms and papers to represent paper content. We use the weight matrix $\mathbf{R}^{(\mathcal{T})} \in \mathbb{R}^{n \times |\mathcal{T}|}$ to denote the paper-term subgraph where $R^{(\mathcal{T})}_{ij}$ is the term frequency of term e_j in paper p_i.

We adopt the concept of **network schema** to describe the heterogeneous bibliographic network at the meta level [22, 23]. An example is shown in Fig. 2.

As shown in [22, 27], meta path-based features in heterogeneous information networks describe a rich set of relation semantics that can capture textual similarity, conceptual relevance and several kinds of social relatedness. A *meta path* is defined over network schema, where nodes are object types and edges are relation types. Table 1 shows some examples that use meta paths to measure paper relevance for the citation recommendation problem. Moreover, structural similarity measures can be defined on each meta path to generate relevance features, as shown in [22, 27].

In general, we represent the meta path-based relevance score between p_i and p_j as $\phi(p_i, p_j)$. Suppose we generate L different meta path-based relevance features by combining different meta paths with different structural similarity measures, we can define a relevance scores matrix $\mathbf{S}^{(i)} \in \mathbb{R}^{n \times L}$ for every paper $p_i \in \mathcal{P}$ where $S^{(i)}_{jl} = \phi^{(l)}(p_i, p_j)$ is the l-th meta path-based relevance score between p_i and p_j[1].

In this work, we cast the citation recommendation problem into the problem of learning a recommendation score function $s(q, p) : \mathcal{Q} \times \mathcal{P} \mapsto \mathbb{R}$ for a query manuscript $q \in \mathcal{Q}$ and a target paper $p \in \mathcal{P}$ based on the heterogeneous bibliographic network. The learned function is later used to

[1]Details of the meta path-based similarity computation in heterogeneous bibliographic network can be found in [22].

compute scores between query and target papers to make a recommendation. Formally, we define the **citation recommendation problem** as follows.

DEFINITION 1 (PROBLEM DEFINITION). *Given a heterogeneous bibliographic network G, and the terms, authors and target venues for a query manuscript $q \in \mathcal{Q}$, we aim to build a recommendation model specifically for q, and recommend a small subset of target papers $p \in \mathcal{P}$ as high quality references for q, by ranking the papers with the score function $s(q, p)$.*

3. THE PROPOSED FRAMEWORK

At a high level, the proposed cluster-based citation recommendation framework consists of two major steps:

1. Learning the model parameters based on known citations by solving a joint optimization problem (Sec. 4).

2. Making paper-specific recommendations for each query manuscript based on the learned ClusCite model, which is introduced in detail in this section.

3.1 Model Overview

We first provide an overview of the proposed model by defining the major components in the score function $s(q, p)$.

Given a query manuscript q, its citations will focus on several interest groups each having its own behavioral patterns in finding relevant and high authority work (Fig. 1). It is desirable to recommend papers that are highly ranked in multiple interest groups of the query, since they best capture diverse information needs. We propose a cluster-based score function to decide relative relevance and importance of target papers in the context of each interest group. It assigns a final recommendation score by integrating scores computed with respect to different interest groups.

Mathematically, suppose paper citations can be softly clustered into K interest groups, based on multiple types of relationships between objects in the heterogeneous bibliographic network, then we define the score function $s(q, p)$ as follows:

$$s(q,p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \Big\{ r^{(k)}(q,p) + f_{\mathcal{P}}^{(k)}(p) \Big\}. \tag{1}$$

Function $s(q, p)$ measures how likely a query manuscript $q \in \mathcal{Q}$ is to cite a target paper $p \in \mathcal{P}$. It is decomposed into a set of cluster-based functions: the cluster-based relevance function $r^{(k)}(q, p) : \mathcal{Q} \times \mathcal{P} \mapsto \mathbb{R}$ measures the relatedness between q and p according to the k-th interest group, and paper relative authority function $f_{\mathcal{P}}^{(k)}(p) : \mathcal{P} \mapsto \mathbb{R}$ computes the relative importance of p within the k-th interest group. The weighted combination of these functions defines the final recommendation score with respect to the group membership indicators of q, *i.e.*, $\{\theta_q^{(k)} : \theta_q^{(k)} > 0\}$, which represent how likely query q is to belong to the K different interest groups.

3.2 Feature Weights for Paper Relevance

As mentioned in Sec. 2, one can compute a rich set of meta path-based features to describe paper relevance under various relation semantics. Each meta path-based feature, could play a distinct role in identifying relevant work in different interest groups.

In Fig. 1, incorporating meta path $P - V - P$ along with textual similarity $P{-}T{-}P$ can effectively suggest related papers under the interest "link prediction (problem)" because only a compact set of venues (*e.g.*, KDD, ICML and ICDM) study this problem. However, if the interest switches to "L-BFGS (algorithm)", using $P - V - P$ probably will hurt the results since a much broarder set of venues involve studying this algorithm, and thus, sharing a venue with the query provides very weak evidence for paper relevance.

Table 2: Learned weights on seven different meta paths for four mined interest groups ($K = 40$).

Meta path	Group 1	Group 2	Group 3	Group 4
$P - V - P$	0.0024	0.0113	0.0158	0.3076*
$P - A - P$	0.0054	0.0006	0.0192	0.1243
$P - A - P \rightarrow P$	0.6133**	0.2159*	0.2254	0.0213
$P - T - P$	0.1227	0.0947	0.1579	0.1095
$P - T - P \rightarrow P$	0.0442	0.5448**	0.3250*	0.0231
$P - T - P \leftarrow P$	0.1938*	0.0870	0.3578**	0.2409**

In order to capture the biased significance of different relevance features for different interest groups, we assign feature weights for each interest group individually, leading to the definition of a cluster-based relevance function as follows:

$$r^{(k)}(q,p) = \sum_{l=1}^{L} w_k^{(l)} \cdot \phi^{(l)}(q,p). \tag{2}$$

For each interest group k, we use a set of weights $\{w_k^{(l)} : w_k^{(l)} > 0\}$ to measure the significance of the L different meta path-based features $\{\phi^{(l)}(q, p)\}$ for the group.

These K feature weights are estimated through a joint optimization problem (Sec. 4). We demonstrate in Table 2 the learned feature patterns over 7 meta paths for 4 example interest groups (* and ** highlight first and second most significant values), using the random walk-based similarity measure on DBLP. All 4 groups show distinct weights on the 7 meta paths, justifying the claim that different interest groups hold different feature weights. In particular, we find meta paths which impose textual similarity (*e.g.*, $P{-}T{-}P \leftarrow P$) as well as references of co-author's papers ($P{-}A{-}P \rightarrow P$) play critical roles in finding relevant papers in these 4 groups, which matches human intuitions very well.

3.3 Object Relative Authority

A paper may have very different visibility or authority among different interest groups even if it has many citations. In the DBLP dataset [25], paper ObjectRank [3] (132 citations) got 47 citations from VLDB but only 12 from WWW, while RankSVM [14] (250 citations) obtained only 27 citations from VLDB but 109 from WWW, implying the bias of authority in different interest groups.

Instead of learning object's group membership and deriving relative authority separately, we propose to estimate them jointly using graph regularization, which preserves consistency over each subgraph. By doing so, paper relative authority serves as a feature for learning interest groups, and better estimated groups can in turn help derive relative authority more accurately (Fig. 3).

We adopt the semi-supervised learning framework [8] that leads to iteratively updating rules as authority propagation between different types of objects.

$$\begin{aligned} \mathbf{F}_{\mathcal{P}} &= G_{\mathcal{P}}\big(\mathbf{F}_{\mathcal{P}},\ \mathbf{F}_{\mathcal{A}},\ \mathbf{F}_{\mathcal{V}}; \lambda_{\mathcal{A}},\ \lambda_{\mathcal{V}}\big), \\ \mathbf{F}_{\mathcal{A}} &= G_{\mathcal{A}}\big(\mathbf{F}_{\mathcal{P}}\big) \ \text{ and } \ \mathbf{F}_{\mathcal{V}} = G_{\mathcal{V}}\big(\mathbf{F}_{\mathcal{P}}\big). \end{aligned} \tag{3}$$

We denote relative authority score matrices for paper, author and venue objects by $\mathbf{F}_{\mathcal{P}} \in \mathbb{R}^{K \times n}$, $\mathbf{F}_{\mathcal{A}} \in R^{K \times |\mathcal{A}|}$ and $\mathbf{F}_{\mathcal{V}} \in R^{K \times |\mathcal{V}|}$. Generally, in an interest group, relative importance of one type of object could be a combination of the relative importance from different types of objects [23].

In our solution, the propagation function $G_{\mathcal{P}}$ updates paper relative authority scores for all groups, following the intuition: High quality papers from an interest group are often published in highly reputed venues, written by authoritative authors and related to other high quality papers, from this group. Trade-off parameters $\lambda_{\mathcal{A}}$ and $\lambda_{\mathcal{V}}$ control the relative importance of paper-author and paper-venue relations. On the other hand, propagation functions $G_{\mathcal{A}}$ and $G_{\mathcal{V}}$ capture the rules: highly regarded authors often write good quality papers, and highly reputed venues often publish good quality papers. We include detailed formulae for the three propagation functions in Sec. 4.

3.4 Paper-Specif c Citation Recommendation

In practice, to derive interest group memberships for newly emerged queries, one has to re-estimate the model using these queries and training data, which is highly inefficient. Moreover, as the number of papers grows rapidly, the size of the model parameter space will increase a lot, making the model learning even more unscalable.

To tackle these two challenges, we leverage group memberships of the query's related attribute objects, *i.e.*, authors, terms and target venues, to approximately represent group membership of the query manuscript.

Intuitively, terms of the query manuscript describe its information needs based on paper content, whereas its author(s) and venue complement the content with research interests and other conceptual information. Specifically, we represent the query's group membership $\theta_q^{(k)}$ by weighted integration of group memberships of its attribute objects.

$$\theta_q^{(k)} = \sum_{\mathcal{X}\in\{\mathcal{A},\mathcal{V},\mathcal{T}\}} \sum_{x\in N_{\mathcal{X}}(q)} \frac{\theta_x^{(k)}}{|N_{\mathcal{X}}(q)|}. \tag{4}$$

We use $N_{\mathcal{X}}(q)$ to denote type $\mathcal{X}$ neighbors for query q, *i.e.*, its attribute objects. How likely a type $\mathcal{X}$ object is to belong to the k-th interest group is represented by $\theta_x^{(k)}$.

Paper-specific citation recommendation can be efficiently conducted for each query manuscript q by applying Eq. (1) along with definitions in Eqs. (2), (3) and (4).

4. MODEL LEARNING

This section introduces the learning algorithm for the proposed citation recommendation model in Eq. (1).

There are three sets of parameters in our model: group memberships for attribute objects; feature weights for interest groups; and object relative authority within each interest group. A straightforward way is to first conduct hard-clustering of attribute objects based on the network and then derive feature weights and relative authority for each cluster. Such a solution encounters several problems: (1) one object may have multiple citation interests, (2) mined object clusters may not properly capture distinct citation interests as we want, and (3) model performance may not be best optimized by the mined clusters.

In our solution, we formulate a joint optimization problem to estimate all model parameters simultaneously, which minimizes prediction error as well as graph regularization. By doing so, we can softly cluster attribute objects in terms of their citation interests and guarantee the learned model can yield good performance on training data.

We explain the joint optimization problem in Sec. 4.1 and design an efficient algorithm to solve it in Sec. 4.2 along with its computational complexity analysis in Sec. 4.3.

4.1 The Joint Optimization Problem

To learn model parameters, we use a citation network as training data, where value 1 indicates observed citation relationships while value 0 represents a mixture of negatives (should not cite) or unobserved (unaware and may cite in the future) examples. Traditional learning methods adopt classification [27] or learning-to-rank [1] objective functions and usually treat all 0s in training data as negative examples, which does not fit the real cases.

Without loss of generality, we adopts weighted square error [11] on the citation matrix as the loss function to measure the prediction performance, which is defined as follows:

$$\begin{aligned}\mathcal{L} &= \sum_{i,j=1}^{n} M_{ij}\Big(Y_{ij} - \sum_{k=1}^{K}\sum_{l=1}^{L}\theta_{p_i}^{(k)} w_k^{(l)} S_{jl}^{(i)} - \sum_{k=1}^{K}\theta_{p_i}^{(k)} F_{\mathcal{P},kj}\Big)^2, \\ &= \sum_{i=1}^{n}\big\|\mathbf{M}_i \odot \Big(\mathbf{Y}_i - \mathbf{R}_i\mathbf{P}(\mathbf{W}\mathbf{S}^{(i)T} + \mathbf{F}_{\mathcal{P}})\Big)\big\|_2^2.\end{aligned} \tag{5}$$

We define the weight indicator matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ for the citation matrix, where M_{ij} takes value 1 if the citation relationship between p_i and p_j is observed and 0 in other cases. By doing so, the model can focus on positive examples and get rid of noise in the 0 values. One can also define other loss functions to optimize with respect to precision or recall.

For ease of optimization, the loss can be further rewritten in a matrix form, where matrix $\mathbf{P} \in \mathbb{R}^{(|\mathcal{T}|+|\mathcal{A}|+|\mathcal{V}|)\times K}$ is group membership indicator for all attribute objects while $\mathbf{R}_i \in \mathbb{R}^{n\times((|\mathcal{T}|+|\mathcal{A}|+|\mathcal{V}|))}$ is the corresponding neighbor indicator matrix such that $\mathbf{R}_i\mathbf{P} = \sum_{\mathcal{X}\in\{\mathcal{A},\mathcal{V},\mathcal{T}\}}\sum_{x\in N_{\mathcal{X}}(p_i)}\frac{\theta_x^{(k)}}{|N_{\mathcal{X}}(p_i)|}$. Feature weights for each interest group are represented by each row of the matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$, *i.e.*, $W_{kl} = w_k^{(l)}$. Hadamard product $\odot$ is used for the matrix element-wise product.

As discussed in Sec. 3.3, to achieve authority learning jointly, we adopt graph regularization to preserve consistency over the paper-author and paper-venue subgraphs, which takes the following form:

$$\begin{aligned}\mathcal{R} = &\frac{\lambda_{\mathcal{A}}}{2}\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{A}|} R_{ij}^{(\mathcal{A})}\Big\|\frac{\mathbf{F}_{\mathcal{P},i}}{D_{ii}^{(\mathcal{P}\mathcal{A})}} - \frac{\mathbf{F}_{\mathcal{A},j}}{D_{jj}^{(\mathcal{A}\mathcal{P})}}\Big\|_2^2 \\ &+ \frac{\lambda_{\mathcal{V}}}{2}\sum_{i=1}^{n}\sum_{j=1}^{|\mathcal{V}|} R_{ij}^{(\mathcal{V})}\Big\|\frac{\mathbf{F}_{\mathcal{P},i}}{D_{ii}^{(\mathcal{P}\mathcal{V})}} - \frac{\mathbf{F}_{\mathcal{V},j}}{D_{jj}^{(\mathcal{V}\mathcal{P})}}\Big\|_2^2.\end{aligned} \tag{6}$$

The intuition behind the above two terms is natural: Linked objects in the heterogeneous network are more likely to share similar relative authority scores [13]. To reduce impact of node popularity, we apply a normalization technique on authority vectors, which helps suppress popular objects to keep them from dominating the authority propagation. Each element in the diagonal matrix $\mathbf{D}^{(\mathcal{P}\mathcal{A})} \in \mathbb{R}^{n\times n}$ is the degree of paper p_i in subgraph $R^{(\mathcal{A})}$ while each element in $\mathbf{D}^{(\mathcal{A}\mathcal{P})} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ is the degree of author a_j in subgraph $R^{(\mathcal{A})}$. Similarly, we can define the two diagonal matrices for subgraph $R^{(\mathcal{V})}$.

Integrating the loss in Eq. (5) with graph regularization in Eq. (6), we formulate a joint optimization problem following the semi-supervised learning framework [8]:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{W},\mathbf{F}_{\mathcal{P}},\mathbf{F}_{\mathcal{A}},\mathbf{F}_{\mathcal{V}}} &\frac{1}{2}\mathcal{L} + \mathcal{R} + \frac{c_p}{2}\|\mathbf{P}\|_F^2 + \frac{c_w}{2}\|\mathbf{W}\|_F^2 \\ &\text{s.t.}\quad \mathbf{P} \geq 0;\quad \mathbf{W} \geq 0.\end{aligned} \tag{7}$$

To ensure stability of the obtained solution, Tikhonov regularizers are imposed on variables $\mathbf{P}$ and $\mathbf{W}$ [4], and we use $c_p,\ c_w > 0$ to control the strength of regularization. In addition, we impose non-negativity constraints to make sure

learned group membership indicators and feature weights can provide semantic meaning as we want.

4.2 The ClusCite Algorithm

Directly solving Eq (7) is not easy because the objective function is non-convex. We develop an alternative minimization algorithm, called **ClusCite**, which alternatively optimizes the problem with respect to each variable.

The learning algorithm essentially accomplishes two things simultaneously and iteratively: Co-clustering of attribute objects and relevance features with respect to interest groups, and authority propagation between different objects. During an iteration, different learning components will mutually enhance each other (Fig. 3): Feature weights and relative authority can be more accurately derived with high quality interest groups while in turn they serve a good feature for learning high quality interest groups.

First, to learn group membership for attribute objects, we take the derivative of the objective function in Eq. (7) with respect to $\mathbf{P}$ while fixing other variables, and apply the Karush-Kuhn-Tucker complementary condition to impose the non-negativity constraint [7]. With some simple algebraic operations, a multiplicative update formula for $\mathbf{P}$ can be derived as follows:

$$P_{jk} \leftarrow P_{jk} \frac{\Big[\sum_{i=1}^{n} \mathbf{R}_i^T \tilde{\mathbf{Y}}_i \mathbf{S}^{(i)} \mathbf{W}^T + \mathbf{L}_{\mathbf{P}1}^{+} + \mathbf{L}_{\mathbf{P}2}^{-}\Big]_{jk}}{\Big[\mathbf{L}_{\mathbf{P}0} + \mathbf{L}_{\mathbf{P}1}^{-} + \mathbf{L}_{\mathbf{P}2}^{+} + c_p \mathbf{P}\Big]_{jk}}, \quad (8)$$

where matrices $\mathbf{L}_{\mathbf{P}0}$, $\mathbf{L}_{\mathbf{P}1}$ and $\mathbf{L}_{\mathbf{P}2}$ are defined as follows:

$$\begin{aligned}
\mathbf{L}_{\mathbf{P}0} &= \sum_{i=1}^{n} \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \mathbf{W} \tilde{\mathbf{S}}^{(i)T} \tilde{\mathbf{S}}^{(i)} \mathbf{W}^T;\ \mathbf{L}_{\mathbf{P}1} = \sum_{i=1}^{n} \mathbf{R}_i^T \tilde{\mathbf{Y}}_i \mathbf{F}_{\mathcal{P}}^T; \\
\mathbf{L}_{\mathbf{P}2} &= \sum_{i=1}^{n} \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \tilde{\mathbf{F}}_{\mathcal{P}}^{(i)} \tilde{\mathbf{F}}_{\mathcal{P}}^{(i)T} + \sum_{i=1}^{n} \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \mathbf{W} \tilde{\mathbf{S}}^{(i)T} \mathbf{F}_{\mathcal{P}}^T \\
&\quad + \sum_{i=1}^{n} \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \mathbf{F}_{\mathcal{P}} \tilde{\mathbf{S}}^{(i)} \mathbf{W}^T.
\end{aligned}$$

In order to preserve non-negativity throughout the update, $\mathbf{L}_{\mathbf{P}1}$ is decomposed into $\mathbf{L}_{\mathbf{P}1}^{-}$ and $\mathbf{L}_{\mathbf{P}1}^{+}$ where $A_{ij}^{+} = (|A_{ij}| + A_{ij})/2$ and $A_{ij}^{-} = (|A_{ij}| - A_{ij})/2$. Similarly, we decompose $\mathbf{L}_{\mathbf{P}2}$ into $\mathbf{L}_{\mathbf{P}2}^{-}$ and $\mathbf{L}_{\mathbf{P}2}^{+}$, but note that the decomposition is applied to each of the three components of $\mathbf{L}_{\mathbf{P}2}$, respectively. We denote the masked $\mathbf{Y}_i$ as $\tilde{\mathbf{Y}}_i$, which is the Hadamard product of $\mathbf{M}_i$ and $\mathbf{Y}_i$. Similarly, $\tilde{\mathbf{S}}^{(i)}$ and $\tilde{\mathbf{F}}_{\mathcal{P}}^{(i)}$ denote row-wise masked $\mathbf{S}^{(i)}$ and $\mathbf{F}_{\mathcal{P}}$ by $\mathbf{M}_i$.

Second, to learn feature weights for interest groups, the multiplicative update formula for $\mathbf{W}$ can be derived following a similar derivation as that of $\mathbf{P}$, taking the form:

$$W_{kl} \leftarrow W_{kl} \frac{\Big[\sum_{i=1}^{n} \mathbf{P}^T \mathbf{R}_i^T \tilde{\mathbf{Y}}_i \mathbf{S}^{(i)} + \mathbf{L}_{\mathbf{W}}^{-}\Big]_{kl}}{\Big[\sum_{i=1}^{n} \mathbf{P}^T \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \mathbf{W} \tilde{\mathbf{S}}^{(i)T} \tilde{\mathbf{S}}^{(i)} + \mathbf{L}_{\mathbf{W}}^{+} + c_w \mathbf{W}\Big]_{kl}}, \quad (9)$$

where we have $\mathbf{L}_{\mathbf{W}} = \sum_{i=1}^{n} \mathbf{P}^T \mathbf{R}_i^T \mathbf{R}_i \mathbf{P} \mathbf{F}_{\mathcal{P}} \tilde{\mathbf{S}}^{(i)}$.

Similarly, to preserve non-negativity of $\mathbf{W}$, $\mathbf{L}_{\mathbf{W}}$ is decomposed into $\mathbf{L}_{\mathbf{W}}^{+}$ and $\mathbf{L}_{\mathbf{W}}^{-}$, which can be computed same before.

Finally, we derive the authority propagation functions in Eq. (3) by optimizing the objective function in Eq. (7) with respect to the authority score matrices of papers, authors and venues. Specifically, we take the derivative of the objective function with respect to $\mathbf{F}_{\mathcal{P}}$, $\mathbf{F}_{\mathcal{A}}$ and $\mathbf{F}_{\mathcal{V}}$, and follow traditional semi-supervised learning frameworks [8] to derive the update rules, which take the form:

$$\begin{aligned}
\mathbf{F}_{\mathcal{P}} &= G_{\mathcal{P}}(\mathbf{F}_{\mathcal{P}}, \mathbf{F}_{\mathcal{A}}, \mathbf{F}_{\mathcal{V}}; \lambda_{\mathcal{A}}, \lambda_{\mathcal{V}}) \\
&= \frac{1}{\lambda_{\mathcal{A}} + \lambda_{\mathcal{V}}} \Big(\lambda_{\mathcal{A}} \mathbf{F}_{\mathcal{A}} \mathbf{S}_{\mathcal{A}}^T + \lambda_{\mathcal{V}} \mathbf{F}_{\mathcal{V}} \mathbf{S}_{\mathcal{V}}^T + \mathbf{L}_{\mathbf{F}_{\mathcal{P}}} \Big) \quad (10) \\
\mathbf{F}_{\mathcal{A}} &= G_{\mathcal{A}}(\mathbf{F}_{\mathcal{P}}) = \mathbf{F}_{\mathcal{P}} \mathbf{S}_{\mathcal{A}}; \quad (11) \\
\mathbf{F}_{\mathcal{V}} &= G_{\mathcal{V}}(\mathbf{F}_{\mathcal{P}}) = \mathbf{F}_{\mathcal{P}} \mathbf{S}_{\mathcal{V}}. \quad (12)
\end{aligned}$$

where we have normalized adjacency matrices and the paper authority guidance terms defined as follows:

$$\begin{aligned}
\mathbf{S}_{\mathcal{A}} &= (\mathbf{D}^{(\mathcal{PA})})^{-1/2} \mathbf{R}^{(\mathcal{A})} (\mathbf{D}^{(\mathcal{AP})})^{-1/2} \\
\mathbf{S}_{\mathcal{V}} &= (\mathbf{D}^{(\mathcal{PV})})^{-1/2} \mathbf{R}^{(\mathcal{V})} (\mathbf{D}^{(\mathcal{VP})})^{-1/2} \\
\mathbf{L}_{\mathbf{F}_{\mathcal{P}}} &= \sum_{i=1}^{n} \mathbf{P}^T \mathbf{R}_i^T \Big\{ \tilde{\mathbf{Y}}_i - \mathbf{R}_i \mathbf{P} \big(\mathbf{W} \tilde{\mathbf{S}}^{(i)} + \tilde{\mathbf{F}}_{\mathcal{P}}^{(i)} \big) \Big\}
\end{aligned}$$

Using normalized adjacency matrices $\mathbf{S}_{\mathcal{A}}$ and $\mathbf{S}_{\mathcal{V}}$ to propagate relative authority can suppress popular objects in the network. In this way, they will not dominate the authority propagation. At each iteration, the guidance term $\mathbf{L}_{\mathbf{F}_{\mathcal{P}}}$ adjusts paper relative authority such that the model can fit known citations in a more accurate way.

Algorithm 1 summarizes the ClusCite algorithm. For convergence analysis, ClusCite essentially applies block coordinate descent on the optimization problem in Eq. (7). The proof procedure in [26] can be adopted to prove convergence for ClusCite (to the local minimum). For lack of space, we do not include it here.

Figure 3: Correlation between paper relative authority and # ground truth citations, during different iterations.

Fig. 3 illustrates the quality change of estimated paper relative authority. Given an interest group, citations from test set to training papers serve as our ground truth for the relative authority of this group. We study the change of correlation between estimated relative authority and the ground truth, during different iterations. The initialization (global citation count) shows poor quality based on the correlation. As the algorithm iterates, we observe significant enhancement on the correlation, which justifies the effectiveness of the proposed authority propagation approach.

4.3 Computational Complexity Analysis

In this section, we analyze the computational complexity of the proposed ClusCite algorithm. Let d denote the total number of attribute objects and $|E|$ the total number of links in the heterogeneous network. First, it takes $O(K(n+d))$ time to initialize all the variables and $O(|E|L)$ time to pre-compute the constants in the update formula. In addition, we apply the fact that: $\sum_n \mathbf{A}_n \mathbf{X} \mathbf{B}_n = \mathbf{C}$ is equivalent to $(\sum_n \mathbf{B}_n^T \bigotimes \mathbf{A}_n)\,\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{C})$, in our implementation so that we can avoid summations over all papers by pre-computing several matrix Kronecker products ($\bigotimes$). This step takes totally $O(L^2|E|^2/n + L|E|^3/n^2)$ time.

We then study the time complexity at each iteration of ClusCite with pre-computed matrices. Learning the group

Algorithm 1 Model Learning by ClusCite

Input: adjacency matrices $\{\mathbf{Y}, \mathbf{S}_{\mathcal{A}}, \mathbf{S}_{\mathcal{V}}\}$, neighbor indicator $\mathbf{R}$, mask matrix $\mathbf{M}$, meta path-based features $\{\mathbf{S}^{(i)}\}$, parameters $\{\lambda_{\mathcal{A}}, \lambda_{\mathcal{V}}, c_w, c_p\}$, number of interest groups K
Output: group membership $\mathbf{P}$, feature weights $\mathbf{W}$, relative authority $\{\mathbf{F}_{\mathcal{P}}, \mathbf{F}_{\mathcal{A}}, \mathbf{F}_{\mathcal{V}}\}$

1: Initialize $\mathbf{P}$, $\mathbf{W}$ with positive values, and $\{\mathbf{F}_{\mathcal{P}}, \mathbf{F}_{\mathcal{A}}, \mathbf{F}_{\mathcal{V}}\}$ with citation counts from training set
2: **repeat**
3: Update group membership $\mathbf{P}$ by Eq. (8)
4: Update feature weights $\mathbf{W}$ by Eq. (9)
5: Compute paper relative authority $\mathbf{F}_{\mathcal{P}}$ by Eq. (10)
6: Compute author relative authority $\mathbf{F}_{\mathcal{A}}$ by Eq. (11)
7: Compute venue relative authority $\mathbf{F}_{\mathcal{V}}$ by Eq. (12)
8: **until** objective in Eq. (7) converges

membership matrix $\mathbf{P}$ by Eq. (8) takes $O(L|E|^3/n^3 + L^2|E|^2/n^2 + Kdn)$ time. Learning the feature weights $\mathbf{W}$ by Eq. (9) takes $O(L|E|^3/n^3 + L^2|E|^2/n^2 + Kdn)$ time. Updating all three relative authority matrices takes $O(L|E|^3/n^3 + |E| + Kdn)$ time. Let the number of iterations to compute ClusCite be T ($T \ll n$). The total time complexity is $O(L|E|^3/n^2 + L^2|E|^2/n + T|E| + TKdn)$. In our experiments, ClusCite usually converges within 50 iterations.

5. EXPERIMENTS

In this section, we evaluate the recommendation performance of the proposed method on real world data and conduct case studies to demonstrate its effectiveness.

5.1 Data Preparation

In the experiments, two different bibliographic datasets are used: the DBLP dataset[2] [25] and the PubMed dataset[3]. Statistics of the two constructed heterogeneous bibliographic networks are summarized in Table 3.

5.1.1 Heterogeneous Bibliographic Networks

Tang *et al.* [25] extracted citation information and built a DBLP citation dataset. We generated a subset of the aforementioned dataset by filtering out papers with incomplete meta information or less than 5 citations. Keywords and key phrases are extracted from paper titles and abstracts using the TF-IDF measure and the TextBlob noun phrase extractor[4]. The PubMed Central dataset is processed by the same method as described above to generate a subset[5]. We converted both datasets into heterogeneous bibliographic networks according to the network schema in Fig. 2.

5.1.2 Training and Evaluation Sets

We split the network to generate training, validation and testing subsets according to the paper publication year. We considered three time intervals T_0, T_1 and T_2. The subnetwork associated with papers in T_0 was used for model training. Papers in T_1 were then used as the validation set for parameter tuning and papers in T_2 were used as the test set for evaluations. Tables 4(a) and 4(b) summarize the statistics of the subsets. During evaluation, we consider citations from the evaluation sets (T_1 and T_2) to the training set (T_0) as the ground truth. Such an evaluation practice is more realistic because a citation recommendation system only knows the related attribute objects of a newly written manuscript. Also, it predicts future citations based on models which are learned from past citations.

[2] http://arnetminer.org/DBLP_Citation
[3] http://www.ncbi.nlm.nih.gov/pmc/
[4] http://textblob.readthedocs.org/en/latest/
[5] https://github.com/shanzhenren2/PubMed_subset

Table 3: Statistics of two bibliographic networks.

Data sets	DBLP	PubMed
# papers	137,298	100,215
# authors	135,612	212,312
# venues	2,639	2,319
# terms	29,814	37,618
# relationships	∼2.3M	∼3.6M
Paper avg citations	5.16	17.55

Table 4: Training, validation and testing paper subsets from the DBLP and PubMed datasets

(a) The DBLP dataset

Subsets	Train	Validation	Test
Years	T_0=[1996, 2007]	T_1=[2008]	T_2=[2009, 2011]
# papers	62.23%	12.56%	25.21%

(b) The PubMed dataset

Subsets	Train	Validation	Test
Years	T_0=[1966, 2008]	T_1=[2009]	T_2=[2010, 2013]
# papers	64.50%	7.81%	27.69%

5.1.3 Feature Generation

In the experiments, without loss of generality, we selected 15 different meta-paths between paper objects including $(P-X-P)^y$, $P-X-P \rightarrow P$ and $P-X-P \leftarrow P$ where $X = \{A, V, T\}$ and $y = \{1, 2, 3\}$. Note that $(P-X-P)^2$ denotes $P-X-P-X-P$. We used two different structural similarity measures: PathSim [22] measure and the random-walk based measure [27]. We applied the random-walk based measure to all meta-paths and the PathSim measure to only symmetric meta-paths due to its requirement. This provides us with 24 meta-path based relevance features. Note that all the "cited" and "citing" relations in the meta-paths were only measured between papers in the training set.

5.2 Experimental Settings

We provide details on the experimental settings for conducting evaluations on all the methods.

5.2.1 Compared Methods

We compared the proposed method (ClusCite) with its variation which considered only relevance features (ClusCite-Rel). Several widely deployed or state-of-the-art citation recommendation approaches were also implemented, including content-based methods, link-based methods and hybrid methods. All compared methods were first tuned on validation set to pick the tuning parameters.

BM25: BM25 is a text-based method, which computes similarity scores using only text information.

PopRank [19]: PopRank is a link-based method which derives an object's importance based on authority propagation in the heterogeneous bibliographic network.

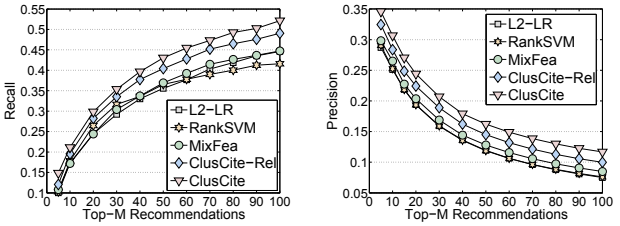
TopicSim: We measure similarity between papers with topic modeling technique (LDA) and return the papers with the most similar topic distribution compared with the query.

Link-PLSA-LDA [18]: Link-PLSA-LDA[6] is a hybrid method that leverages both document text and citation links

[6] https://sites.google.com/site/rameshnallapati/software

Table 5: Recommendation performance comparisons on DBLP and PubMed datasets in terms of Precision, Recall and MRR. We set the number of interest groups to be 200 ($K = 200$) for ClusCite and ClusCite-Rel.

Method	DBLP					PubMed				
	P@10	P@20	R@20	R@50	MRR	P@10	P@20	R@20	R@50	MRR
BM25	0.1260	0.0902	0.1431	0.2146	0.4107	0.1847	0.1349	0.1754	0.2470	0.4971
PopRank	0.0112	0.0098	0.0155	0.0308	0.0451	0.0438	0.0314	0.0402	0.0814	0.2012
TopicSim	0.0328	0.0273	0.0432	0.0825	0.1161	0.0761	0.0685	0.0855	0.1516	0.3254
Link-PLSA-LDA	0.1023	0.0893	0.1295	0.1823	0.3748	0.1439	0.1002	0.1589	0.2015	0.4079
L2-LR	0.2274	0.1677	0.2471	0.3547	0.4866	0.2527	0.1959	0.2504	0.3981	0.5308
RankSVM	0.2372	0.1799	0.2733	0.3621	0.4989	0.2534	0.1954	0.2499	0.382	0.5187
MixFea	0.2261	0.1689	0.2473	0.3636	0.5002	0.2699	0.2025	0.2519	0.4021	0.5041
ClusCite-Rel	0.2402	0.1872	0.2856	0.4015	0.5156	0.2786	0.2221	0.2753	0.4305	0.5524
ClusCite	**0.2429**	**0.1958**	**0.2993**	**0.4279**	**0.5481**	**0.3019**	**0.2434**	**0.3129**	**0.4587**	**0.5787**



(a) Recall on DBLP (b) Precision on PubMed

Figure 4: Performance comparisons measured by Recall and Precision at different positions.

when modeling topics. The candidates were ranked in terms of the conditional probability of citations from the query manuscript to the candidate papers.

L2-LR [27]: This technique changes the problem into classification with a linearly weighted combination of meta path-based relevance features. Positive examples are observed citations and negative examples are randomly sampled paper pairs.

RankSVM [14]: RankSVM considers the preference between paper-paper relationships, instead of assuming all unobserved relationships are negative examples.

MixFea: the candidates were ranked by a linear combination of meta path-based relevance features, topic distributions and PopRank's features. We used RankSVM to estimate feature weights.

ClusCite: candidates were ranked based on the scores computed by Eq. (1). We set the number of interest groups $K = 200$, $c_p = 10^{-6}$, $c_w = 10^{-7}$, $\lambda_{\mathcal{A}} = \lambda_{\mathcal{V}} = 0.3$ after tuning them on validation sets (Fig. 5 and Sec. 5.6).

ClusCite-Rel: candidates were ranked based on the proposed model with only meta path-based relevance features, *i.e.*, by dropping $\mathbf{F}_{\mathcal{P}}$ in Eq. (1). It used the same settings on K, c_p and c_w as those of ClusCite.

5.2.2 Evaluation Metrics

We employed Precision and Recall at position M (P@M and R@M) as the evaluation metrics. Recall@M is defined as the percentage of original citing papers that appear in the top-M recommended list. A high recall with a lower M indicates a better citation recommendation system. Precision@M was also used to measure the effectiveness of the recommendation system by checking whether the original citing papers were ranked high for the query manuscript.

Furthermore, it is desirable that ground truth papers should appear earlier in the top-M recommended list. Therefore, Mean Reciprocal Rank (MRR) was also employed over the target papers, which is defined as $\text{MRR} = \frac{1}{|\mathcal{Q}_T|}\sum_{q\in\mathcal{Q}_T}\frac{1}{\text{rank}(q)}$, where $\mathcal{Q}_T$ is the testing set and rank(q) denotes the rank of its first ground truth paper (positive example).

5.3 Performance Comparison

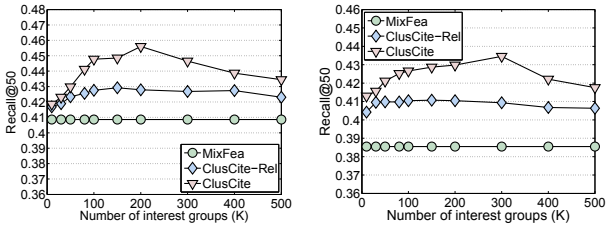
We now compare the proposed recommendation model (ClusCite) with its variation (ClusCite-Rel) and other baselines in terms of the citation recommendation performance.

First, we compare the proposed methods with seven different baselines using Precision@10, 20, Recall@20, 50 and MRR. Table 5 summarizes the comparison results on both DBLP and PubMed datasets. Overall, the proposed ClusCite method and its variation ClusCite-Rel outperform other methods on all metrics. In particular, ClusCite obtains a 17.68% improvement in Recall@50 and 9.57% improvement in MRR compared to the best baseline on the DBLP dataset. On the PubMed dataset, it improves Recall@20 by 20.19% and MRR by 14.79% compared to MixFea. Even though MixFea has incorporated a rich set of features, ClusCite obtained superior performance because it not only explores citation behaviors by learning group-based feature weights over different relation semantics, but also integrates relative paper authority to augment the recommendation process.

The ClusCite-Rel method outperforms all other baselines and improves Recall@50 by 10.42% compared to the best baseline, MixFea, on the DBLP dataset. Comparing ClusCite-Rel with methods such as RankSVM and L2-LR, one can clearly notice the performance gain from distinguishing relevance feature weights for different interest groups. ClusCite always outperforms ClusCite-Rel, improving MRR by 12.21% and Recall@50 by 6.57% on the DBLP dataset. The enhancement mainly comes from utilizing paper relative authority with respect to different interest groups. Also, the derived relative authority can assist recommendation since it is jointly learned through the unified optimization.

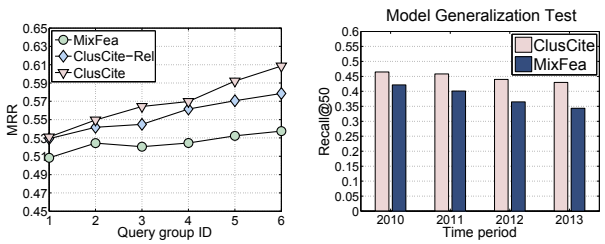
MixFea is another method that incorporates paper authority information, but it does not distinguish paper authority in different interest groups. However, it still obtained better results than RankSVM and L2-LR did in most cases. This demonstrates the effectiveness of paper authority information in the citation recommendation process. Furthermore, poor performance of PopRank shows that using only global authority is not sufficient to conduct good citation recommendation. Different from the conclusions in [10], We found that Link-PLSA-LDA and TopicSim can only achieve 0.0893 and 0.0273 for Precision@20 (compared to 0.1677 with L2-LR), respectively. Also, BM25 outperformed both of the topic-based methods in all cases. This shows that topic-based features are not good enough for finding relevant papers, since the features may be of coarse granularity.

For more comprehensive comparisons, we computed the precision and recall at different positions (5 to 100) to study the trends in performance changes. Due to space limits, Fig. 4 only shows the comparison results of Recall on DBLP and comparison results of Precision on PubMed, respec-



(a) DBLP (b) PubMed

Figure 5: Performance change (Recall@50) on validation sets, with respect to number of interest groups (K) (MixFea as baseline).



(a) # attribute objects (b) time periods of queries

Figure 6: Performance change with number of attribute objects and time periods of query papers.

tively. For both precision and recall, the performance gap between ClusCite and ClusCite-Rel gets slightly larger as more candidates are returned. This indicates that authority information played a critical role in identifying papers with moderate relevance to the query (people may cite relevant papers even though they are new and less reputed, but they prefer authoritative ones among the less relevant papers).

5.4 Performance Analysis

In this section, we analyze the performance of ClusCite, ClusCite-Rel and MixFea in different recommendation scenarios. We ran the following experiments on both datasets and observed similar performance changes in both. However, in the interest of brevity, we only present results from the PubMed dataset for some analyses.

First, we studied performance change with respect to the number of interest groups for ClusCite and ClusCite-Rel. As presented in Fig. 5(a) and 5(b), although not very sensitive to K, these two methods did perform differently when the number of groups were varied. Also, the performance changes were more notable at smaller K, *i.e.*$K < 100$. This indicates that the proposed methods cannot determine citation behavior well when the number of groups is small. On the other hand, a large K (*e.g.*$K > 300$) caused a performance drop due to the insufficiency of training data in deriving interest groups. We found that ClusCite achieved the best performance when the number of groups was $K = 200$ while ClusCite-Rel obtained the best performance with a large number $K = 300$. This shows that biomedical domain has more diverse citation behavior patterns.

In ClusCite and ClusCite-Rel, the paper-specific recommendation makes a prediction for a query based on its attribute objects. Therefore, we want to examine their performance change by studying the correlation between recommendations of the two proposed methods and the number of attribute objects in the query manuscript. We divided the test set into 6 groups with respect to the number of attribute objects. The resulting query groups had an average number of attribute objects ranging from 6.46 (group 1) to 18.98 (group 6). The results by MRR are summarized in Fig. 6(a). Overall, ClusCite outperformed ClusCite-Rel, and both outperformed MixFea. The proposed methods achieved a larger performance improvement when the number of attribute objects increased (*e.g.*from 0.02 in group 1 to 0.08 in group 6) while the performance of MixFea seemed less sensitive between different query groups. This demonstrates that with more attribute objects provided by the query manuscript, the proposed method can make better paper-specific recommendations because richer attribute objects provide better estimation on group membership of the query manuscript.

Finally, we tested the model generalization by evaluating performance on test papers from different time periods. We generated four test subsets using papers in T_2 of the PubMed dataset where each subset consists of papers from one specific year. By applying the methods on each subset, we want to study how the model, learned from papers in T_0, can predict citations for future papers. The study results are shown in Fig. 6(b). Overall, the performance of both methods dropped when recommending for newer papers but ClusCite always outperformed MixFea. Recall@50 of MixFea decreased by 16.42% from year 2010 to 2013 while Recall@50 of ClusCite dropped only about 7.72%, which indicates the better generalization of the proposed method.

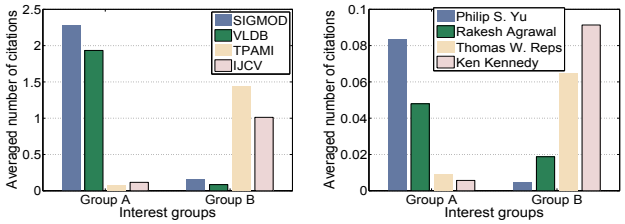
5.5 Case Studies

To demonstrate the effectiveness of mining hidden interest groups, we conduct two sets of case studies on the DBLP dataset to show citation behavioral patterns (Fig. 7) and relative authority ranking of authors and venues, (Table 6) within an example of mined interest groups.

First, we show that the learned interest groups have distinct citation behavioral patterns and can satisfy different information needs. We apply K-means clustering on all objects' group membership indicators and derive their most likely groups (we set $K = 40$). Two representative groups were picked where group A contained 8,345 papers and 208 venues and group B contained 10,922 papers and 291 venues. We found that major venues in group A were database venues (*e.g.*, "SIGMOD" and "VLDB") and those in group B were computer vision venues (*e.g.*, "TPAMI" and "IJCV"). To study how the four venues were cited by papers in the two interest groups, we calculated the average number of citations from papers in group A and B to the four venues, respectively. The results are in Fig. 7(a). One can see that papers in group A prefer to cite database papers while those in group B cite computer vision papers more frequently.

Following a similar procedure, we selected two more representative groups and studied their papers' citations on four different authors: data mining researchers "Pillip S. Yu" and "Rakesh Agrawal" from group A and programming language researchers "Thomas W. Reps" and "Ken Kenndedy" from group B. The average number of citations for these four authors are summarized in Fig. 7(b). Similar behavioral patterns were observed that papers in group A cite data mining researchers more frequently while papers in group B prefer programming language researchers. The derived interest groups show two different behavioral patterns on citations, and justify that they can capture different citation interests.

Second, we study the effectiveness of the relative authority propagation process in the proposed ClusCite algorithm. By setting the number of interest groups as $K = 40$, we apply ClusCite on the training data and obtain relative authority scores for authors ($\mathbf{F}_{\mathcal{A}}$) and venues ($\mathbf{F}_{\mathcal{V}}$). We can list the top ranked objects based on their relative authority scores within different interest groups. Table 6 shows the ranked lists for two example interest groups. One can easily identify



(a) Citations on venues (b) Citations on authors

Figure 7: Case studies on citation behavioral patterns among different interest groups. We show the averaged number of citations on four venues and four authors, for two groups of papers.

the research areas that these two interest groups belong to: Group I is on database and information system while Group II is on computer vision and multimedia. There is a high degree of consensus between the ranking list generated by ClusCite and the top venues and reputed authors in each research area. This demonstrates that the relative authority propagation can generate meaningful authority scores with respect to different interest groups.

Table 6: Top-5 authority venues and authors from two example interest groups derived by ClusCite.

Rank	Venue		Author	
	Group I (database and information system)			
1	VLDB	0.0763	Hector Garcia-Molina	0.0202
2	SIGMOD	0.0653	Christos Faloutsos	0.0187
3	TKDE	0.0651	Elisa Bertino	0.0180
4	CIKM	0.0590	Dan Suciu	0.0179
5	SIGKDD	0.0488	H. V. Jagadish	0.0178
	Group II (computer vision and multimedia)			
1	TPAMI	0.0733	Richard Szeliski	0.0139
2	ACM MM	0.0533	Jitendra Malik	0.0122
3	ICCV	0.0403	Luc Van Gool	0.0121
4	CVPR	0.0401	Andrew Blake	0.0117
5	ECCV	0.0393	Alex Pentland	0.0114

5.6 Parameter Study

In this section, we study the impact of four parameters: c_p and c_w in ClusCite and ClusCite-Rel, and $\lambda_{\mathcal{A}}$ and $\lambda_{\mathcal{V}}$ in ClusCite, on validation sets. The number of interest groups are set as $K = 200$. MixFea, the best baseline, is the only one used here. For conciseness, only DBLP dataset results are presented in Fig. 8, where the x-axes are in log scale.

In the joint optimization problem in Eq. (7), c_p and c_w control the strength of Tikhonov regularizers on group membership indicators and relevance feature weights. A larger value imposes a higher penalty on the magnitude of variable values. We vary one of these two parameters while fixing the other as zero. For ClusCite, we set $\lambda_{\mathcal{A}} = \lambda_{\mathcal{V}} = 0.1$. Both ClusCite and ClusCite-Rel show robust performance over a large range of c_w (Fig. 8(a)) and achieve significant improvement compared to MixFea. We observe a similar trend when varying c_w (Fig. 8(b)) but ClusCite performs slightly better when $c_w = 10^{-7}$. Such changes are because $\mathbf{W}$ plays a role in balancing relevance and authority scores for the ClusCite model while scaling of $\mathbf{P}$ will not affect the ranking results.

ClusCite has two more parameters $\lambda_{\mathcal{A}}$ and $\lambda_{\mathcal{V}}$, which control relative importance of authority information from authors and venues, respectively. By setting one to zero and varying the other, we aim to see a performance change when only one information source is utilized in the authority propagation process. Using ClusCite-Rel and MixFea as baselines, one can see that both information sources help improve the performance of ClusCite significantly. ClusCite achieves the best performance when $\lambda_{\mathcal{A}} = 0.3$ (Fig. 8(c)) and $\lambda_{\mathcal{V}} = 0.3$ (Fig. 8(d)). In particular, we found that applying venue information to authority propagation led to better results.

6. RELATED WORK

6.1 Citation Recommendation

Existing work leverages different kinds of information to recommend citations for a query manuscript, from paper content, known citations to authors and venues of a paper.

Traditional keyword-based approaches have difficulty in finding conceptually similar work due to the ambiguity of short-text queries [20, 5]. One can notice that the performance of BM25 in our experiments is much worse than those of the hybrid methods like L2-LR. Using citation local contexts, *i.e.*, text surrounding the citation positions, context-based methods can capture diverse information needs more precisely [24, 10, 12]. However, the local context might be irrelevant to the ideas of cited paper. Moreover, picking the size of each context window is non-trivial. Also, it will be interesting to study different intents and purposes behind the citation contexts to leverage them more accurately.

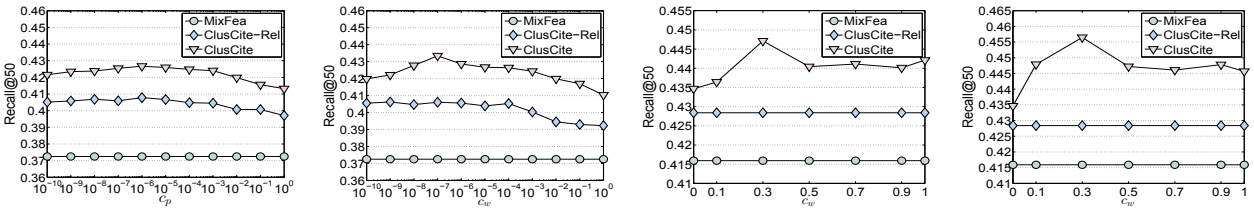
On the other hand, known citations can be used to measure paper structural similarity. Traditional link prediction techniques [16, 2] and collaborative filtering techniques [11] encounter cold-start issue since in practice little or no citations are provided for query manuscript. Heterogenous link prediction techniques [17, 6] tackle this issue by taking advantages of multiple types of relationships between papers, authors and venues. However, these link-based methods cannot achieve satisfactory results without considering content-based features. Therefore, recent studies start integrating both content and structure information to augment the performance [20]. Latent topic models are used to predict citations for new documents by modeling citation links jointly[18, 24]. However, topical similarity may be too coarse to serve as good evidence for citation prediction and experimental results on TopicSim and Link-PLSA-LDA show their limited performance. Yu *et al.* [27] derive a rich set of meta-path based features from heterogeneous bibliographic networks in modeling citation recommendation, which can capture text-based similarity, conceptual relevance as well as several types of social relatedness.

Aforementioned methods consider only paper relevance but ignore another critical information for citation recommendation, namely the importance and quality of target papers [20]. Bethard and Jurafsky [5] built a literature search system by learning a linearly weighted model over both relevance and authority features.

Our work considers diverse citation interests by imposing different feature patternsaccording to the interest groups of each query (see comparisons between ClusCite-Rel and L2-LR [27]). Yu *et al.* [28] study personalized entity recommendation, which shares the similar idea of building local retrieval model for each cluster specifically. Our work is also related to [5] in terms of incorporating paper authority, but we derive paper relative authority within each group specifically (see comparisons between ClusCite and MixFea).

6.2 Authority Ranking on Graphs

Ranking objects on graphs by their importance and popularity has been extensively studied [15, 19] and combined with keyword search system [3]. In particular, Sun and



(a) Varying c_p (b) Varying c_w (c) Varing $\lambda_{\mathcal{A}}$ (d) Varing $\lambda_{\mathcal{V}}$

Figure 8: Performance change (Recall@50) of ClusCite-Rel and ClusCite on DBLP validation set when varying parameters c_p and c_w for both methods, and $\lambda_{\mathcal{A}}$ and $\lambda_{\mathcal{V}}$ for ClusCite. MixFea is used as a baseline.

Giles [21] consider both citation impacts as well as venue influence when propagating paper authority scores in bibliographic networks. With ranking supervision, graph-based semi-supervised ranking frameworks can be further applied[1, 8]. However, these methods do not capture the bias of authority when topics or interests of the query change. (see comparison between PopRank [19] and ClusCite)

Haveliwala [9] personalizes the PageRank algorithm by considering query topics to derive query-specific authority score. Similar ideas were explored when performing clustering [23] and classification [13] in heterogeneous information networks, where object relative authority served as features for representing classes. To our best knowledge, the proposed method is the first to learn object relative authority through optimizing the citation recommendation model, based on multiple types of relationships in heterogeneous bibliographic networks.

7. CONCLUSION AND FUTURE WORK

In this paper, we study citation recommendation in the context of heterogeneous bibliographic networks and propose a novel cluster-based citation recommendation framework to satisfy a user's diverse citation intents. By organizing paper citations into interest groups, the proposed method is able to determine the significance of different structural relevance features for each group, and derive paper's relative authority within each group. In this way, we can make paper-specific recommendations to capture each query's diverse information needs. We formulate a joint optimization problem to learn model parameters by taking advantage of multiple relationships in the network, and develop an efficient algorithm to solve it. Performance evaluation results show a significant improvement compared to state-of-the-art methods and the case studies demonstrate the effectiveness of the proposed method.

Interesting future work includes extending the proposed clustering-based recommendation framework for Web search tasks or entity recommendation so that one can capture local relevance and authority jointly. In addition, there is potential to adjust the network structure for each interest group so that relative authority can more accurately propagate within the corresponding sub-networks. Finally, one can integrate object authority information with each meta path instance to design novel features for citation recommendation.

8. ACKNOWLEDGEMENTS

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, U.S. National Science Foundation grants CNS-0931975, IIS-1017362, IIS-1320617, IIS-1354329, NASA NRA-NNH10ZDA001N, DTRA, and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

9. REFERENCES

[1] A. Agarwal, S. Chakrabarti, and S. Aggarwal. Learning to rank networked entities. In *SIGKDD*, 2006.
[2] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM*, 2011.
[3] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *VLDB*, 2004.
[4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.
[5] S. Bethard and D. Jurafsky. Who should I cite: learning literature search models from citation behavior. In *CIKM*, 2010.
[6] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *ICDM*, 2012.
[7] Q. Gu, J. Zhou, and C. Ding. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *SDM*, 2010.
[8] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *SIGIR*, 2009.
[9] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, 2002.
[10] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles. Context-aware citation recommendation. In *WWW*, 2010.
[11] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM*, 2008.
[12] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach. Recommending citations: translating papers into references. In *CIKM*, 2012.
[13] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *SIGKDD*, 2011.
[14] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.
[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
[16] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.
[17] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon. Supervised link prediction using multiple sources. In *ICDM*, 2010.
[18] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *SIGKDD*, 2008.
[19] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *WWW*, 2005.
[20] T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *SIGIR*, 2007.
[21] Y. Sun and C. L. Giles. Popularity weighted ranking for academic digital libraries. In *ECIR*. 2007.
[22] Y. Sun, J. Han, X. Yan, S. P. Yu, and T. Wu. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. In *VLDB*, 2011.
[23] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *SIGKDD*, 2009.
[24] J. Tang and J. Zhang. A discriminative approach to topic-based citation recommendation. In *PAKDD*. 2009.
[25] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *SIGKDD*, 2008.
[26] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.
[27] X. Yu, Q. Gu, M. Zhou, and J. Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, 2012.
[28] 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 *WSDM*, 2014.