Coranking the Future Influence of Multiobjects in Bibliographic Network Through Mutual Reinforcement

SENZHANG WANG, Beihang University
SIHONG XIE, University of Illinois at Chicago
XIAOMING ZHANG, Beihang University
ZHOUJUN LI, Beihang University; Capital Normal University
PHILIP S. YU, University of Illinois at Chicago; Tsinghua University
YUEYING HE, National Computer Network Emergency Response Technical Team/Coordination Center of China

Scientific literature ranking is essential to help researchers find valuable publications from a large literature collection. Recently, with the prevalence of webpage ranking algorithms such as PageRank and HITS, graph-based algorithms have been widely used to iteratively rank papers and researchers through the networks formed by citation and coauthor relationships. However, existing graph-based ranking algorithms mostly focus on ranking the current importance of literature. For researchers who enter an emerging research area, they might be more interested in new papers and young researchers that are likely to become influential in the future, since such papers and researchers are more helpful in letting them quickly catch up on the most recent advances and find valuable research directions. Meanwhile, although some works have been proposed to rank the prestige of a certain type of objects with the help of multiple networks formed of multiobjects, there still lacks a unified framework to rank multiple types of objects in the bibliographic network simultaneously. In this article, we propose a unified ranking framework *MRCoRank* to corank the future popularity of four types of objects: papers, authors, terms, and venues through mutual reinforcement. Specifically, because the citation data of new publications are sparse and not efficient to characterize their innovativeness, we make the first attempt to extract the text features to help characterize innovative papers and authors. With the observation that the current trend is more indicative of the future trend of citation and coauthor relationships, we then construct time-aware weighted graphs to quantify the importance of links established at different times on both citation and coauthor graphs. By leveraging both the constructed text features and time-aware graphs, we finally fuse the rich information in a mutual reinforcement ranking framework to rank the future importance of multiobjects simultaneously. We evaluate the proposed model through extensive experiments on the ArnetMiner dataset containing more than 1,500,000 papers. Experimental results verify the effectiveness of *MRCoRank* in coranking the future influence of multiobjects in a bibliographic network.

This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 61170189, 61370126, 61202239), National High Technology Research and Development Program of China under grant (No. 2015AA016004), Major Projects of the National Social Science Fund of China under grant (No. 14&ZH0036), Science and Technology Innovation Ability Promotion Project of Beijing (PXM2015-014203-000059), Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2015ZX-16), and US NSF through grants III-1526499, CNS-1115234, and OISE-1129076.
Authors' addresses: S. Z. Wang and X. M. Zhang, State Key Laboratory of Software Development Environment, Beihang University; emails: {szwang, yolixs}@buaa.edu.cn; Z. J. Li (Corresponding author), (1) State Key Laboratory of Software Development Environment, Beihang University, and (2) Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University; email: lizj@buaa.edu.cn; S. H. Xie, Department of Computer Science, University of Illinois at Chicago; email: sxie6@uic.edu; P. S. Yu, (1) Department of Computer Science, University of Illinois at Chicago, and (2) Institute for Data Science, Tsinghua University; email: psyu@uic.edu; Y. Y. He, National Computer Network Emergency Response Technical Team/Coordination Center of China; email: hyy@cert.org.cn.

DOI: http://dx.doi.org/10.1145/2897371

Categories and Subject Descriptors: C.2.14 [**AI Technology**]: Data Mining and Knowledge Discovery; C.1.27 [**Systems and Applications**]: Social and Information Networks

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Influence mining, mutual reinforcement, literature ranking

ACM Reference Format:
Senzhang Wang, Sihong Xie, Xiaoming Zhang, Zhoujun Li, Philip S. Yu, and Yueying He. 2016. Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. ACM Trans. Intell. Syst. Technol. 7, 4, Article 64 (May 2016), 28 pages.
DOI: http://dx.doi.org/10.1145/2897371

1. INTRODUCTION

As one of the centric research issues in scientometrics, literature ranking has been extensively studied to help researchers catch up on the most recent advances [Garfield 1972; Walker et al. 2007; Nerur et al. 2005; Jiang et al. 2013]. Although remarkable efforts have been devoted to ranking the current importance of papers and researchers [Zhou et al. 2007; Jiang et al. 2012; Ding et al. 2009; Zhang et al. 2011], how to identify potentially influential new papers and young researchers and predict their future influence is less touched upon. Ranking the future importance of scientific literature is essential in the following two scenarios: First, it may facilitate researchers quickly entering an emerging research area and exploiting new research fields. For example, in order to grasp the essence of an emerging research direction like social computing or cloud computing, researchers may be particularly interested in the questions: "Which papers published recently may become popular in the future that I should read?" and "Which young researchers will probably become influential so that I should follow their works or cowork with them?" Second, accurately predicting potentially influential new papers and young researchers may direct policymakers to select valuable candidates for research funding [Jiang et al. 2013]. Motivated by these real applications, instead of ranking the current importance of all the literature and researchers, in this article we focus on *predicting the future influence of new publications and young researchers*.

Traditional scientific literature ranking models can be roughly divided into citation-count-based methods [Garfield 1972; Nerur et al. 2005; Hirsch 2005; Egghe 2006] and graph-based ranking methods [Zhou et al. 2007; Jiang et al. 2012; Sayyadi and Getoor 2009; Walker et al. 2007; Ding et al. 2009; Li et al. 2008; Zhang et al. 2011; Bras-Amorós et al. 2010]. Citation count is a simple but useful measurement to rank the importance of papers and authors [Garfield 1972; Nerur et al. 2005]. Based on citation count, some more complicated metrics are proposed, such as h-index [Hirsch 2005], g-index [Egghe 2006], and s-index [Silagadze 2011]. The major limitation of citation-count-based methods is that they ignore the available structure information such as citation and coauthor graphs. To leverage the network structure information, recently many studies have focused on applying graph-based approaches to literature ranking [Zhou et al. 2007; Jiang et al. 2012; Zhang et al. 2011]. For example, Zhou et al. [2007] proposed to combine citation, authorship, and coauthorship networks to simultaneously rank publications and authors. Jiang et al. [2012] leveraged networks of papers, authors, and venues to set up a unified mutual reinforcement model to rank papers, authors, and venues. Graph-based methods can usually obtain more reasonable ranking results, because they take both the popularity (citation count) and prestige (link information) of publications into consideration.

Although plenty of literature ranking models have been proposed, there still lacks a unified framework to rank the future trend of multiple objects in the highly dynamic and complex heterogenous bibliographic network. Most existing approaches focus on the current influence ranking, such that they usually bias classic old papers and famous

researchers. Papers and researchers that are already widely known can be easily searched. Instead of finding such papers and researchers, people might be more interested in the new papers and young researchers and hope to identify the potentially influential ones in the future. In such a case, the traditional ranking model may not work well. Although some attempts have been made to rank the future importance of publications [Sayyadi and Getoor 2009; Walker et al. 2007], they only aim to rank one type of object. Coranking the future influence of multiple objects in the bibliographic network remains an open problem. Another issue of previous related works is that the text information of the papers is largely ignored, while the text may also provide important clues to improve the ranking results. Different from webpage texts, the texts of papers are much more formal and less noisy [Si et al. 2013; Liu et al. 2005]. Properly using the text information may help us better discover potentially popular research topics, based on which related papers and researchers can be also identified. In this article, we will study whether the text information helps, and if it does, how to use such information. For the first time, to the best of our knowledge, we extract and utilize the texts of papers and fuse them with the link structure information in a unified ranking model to better predict the future prestige of papers, authors, and terms simultaneously.

Future influence ranking provides us opportunities to better capture the future research directions and presents new challenges. First, the literature networks are rather dynamic and can evolve over time. For example, the citation and coauthor graphs change all the time since the papers keep getting new citations and authors cowork with different authors. It is challenging to capture the dynamic nature of the involving literature networks for better predicting their future trend. Most previous works ignored the dynamic property of the networks, such that the ranking result is usually biased to old articles. The top-ranked papers are usually overwhelmed by the classical ones published many years ago. A similar problem also exists in author ranking. Therefore, it is traditionally hard to effectively rank the valuable new papers and influential young researchers by simply modeling the literature networks as static graphs. Although some previous works have attempted to explore additional information, such as time information for help [Li et al. 2008; Bras-Amorós et al. 2010], the evolving citation and coauthor links are still largely ignored.

The second challenge is, as we mentioned earlier, the text information of papers may also be helpful, but how to model them is a nontrivial problem. Previous models, especially the graph-based ranking approaches, only explore the structure information of the citation and coauthor networks but ignore the content information. For the new published papers with only a handful of citations, traditional approaches might be less effective. In such a case, the texts of the papers may provide us clues to judge their innovativeness and novelty. Generally, more innovative papers are more likely to address new problems or discuss new topics. Thus, such papers may contain more novel text features. For example, social-media-related words and phrases such as "social network," "social media," "Twitter," and "Facebook" have become increasingly popular in recent publications. Early papers on this topic are very innovative, and most of them have obtained hundreds of citations. As for the researchers, similarly, the topics they work on may largely reflect their future trend of influence. Young researchers who exploit relatively new research areas are much more likely to attract more attention and become famous. Therefore, effectively identifying the pioneering papers by capturing their novel text features may potentially help us find influential young researchers. However, the challenge is, *what text features could be helpful and how do we model them?*

In this article, we propose a unified model to rank the future importance of four types of objects simultaneously: papers, authors, venues, and text features. In particular, in order to profile the dynamic nature of various literature graphs, we first construct time-aware weighted literature networks by assigning different weights on the links based

on their establishing time. We believe that new established links are more indicative of their future trend. For example, we discover that the papers that are frequently cited by new papers may probably continue to obtain more citations in the near future than those whose most citations are old. To highlight the more recent links, we give them higher weights according to an exponential decay function in terms of time. To utilize the text information for help, we then present a burst-detection-based method to measure the innovative degree of two kinds of text features, words and word pairs. The high-level idea is that the words or word pairs that have become increasingly popular in recent years can be considered as sensors to indicate the innovativeness of the papers. Papers with more such words or word pairs are more likely to discuss new topics; therefore, they are more likely to obtain many citations in the future. By mapping the text features and papers or authors to bipartite graphs, we further construct paper-text feature and author-text feature graphs. Finally, by combining all the aforementioned constructed graphs, we propose a unified ranking model MRCoRank. Similar to the HITS algorithm, MRCoRank employs the mutual reinforcement relationships across networks of papers, authors, venues, and text features. The intuition is that potentially influential researchers using many novel text features of rising popularity in high-quality venues lead to the potentially important papers; potentially important papers published in high-quality venues and containing many novel text features of rising popularity lead to influential researchers; future venues with good prestige attract influential researchers submitting influential papers; and future popular text features are widely used by future influential researchers in their potentially important papers. In addition, an extra advantage of MRCoRank that existing ranking models do not have is that it can also help us discover research topics of rising popularity by clustering text features based on their co-occurrence.

We summarize the main contributions of this article as follows:

- We propose to characterize the innovative papers and authors by their innovative text features. To extract innovative text features, we propose a burst-detection-based method to measure their innovative degree. To the best of our knowledge, this is the first attempt to leverage the paper content information for literature ranking.
- To capture the dynamic and evolving nature of literature networks, we use the time information in both citation and coauthor graphs. The proposed ranking algorithm is conducted on the time-aware weighted networks instead of the original static graphs.
- A unified ranking model named MRCoRank is proposed by incorporating the extracted text features and constructed weighted graphs. As a mutual reinforcement ranking framework, MRCoRank ranks the future influence of papers, authors, venues, and text features simultaneously. In addition, our approach can also be applied to help identify potentially popular research topics.
- We conduct comprehensive evaluations on the ArnetMiner dataset with more than 1,500,000 papers and over 2,000,000 citations. The results demonstrate that MRCoRank outperforms existing state-of-the-art algorithms, including FutureRank and MutualRank on ranking new papers and young researchers.

The remainder of this article is organized as follows. Next we will review related works. Section 3 will describe how we model the time and content information. Then, we introduce the unified ranking model in Section 4. The experiment and evaluation are given in Section 5. Finally, we conclude this article in Section 6.

2. RELATED WORK

The earliest work on scientific literature ranking was the citation count method proposed by Garfield [1972]. Although very simple, citation count is widely used to measure the importance of papers and researchers. Based on citation count, several more complicated metrics are then proposed, such as the h-index proposed by Hirsch [2005], the

g-index proposed by Egghe [2006], the c-index proposed by Bras-Amorós et al. [2010], and the s-index proposed by Silagadze [2011].

The major limitation of citation-count-based methods is that they only consider articles' popularity but ignore their prestige. With the increasing popularity of the PageRank algorithm, some works tried to model a literature collection as a network and apply a PageRank-like approach to obtain an authority vector for papers or authors by iteratively computing the adjacency matrix. For example, Ding et al. [2009] proposed to apply the PageRank algorithm on the coauthor network to rank the influence of researchers. Li et al. [2008] applied the the PageRank method to the citation network to rank the importance of articles. Similarly, Chen et al. [2007] and Ma et al. [2008] also brought the PageRank algorithm to the citations network of papers to access the relative importance of the publications.

The PageRank method can only work on one type of network, which limits its effectiveness in ranking different kinds of objects. The literature network is usually heterogeneous and contains several different types of related graphs, including the coauthor graph, the citation graph, and the venue-publication graph [Jiang et al. 2012]. Recently, some studies began to consider exploring heterogenous networks to rank multiple entities simultaneously [Zhou et al. 2007; Jiang et al. 2012; Zhang et al. 2011; Ng et al. 2011]. Ng et al. [2011] proposed a coranking scheme, MultiRank, for objects and relations in multirelational data. The Co-Rank algorithm proposed by Zhou et al. [2007] combined the citation network and coauthorship network to improve the ranking results for both authors and articles. Jiang et al. [2012] proposed a unified mutual reinforcement ranking model that involves intra- and internetwork information for ranking papers, authors, and venues. These methods benefit from different graphs, and therefore can usually achieve better ranking results. Very similar to the mutual reinforcement model, Li et al. proposed a unified multimodal interaction-based framework to fulfill four different tasks simultaneously: content-based image retrieval, image annotation, text-based image retrieval, and query expansion. Similar to the idea of mutual reinforcement, the proposed model also assumes that the solution for one type of task can be reinforced by considering the other three types of tasks simultaneously.

Some efforts have also been made to rank the future popularity of publications [Sayyadi and Getoor 2009; Wang et al. 2013; Walker et al. 2007; Wang et al. 2013, 2014]. Walker et al. [2007] proposed to add the publication time of the articles to the ranking model to predict the future citation count of papers. Similarly, FutureRank aims to predict the future popularity of scientific articles [Sayyadi and Getoor 2009]. FutureRank ranked the future prestige scores of papers by the citation network, the authorship information, and the publication time information. The assumption of FutureRank is that recently published papers are more likely to obtain more citations than old ones. Another recent related work is conducted by Wang et al. [2013]. They add the time information to the author-paper relationship to rank the future citations of papers. However, the limitation of the aforementioned works is that the time information is not fully utilized in various literature networks. For example, the citation and coauthor relationships are also time sensitive, but no work has studied these properties to the best of our knowledge.

3. TIME-WEIGHTED LINKS CONSTRUCTION AND TEXT FEATURES EXTRACTION

In this section, we will introduce how to model the time and content information to help us better rank the future influence of scientific literature. First, we will explain why the content information is helpful by data analysis. Then we will propose to use two types of text features, words and word pairs, to characterize papers and authors. For each text feature, we propose a burst-detection-based method to quantitatively

Table I. Notations

Notation	Description
$\boldsymbol{P}$	The set of paper collection
$\boldsymbol{A}$	The set of author collection
$\boldsymbol{V}$	The set of venue collection
$\boldsymbol{F}$	The set of text feature collection
N	The number of papers
M	The number of authors
L	The number of venues
K	The number of text features
f_i	The i_{th} text feature
w_i	The weight of text feature f_i
$\boldsymbol{E}$	The vector indicating the innovative degree of text features
$\boldsymbol{A\_P}$	The vector indicating the future authority of papers
$\boldsymbol{A\_A}$	The vector indicating the future authority of authors
$\boldsymbol{A\_V}$	The vector indicating the future authority of venues
$\boldsymbol{A\_F}$	The vector indicating the future authority of text features
$\boldsymbol{M^{PP}}$	The $\|N\| \times \|N\|$ matrix indicating citation graph
$\boldsymbol{M^{AA}}$	The $\|M\| \times \|M\|$ matrix indicating coauthor graph
$\boldsymbol{M^{PF}}$	The $\|N\| \times \|K\|$ matrix indicating paper-text feature graph
$\boldsymbol{M^{AF}}$	The $\|M\| \times \|K\|$ matrix indicating author-text feature graph
$\boldsymbol{M^{VA}}$	The $\|M\| \times \|L\|$ matrix indicating venue-author feature graph
$\boldsymbol{M^{PV}}$	The $\|N\| \times \|L\|$ matrix indicating venue-paper feature graph

measure its innovative degree. Next, we will present how to construct the time-aware citation and coauthor relationships.

Before describing the proposed approach, we first give some notations in Table I that will be used later.

3.1. Text Feature Extraction

When a new research topic emerges, only a small number of researchers focus on it and publish related papers. Then gradually, more and more researchers become interested in it and begin to follow the pioneers' works. Finally, with more and more related papers published, the topic becomes less fresh and researchers turn to other new problems. From the perspective of citation count, papers published early are more likely to get numerous citations, because more and more papers published later cite them. Contrarily, it becomes harder and harder for the latecomers to get citations, since the topic is outdated and there are too many related papers already.

To support this point, we give a statistical result based on real citation data in Figure 1. The left histogram of Figure 1 shows the number of yearly published papers whose titles contain"associate rule" from 1994 to 2010, and the right histogram shows their corresponding average citation counts. The left histogram demonstrates that the number of yearly published papers on the topic of "associate rule" increases rapidly in the first decade and reaches its peak in 2008. Then, it begins to decrease. Interestingly, the trend pattern shown in the right figure is almost the opposite to the left. The most-cited papers are those published very early. With the increase of published papers on this topic, the citation count of these following papers decreases dramatically. Figure 1 shows that more innovative papers with numerous citations are usually the earlier works addressing new research issues.

Therefore, effectively identifying the early papers about emerging new topics may greatly help us to recognize their potential popularity and guide us in selecting research directions. But the challenge is how to find the pioneering papers early on. Our approach

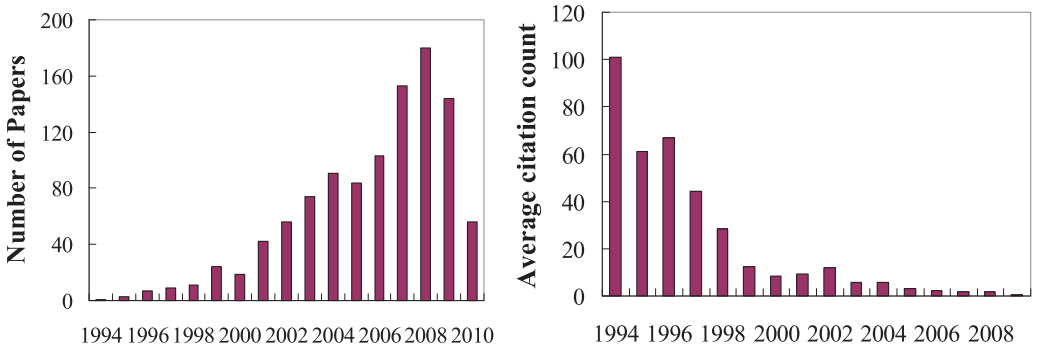


Fig. 1. The number of yearly published papers whose titles contain "associate rule" and their average citation count.

is to consider the text features of papers as the sensors of their innovativeness. A paper addressing a new problem or discussing a new topic may contain more new text features such as new terminologies or new phrases, and hence are more likely to get more citations. To identify such indicative text features, we propose a burst-detection-based schema to quantitatively measure the innovative degree of these text features. Next we will describe what text features we will extract and how to measure their innovativeness.

First, two types of text features, single words and two words co-occurring in the same sentence, are extracted from the titles and abstracts of papers. Some new words emerge and become hot with the emerging of new topics. For example, social networks have gained significant popularity and many related papers are published each year [Wang et al. 2014]. Most of these papers may contain the following words: "Twitter," "Facebook," "Weibo," and "social." Most of these words are relatively new to traditional topics. We also use words' co-occurrence as the text feature. Two explosive co-occurring words may imply the combination of two different topics that may be very innovative. For example, the word pairs "deep-sentiment" and "learning-sentiment" may imply the combination of the topics "deep learning" and "sentiment analysis." In addition, although each separate word may be not new, the co-occurrence of the two words may be new. Taking "deep learning" as an example again: a sentence containing only "deep" or "learning" can hardly to be considered as innovative, but it is probably relatively innovative when the two words appear in the same sentence simultaneously.

Then, we propose to measure the innovativeness of the two kinds of text features by a burst-detection-based method. Burst detection is widely used in event detection in social media [Yao et al. 2010; Kleinberg 2002; Weng et al. 2011; Wang et al. 2015]. Here we apply this technique to measure the innovative degree of each text feature. In event detection, a term is defined as bursty if it frequently occurs in a specified time window but rarely occurs in the past [Weng et al. 2011]. Similarly, we say the paper's text feature is innovative if its frequency increases remarkably in a specified time window. Most of the previous researches on burst feature identification only focused on identifying whether the tags or terms are in stable/burst state for a given time interval [Kleinberg 2002; Weng et al. 2011; Garfield 1972]. In this article, we to need not only identify the burst state of features but also measure their burst degree for predicting their future trend of popularity.

We assume the frequency of each text feature follows the Poisson distribution. Poisson distribution is a discrete probability distribution that can be used to represent the probability of a given number of events occurring in a fixed interval of time. Hence, the

distribution of frequency of text features can be represented as

$$f^i(k, \lambda) = P(x_i = k) = \frac{\lambda_i^k e^{-\lambda_i}}{k!}, \tag{1}$$

where x_i denotes the frequency of text feature i, and λ_i is the mean frequency of x_i. Taking the text feature "cluster" as an example, x_i denotes how many times the word "cluster" has appeared in the papers published in i year, and λ_i is the average appearing times of "cluster" in all the years. The maximum likelihood estimation of λ_i is the sample mean $\widetilde{\lambda_i} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the maximum likelihood estimation of all the features is $\widetilde{\lambda} = \frac{1}{K}\sum_{i=1}^{K} \widetilde{\lambda_i}$.

Definition 1. Regarding each year as a time window and given the feature frequency $x_i^{<t_{j-1},t_j>}$ of the text feature x_i (word or word pair) in the j_{th} time window $<t_{j-1}, t_j>$, we define the degree of innovativeness of feature x_i in the window $<t_{j-1}, t_j>$ as

$$E_i^{<t_{j-1},t_j>} = \frac{|x_i^{<t_{j-1},t_j>} - \widetilde{\lambda_i}|}{\widetilde{\lambda}} \cdot \left[\sum_{s=1}^{u}\left(\frac{x_i^{<t_{j-1},t_j>} - x_i^{<t_{j-s-1},t_{j-s}>}}{\widetilde{\lambda_i}}\right)\frac{1}{s}\right], \tag{2}$$
$$e^{-\rho(t_j - t_0)}$$

where $\widetilde{\lambda_i}$ is the estimated mean frequency of text feature x_i, λ_i is the estimated mean frequency of all the text features, u is the number of previous time windows, and ρ is the time-decaying parameter.

This measurement contains three parts. The first part is the absolute value between feature frequency $x_i^{<t_{j-1},t_j>}$ and the estimated mean frequency $\widetilde{\lambda_i}$. It means that a higher feature frequency will benefit its innovativeness. The second part is the difference between the feature's current frequency $x_i^{<t_{j-1},t_j>}$ and the frequencies in its nearest past s time windows. u is a parameter that limits the number of previous time windows and is set to 3. This part means if the current feature frequency has a significant increment compared with its previous u nearest neighbors, its innovative degree is considered to be high. Meanwhile, to highlight the very early features occurring recently, we use a time-weighted exponential function as the third part. t_0 is the time when the text feature first appears in the paper collection. Figure 2 shows an illustration of our idea. From Figure 2, one can see that early papers (paper 1) on a particular new topic with the text feature x_i get many citations, while papers (paper 3) published later are harder to obtain citations. From the perspective of the text feature curve shown in the lower part, the frequency curve of feature x_i in the early stage shows a sharp increasing pattern, while in the late stage it drops quickly.

Based on the extracted text features and their innovative degrees, the papers and authors can be characterized as follows.

Definition 2. The paper P_i can be characterized as a set of text features $\{f_1^i, \ldots f_l^i, \ldots f_n^i\}$. Each text feature f_l^i can be denoted as a triple $(w_l^{P_i}, E_l^{<t_{j-1},t_j>}, t_i)$, where $w_l^{P_i}$ is the $tf - idf$ weight of the feature i, $E_i^{<t_{j-1},t_j>}$ is the innovative degree of feature i in the j_{th} time window $<t_{j-1}, t_j>$, and t_i is the paper publishing time.

Definition 3. The author A_i can also be characterized as a set of text features $\{f_1^i, \ldots f_l^i, \ldots f_m^i\}$. Each feature f_l^i here can be denoted as such a tuple $(w_l^{Ai}, E_l^{<t_{j-1},t_j>})$.

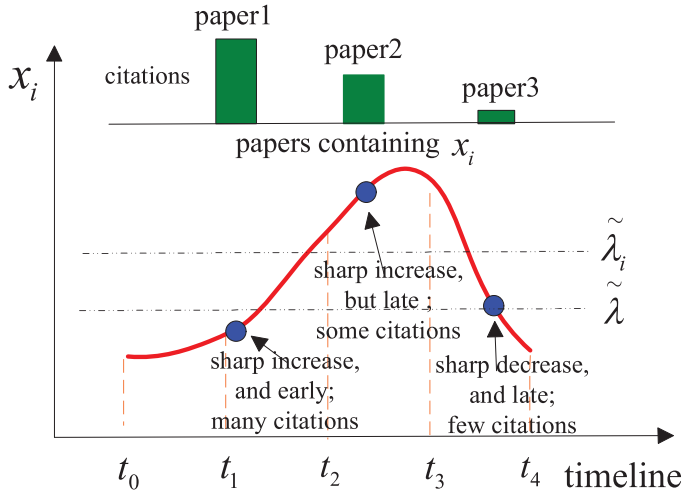


Fig. 2. An illustration of innovativeness measure of text feature x_i. The x-axis is the time, and the y-axis is the frequency of text feature x_i. The upper histogram represents the papers of different years containing the text feature x_i and height of the histogram represents their citation count. The lower red curve shows how the the frequency of text feature x_i changes with time.

Here w_l^{Ai} is $tf - idf$ like the weight of feature i, and $E_l^{<t_{j-1},t_j>}$ is the innovative degree of feature i in the j_{th} time window.

Similar to tf-idf, which is used as a weighting factor to reflect how important a word is to a document in a corpus, we define the weight $w_k^{A_i}$ of feature f_k^i as follows to reflect how important it is to the author A_i in all the authors $\mathbf{A}$.

Definition 4. Given a text feature f_k^i of the author A_i, we define its tf-idf like weight of importance to the author A_i as follows:

$$w_k^{A_i} = tf\left(f_k^i, A_i\right) \cdot idf\left(f_k^i, \mathbf{A}\right), \tag{3}$$

where

$$tf\left(f_k^i, A_i\right) = \frac{f(f_k^i, A_i)}{max\left\{f\left(f_j^i, A_i\right) : f_j^i \in f(A_i)\right\}} \tag{4}$$

$$idf(f_k, \mathbf{A}) = log\frac{|\mathbf{A}|}{|A_n \in \mathbf{A} : f_k \in A_n|}. \tag{5}$$

$f(f_k^i, A_i)$ is the frequency of feature f_k used by author A_i and $max\{f(f_j^i, A_i) : f_j^i \in f(A_i)\}$ returns the highest feature frequency of author A_i. $idf(f_k^i, \mathbf{A})$ is a measure of whether the feature is common or rare across all the authors. The numerator is the number of authors and the denominator is the number of authors using the feature. The two parts are multiplied as the weight of feature f_k to author A_i.

Based on the these definitions, the innovative degree of the paper P_i published at time t can be considered as the sum of all its text features' burst degree:

$$I_{P_i} = \sum_{j=1}^{n} w_j^{P_i} E_j^t. \tag{6}$$

Similarly, the innovation degree of author A_i at time t can be calculated as

$$I_{A_i} = \sum_{k=1}^{m} w_k^{A_i} E_k^t. \tag{7}$$

3.2. Time-Aware Citation and Coauthor Relationships

Some previous related works have tried to use the time information for literature ranking [Li et al. 2008; Sayyadi and Getoor 2009]. For example, to predict the future prestige of papers, Sayyadi and Getoor assume that new published papers are more likely to get more citations than older ones in the future [Sayyadi and Getoor 2009. Nevertheless, not all the new papers will obtain more citations than old ones. Actually, most papers, no matter new or old, get a small number of citations. Only a few papers are frequently cited.

In this article, instead of utilizing the paper publishing time, we apply the time when links were established, such as the time when a paper cites another paper. We assume that the papers frequently cited recently are much more likely to keep obtaining new citations than those whose citations are mostly old.

We denote the paper P_i citing paper P_j at time T_{cite} as $C_{i\to j}(T_{cite})$ with value 1. We propose to utilize the following exponentially decaying equation to measure the weight-of-citation relationship between P_i and P_j:

$$TW_{P_i\to P_j} = e^{-\rho(T_{current}-T_{cite})} C_{i\to j}(T_{cite}). \tag{8}$$

Here ρ is a predefined decaying parameter. In this article, ρ is set to 2.

A researcher who recently coauthors with influential researchers is more likely to keep coauthoring new papers with them. We denote that the author A_i coauthors the paper P_k with A_j at time T_{co} as $A_{i-j}^{P_k}(T_{co})$ with value 1. The time-aware weighted coauthor relationship between them on paper P_k can be represented as

$$TW_{A_i-A_j}^{P_k} = e^{-\rho(T_{current}-T_{co})} A_{i-j}^{P_k}(T_{co}). \tag{9}$$

Two authors may coauthor many papers; thus, the time-weighted coauthor relationship between them over all the papers can be denoted as

$$TW_{A_i-A_j} = \sum_{P_k\in co(A_i,A_j)} e^{-\rho(T_{current}-T_{co}^{P_k})} A_{i-j}^{P_k}(T_{co}), \tag{10}$$

where $co(A_i, A_j)$ is the set of papers coauthored by A_i and A_j.

The venue-paper and venue-author graphs are also time sensitive due to the fact that the prestige of venues evolves over time. Papers published in recent years are more indicative of the future level of the venues than those published many years ago. Hence, the weight of a paper P_j contributing to the future prestige of a venue V_i is related to time and we model it as

$$TW_{V_i-P_j} = e^{-\rho(T_{current}-T_{pub})} V_{i-j}(T_{pub}). \tag{11}$$

$V_{i-j}(T_{pub})$ denotes that paper P_j is published in venue V_i at time T_{pub} with value 1.

Similarly, if a researcher A_j has a paper P_k published in venue V_i at time $T_{pub}^{P_k}$, the weight of the researcher A_j contributing to the future prestige of venue V_i can be modeled as

$$TW_{V_i-A_j}^{P_k} = e^{-\rho(T_{current}-T_{pub}^{P_k})} V_{i-j}^{P_k}\left(T_{pub}^{P_k}\right). \tag{12}$$

A researcher may have many papers published in one venue; hence, the weight of edge from researcher A_j to venue V_i can be calculated by

$$TW_{V_i-A_j} = \sum_{P_k \in pub(V_i, A_j)} e^{-\rho(T_{current}-T_{pub}^{P_k})} V_{i-j}^{P_k}(T_{pub}), \tag{13}$$

where $pub(V_i, A_j)$ denotes the set of papers published in venue V_i by author A_j.

4. MRCORANK: A UNIFIED MUTUAL REINFORCEMENT MODEL FOR CORANKING

In this section, we introduce how to integrate the time-aware weighted graphs and rich text features into a unified Mutual Reinforcement model for Coranking the future importance of scientific articles, authors, venues, and text features simultaneously (MRCoRank). We first briefly summarize the dependency rules of the ranking model on a high level. Then we introduce the constructed various graphs with time-weighted edges on which the MRCoRank runs. Finally, we describe the MRCoRank algorithm in detail.

In brief, the MRCoRank model is based on the mutual reinforcement rules among the following four types of entities connected by the heterogeneous literature networks: papers, authors, text features, and venues. The future importance ranking of the four types of entities is based on the following dependency rules:

- Influential papers are published in high-quality venues, are frequently cited by new published high-quality papers, are usually written by well-known researchers, and contain many innovative text features.
- Influential researchers publish many new high-quality papers in top venues, coauthor papers with other influential researchers, and always catch up on the most recent advances by studying new problems or topics.
- Important venues attract many influential researchers publishing high-quality papers.
- Recent citations are more indicative of their future citations; the influence of recent coauthors is more indicative of the influence of their future coauthors; and the recently published papers are more indicative of a venue's future prestige.

4.1. Literature Graphs

Before describing the algorithm in detail, we first give a brief introduction to the graphs used in the proposed approach. There are four types of nodes, that is, authors, papers, venues, and text features, forming seven types of graphs, that is, coauthor graph, paper citation graph, author-paper graph, venue-paper graph, venue-author graph, author-text feature graph, and paper-text feature graph.

Time-aware coauthor graph M^{AA}. There exists an edge e_{ij} if A_i and A_j coauthor at least one paper. The matrix representation can be defined as

$$M_{ij}^{AA} = \begin{cases} TW_{A_i-A_j} & \text{if } A_i \text{ coauthors papers with } A_j \\ 0 & \text{otherwise.} \end{cases}$$

The coauthor network is an undirected and time-weighted graph, and the weight of each edge is defined in Equation (9).

Time-aware paper citation graph M^{PP}. There exists an edge e_{ij} if paper P_i cites paper P_j. The adjacency matrix of the graph is denoted as

$$M_{ij}^{PP} = \begin{cases} TW_{P_i \to P_j} & \text{if paper } P_i \text{ cites } P_j \\ 0 & \text{otherwise.} \end{cases}$$

The citation graph is a directed and time-weighted graph. The weight of each edge is defined in Equation (8).

Time-aware venue-paper graph M^{VP}. There exists an edge e_{ij} if the paper P_i is published in venue V_j. The adjacency matrix of the venue-paper graph is denoted as

$$M_{ij}^{VP} = \begin{cases} TW_{P_i \to V_j} & \text{if paper } P_i \text{ is published in venue } V_j \\ 0 & \text{otherwise.} \end{cases}$$

The venue-paper graph is a weighted bipartite graph.

Time-aware venue-author graph M^{VA}. There exists an edge e_{ij} if author A_i has a paper published in venue V_j. The adjacency matrix of the venue-author graph is denoted as

$$M_{ij}^{VA} = \begin{cases} TW_{A_i \to V_j} & \text{if author } A_i \text{ has a paper published in venue } V_j \\ 0 & \text{otherwise.} \end{cases}$$

Author-paper graph M^{AP}. This graph contains two kinds of nodes, papers and authors. If A_i is the author of paper P_j, there exists an edge e_{ij}. It is a bipartite graph, and the matrix representation can be denoted as

$$M_{ij}^{AP} = \begin{cases} 1 & \text{if } A_i \text{ is the author of paper } P_j \\ 0 & \text{otherwise.} \end{cases}$$

Paper-text feature graph M^{PT}. This graph also contains two kinds of nodes, papers and text features. If paper P_i contains the text feature f_j, there exists an edge. Its matrix representation is

$$M_{ij}^{PT} = \begin{cases} w_{ij} & \text{if paper } P_i \text{ contains the feature } f_j \\ 0 & \text{otherwise.} \end{cases}$$

w_{ij} is the $tf - idf$ weight of text feature f_j in paper P_i.

Author-text feature graph M^{AT}. This graph contains authors and text features. If author A_i uses the text feature f_j as least once in his or her papers, there exists an edge between them. The matrix representation is

$$M_{ij}^{AT} = \begin{cases} w_{ij} & \text{if } A_i \text{ uses the feature } f_j \\ 0 & \text{otherwise.} \end{cases}$$

w_{ij} is the $tf - idf$ like weight of f_j used by A_i defined in Equations (3) to (5).

4.2. Algorithm

We conduct the MRCoRank algorithm iteratively on the aforementioned graphs by the following steps:

1. Initially, the authority vectors of $\boldsymbol{A\_P}$, $\boldsymbol{A\_A}$, $\boldsymbol{A\_V}$, and $\boldsymbol{A\_F}$ of papers, authors, venues, and text features are set to $\frac{\boldsymbol{I}_N}{N}$, $\frac{\boldsymbol{I}_M}{M}$, $\frac{\boldsymbol{I}_L}{L}$, and $\frac{\boldsymbol{I}_K}{K}$. $\boldsymbol{I}_N$, $\boldsymbol{I}_M$, $\boldsymbol{I}_L$, and $\boldsymbol{I}_K$ are unit vectors. Then repeat steps 2 through 5 until it converges.

2. Based on the dependency rules, update the paper authority vector $\boldsymbol{A\_P}^{t+1}$ by the authority vectors of authors $\boldsymbol{A\_A}^t$, papers $\boldsymbol{A\_P}^t$, venues $\boldsymbol{A\_V}^t$, and text feature $\boldsymbol{A\_F}^t$, and the author-paper matrix $\boldsymbol{M^{AP}}$, the paper citation matrix $\boldsymbol{M^{PP}}$, the venue-paper matrix $\boldsymbol{M^{VP}}$, and the paper-text feature matrix $\boldsymbol{M^{PT}}$ (Equation (14)).

3. Update the author authority vector $\boldsymbol{A\_A}^{t+1}$ by the authority vectors of papers $\boldsymbol{A\_P}^t$, authors $\boldsymbol{A\_A}^t$, venues $\boldsymbol{A\_V}^t$, and text features $\boldsymbol{A\_F}^t$, and the coauthor matrix $\boldsymbol{M^{AA}}$, the author-paper matrix $\boldsymbol{M^{AP}}$, the venue-author matrix $\boldsymbol{M^{VA}}$, and the author-text feature matrix $\boldsymbol{M^{AT}}$ (Equation (15)).

4. Update the venue authority vector $\boldsymbol{A\_V}^{t+1}$ by the authority vectors of papers $\boldsymbol{A\_P}^t$, authors $\boldsymbol{A\_A}^t$, and the venue-author matrix $\boldsymbol{M}^{\boldsymbol{VA}}$ and the the venue-paper matrix $\boldsymbol{M}^{\boldsymbol{VP}}$ (Equation (16)).

5. Update the text feature authority vector $\boldsymbol{A\_F}^{t+1}$ by the authority vectors of papers $\boldsymbol{A\_P}^t$ and authors $\boldsymbol{A\_A}^t$, and the innovative degree $\boldsymbol{E}$ of features, the paper-text features matrix $\boldsymbol{M}^{\boldsymbol{PT}}$, and the author-text features matrix $\boldsymbol{M}^{\boldsymbol{AT}}$ (Equation (17)).

Specifically, the iteration process of the MRCoRank algorithm can be formulated as follows:

future authority of paper P_i

$$A\_P_i^{t+1} = \alpha_{pp} \sum_{P_j \in Cite(P_i)} M_{ij}^{PP} A\_P_j^t + \beta_{pa} \sum_{A_j \in Author(P_i)} M_{ij}^{PA} A\_A_i^t + \gamma_{vp}(1-\alpha_{pp}-\beta_{pa}) M_{ij}^{VP} A\_V_i^t + (1-\gamma_{vp})(1-\alpha_{pp}-\beta_{pa}) \sum_{f_j \in Feature(P_i)} A\_F_j^t M_{ij}^{PT} \tag{14}$$

future authority of author A_i

$$A\_A_i^{t+1} = \alpha_{aa} \sum_{A_j \in Coauthor(A_i)} M_{ij}^{AA} A\_A_j^t + \beta_{pa} \sum_{A_i \in Author(P_j)} \left(M_{ij}^{PA}\right)^T A\_P_j^t + \gamma_{va}(1-\alpha_{aa}-\beta_{pa}) M_{ij}^{VA} A\_V_i^t + (1-\gamma_{va})(1-\alpha_{aa}-\beta_{pa}) \sum_{f_j \in Feature(A_i)} A\_F_j^t M_{ij}^{PT} \tag{15}$$

future authority of venue V_i

$$A\_V_i^{t+1} = \alpha_v \sum_{A_j \in PublishIn(V_i)} M_{ij}^{VA} A\_A_j^t + (1-\alpha_v) \sum_{P_j \in PublishIn(V_i)} M_{ij}^{VP} A\_P_j^t \tag{16}$$

future authority of text features f_i

$$A\_F_i^{t+1} = \left[\alpha_f \sum_{A_j \in Author(f_i)} M_{ij}^{TA} A\_A_j^t + (1-\alpha_f) \sum_{P_j \in Paper(f_i)} M_{ij}^{TP} A\_P_j^t \right] E_i^t \tag{17}$$

Here $Cite(P_i)$ denotes the set of papers that cite paper P_i. $Author(P_i)$ denotes the set of authors of paper P_i. $Feature(P_i)$ denotes the set of text features in paper P_i. $Coauthor(A_i)$ denotes the set of coauthors of author A_i. $Feature(A_i)$ denotes the set of text features used by author A_i. $Author(f_i)$ denotes the set of users who use the text feature f_i. $Paper(f_i)$ denotes the set of papers containing text feature f_i.

We further explain these iteration equations as follows. Equation (14) shows how to update the future authority of papers, which is corresponding to aforementioned step 2. The future authority of paper P_i in iteration $t+1$ is determined by four parts: the authorities of all the papers that cite paper P_i in last iteration t, that is, $\sum_{P_j \in Cite(P_i)} M_{ij}^{PP} A\_P_j^t$; the authorities of all the P_i authors in the last iteration t, that is, $\sum_{A_j \in Author(P_i)} M_{ij}^{PA} A\_A_i^t$; the authors of the venue or journal that P_i is published in, that is, $M_{ij}^{VP} A\_V_i^t$; and the authorities of text features appear in paper P_i, that is, $\sum_{f_j \in Feature(P_i)} A\_F_j^t M_{ij}^{PT}$. The updating rules in Equations (15) through (17) are similar to Equation (14); thus, we omit the explanations on them. α_{pp}, α_{aa}, α_v, α_f, β_{pa}, γ_{vp}, and γ_{va} are all parameters with values from 0 to 1. For simplicity, we consider that paper and author are of the same importance in contributing the future authorities of text features and venues. Hence, we set $\alpha_f = \alpha_v = 0.5$. We also assume that the venue

is equally important in contributing the authority of authors and papers, and simply set $\gamma_{vp} = \gamma_{va}$. At the end of each round of iteration, we normalize $\boldsymbol{A\_P}$, $\boldsymbol{A\_A}$, $\boldsymbol{A\_V}$, and $\boldsymbol{A\_F}$. For example, we normalize the paper authority vector in each iteration as follows: $A\_P_i^t \leftarrow \frac{A\_P_i^t}{\sum_{j=1}^{N} A\_P_j^t}$.

Equations (14), (15), (16), and (17) can be rewritten in matrix forms as follows:

$$\begin{aligned}\boldsymbol{A\_P}^{t+1} =& \alpha_{pp}(\boldsymbol{M^{PP}A\_P}^t) + \beta_{pa}(\boldsymbol{M^{PA}A\_A}^t) + \gamma_{vp}(1-\alpha_{pp}-\beta_{pa})(\boldsymbol{M^{VP}A\_V}^t) \\ &+ (1-\gamma_{vp})(1-\alpha_{pp}-\beta_{pa})(\boldsymbol{M^{PT}A\_F}^t)\end{aligned} \tag{18}$$

$$\begin{aligned}\boldsymbol{A\_A}^{t+1} =& \alpha_{aa}(\boldsymbol{M^{AA}A\_A}^t) + \beta_{pa}((\boldsymbol{M^{AP}})^T\boldsymbol{A\_P}^t) + \gamma_{va}(1-\alpha_{aa}-\beta_{pa})(\boldsymbol{M^{VA}A\_V}) \\ &+ (1-\gamma_{va})(1-\alpha_{aa}-\beta_{pa})(\boldsymbol{M^{AT}A\_F}^t)\end{aligned} \tag{19}$$

$$\boldsymbol{A\_V}^{t+1} = \alpha_v(\boldsymbol{M^{VA}A\_A}^t) + (1-\alpha_v)(\boldsymbol{M^{VP}A\_P^t}) \tag{20}$$

$$\boldsymbol{A\_F}^{t+1} = [\alpha_f(\boldsymbol{M^{TA}A\_A}^t) + (1-\alpha_f)(\boldsymbol{M^{TP}A\_P}^t)]\boldsymbol{E^t}. \tag{21}$$

Equations (22), (23), (24), and (25) can be further rephrased as the following equation:

$$\boldsymbol{R}^{t+1} = \boldsymbol{M}\boldsymbol{R}^t, \tag{22}$$

where $\boldsymbol{R} = [\boldsymbol{A\_P^T}, \boldsymbol{A\_A^T}, \boldsymbol{A\_V^T}, \boldsymbol{A\_F^T}]^T$, and

$$\boldsymbol{M} = \tag{23}$$

$$\begin{pmatrix} \alpha_{pp}\boldsymbol{M^{PP}}\boldsymbol{\Lambda_I} & \beta_{pa}\boldsymbol{M^{PA}} & \gamma_{vp}(1-\alpha_{pp}-\beta_{pa})\boldsymbol{M^{VP}} & (1-\gamma_{vp})(1-\alpha_{pp}-\beta_{pa})\boldsymbol{M^{PT}} \\ \beta_{pa}\boldsymbol{M^{AP}} & \alpha_{aa}\boldsymbol{M^{AA}}\boldsymbol{\Lambda_I} & \gamma_{vp}(1-\alpha_{pp}-\beta_{pa})\boldsymbol{M^{VP}} & (1-\gamma_{va})(1-\alpha_{aa}-\beta_{pa})\boldsymbol{M^{AT}} \\ (1-\alpha_v)\boldsymbol{M^{VP}}\boldsymbol{\Lambda_I} & \alpha_v\boldsymbol{M^{VA}}\boldsymbol{\Lambda_I} & \boldsymbol{\Lambda_0} & \boldsymbol{\Lambda_0} \\ (1-\alpha_f)\boldsymbol{\Lambda_E}\boldsymbol{M^{TP}} & \alpha_f\boldsymbol{\Lambda_E}\boldsymbol{M^{TA}} & \boldsymbol{\Lambda_0} & \boldsymbol{\Lambda_0} \end{pmatrix}.$$

$\boldsymbol{\Lambda_I}$ and $\boldsymbol{\Lambda_E}$ are both diagonal matrixes with the diagonal elements $\Lambda_{ii} = 1$ and $\Lambda_{ii} = E_i$, respectively. $\boldsymbol{\Lambda_0}$ is a zero matrix. The matrix $\boldsymbol{M}$ is a transition matrix corresponding to a Markovian process; thus, it is not hard to verify that $\boldsymbol{R}$ is the eigenvector of matrix $\boldsymbol{M}$, and it will converge to the primary eigenvector.

Details of the algorithm are given in Algorithm 1.

5. EXPERIMENTAL RESULTS

In this section, we will show the effectiveness of the proposed model via extensive evaluations. We start by describing the ArnetMiner dataset used in this article and introduce how we process the dataset. Then we will introduce how we set up the experiments, including the evaluation metric and baselines. As there are many parameters in the proposed model, we next study the sensitivity of the model on these parameters. To give an intuitive understanding of the effectiveness of MFCoRank, we then give comprehensive case studies. The following quantitative comparison with various baselines will show the superior performance of MFCoRank in discovering new papers and young researchers over the metric *recommendation intensity*. Finally, we show the effectiveness of the proposed model in discovering research topics of rising popularity by clustering text features based on their co-occurrence.

5.1. Dataset

The publicly available ArnetMiner dataset on publications[1] is used to evaluate the proposed coranking model [Tang et al. 2008]. It contains 1,572,277 papers and 2,084,019

[1]http://arnetminer.org/citation#b541.

ALGORITHM 1: MRCoRank

Input: Time-aware coauthor graph M^{AA}, time-aware paper citation graph M^{PP}, time-aware venue-paper graph M^{VP}, time-aware venue-author graph M^{VA}, author-paper graph M^{AP}, paper-text feature graph M^{PT}, and author-text feature graph M^{AT}; parameters K, α_{pp}, α_{aa}, α_v, α_f, β_{pa}, γ_{vp}, γ_{va}.

Output: The future authority vectors $\boldsymbol{A\_P}$, $\boldsymbol{A\_A}$, $\boldsymbol{A\_V}$, and $\boldsymbol{A\_F}$ for papers, authors, venues, and text features.

Initialization: $\boldsymbol{A\_P} \leftarrow \frac{I\_N}{N}$, $\boldsymbol{A\_A} \leftarrow \frac{I\_M}{M}$, $\boldsymbol{A\_V} \leftarrow \frac{I\_L}{L}$, $\boldsymbol{A\_F} \leftarrow \frac{I\_K}{K}$;

while $k < K$ **do**

  Update the paper authority vector $\boldsymbol{A\_P}^{t+1}$ based on Equation (22);

  Update the author authority vector $\boldsymbol{A\_A}^{t+1}$ based on Equation (23);

  Update the venue authority vector $\boldsymbol{A\_V}^{t+1}$ based on Equation (24);

  Update the text feature authority vector $\boldsymbol{A\_F}^{t+1}$ based on Equation (25);

  Normalize $\boldsymbol{A\_P}$, $\boldsymbol{A\_A}$, $\boldsymbol{A\_V}$, $\boldsymbol{A\_F}$;

  $k = k + 1$;

return $\boldsymbol{A\_P}$, $\boldsymbol{A\_A}$, $\boldsymbol{A\_V}$, $\boldsymbol{A\_F}$;

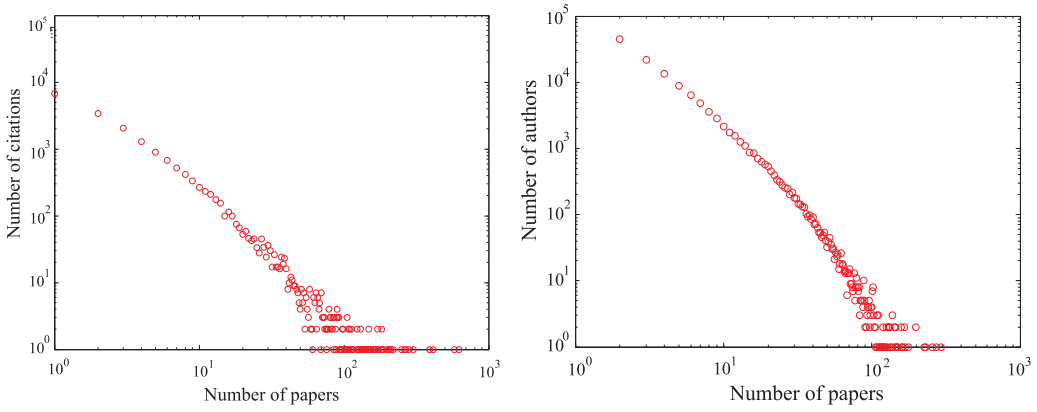
corresponding citations published before 2011. The metadata of each paper contain paper ID, title, abstract, authors, publication year, publication venue, cited papers ID, and citation count. Besides providing the structure information like most bibliographic datasets, the ArnetMiner dataset also contains abstracts of all the papers, which enables us to extract and utilize text features.

Before studying the dataset, we preprocess the dataset as follows. First, as we only rank research papers, we identify survey papers whose titles contain the words "review" or "survey" and eliminate these papers. Survey papers are much easier to obtain a lot of citations for and are easier to search. Second, the papers with no citations and that do not cite other papers are removed, since it is hard to evaluate the influence of papers with no citations currently. Third, the collection contains some workshop proceedings. These proceedings contain all the papers published in the workshop. But in the dataset, the whole proceeding is considered as a "paper." Such proceedings are also removed. In addition, the metadata of most old papers are incomplete. For example, most papers published before 1990 have no citation relationships and abstracts. Therefore, we remove the papers published before 1990 and the papers published after 1990 but with incomplete metadata. For the authors, we only extract and rank the authors in the remaining papers. The authors whose publications are all removed are also cleaned. After the preprocessing, there are 302,336 remaining papers, 39,277 authors, 1,085,181 citations, and over 800 venues or journals.

To have a deeper understanding of the studied dataset, we count the number of published papers for each author and the number of citations for each paper. The distributions of the two relationships are depicted in Figure 3. Both distributions show a power law distribution that is typical in social networks. It means that a few authors have a large number of publications. Most authors, on the contrary, only have a small number of publications. For the papers, a small number of papers get numerous citations, while most get only a few citations.

5.2. Experiment Setup

Ground truth. Almost all the previous works on literature ranking face the same challenge: how to evaluate the results. The key difficulty is that there is little or even no ground truth. A widely used metric to measure the importance of papers and researchers is the citation count [Zhou et al. 2007; Bras-Amorós et al. 2010; Jiang et al. 2012]. In this article, as we aim to rank the future influence of papers and authors, we



(a) # of papers *vs* # of citations (b) # of papers *vs* # of authors

Fig. 3. Statistics of the ArnetMiner literature dataset.

adopt the number of their future citations rather than their current ones as the ground truth [Bras-Amorós et al. 2010]. Then the ground-truth ranking of papers and authors can be obtained by sorting them in the descending order of their future citation counts.

We first divide the dataset into two parts: the ranking part and the evaluation part. Specifically, we select and rank the papers published before 2005 to obtain the ranking lists of papers and authors, and then the ground-truth ranks are obtained by ranking their citation counts from 2005 to 2011. Meanwhile, in order to find what new papers will be the most cited, we select the papers published in the same recent year from the entire ranking list for evaluation. For example, for all the papers published in 2000, which ones will become the most cited? Similarly, we define young researchers as those who begin to publish papers from a specific recent year. For example, for all the researchers starting to publish papers from 2005, who will become more influential? Thus, we only select those starting to publish papers in the same year from the entire ranking list for evaluation.

Evaluation metric. We use the *recommendation intensity (RI)* as the evaluation metric [Jiang et al. 2012, 2013]. The *RI* is based on the following two intuitions: given two ranking results $R1$ and $R2$ on their top-k result list, we think $R1$ is better than $R2$ if (1) $R1$ returns more entities matching the ground-truth ranking, and (2) the matched entities are at the front of the top-k list. Assume R is the list of top-k returned entities of a ranking approach, and L is the list of ground truths; then for each entity P_i in R with the ranked order o_r, the *recommendation intensity* of P_i at k can be defined as

$$RI(P_i)@k = \begin{cases} 1 + (k - o_r)/k & P_i \in L \\ 0 & P_i \notin L. \end{cases} \tag{24}$$

This means that if the entity P_i is in the top-k ground-truth list R and is ranked higher (smaller o_r), then its *recommendation intensity* is higher.

Based on each entity's *recommendation intensity* in the list R, the *recommendation intensity* of the list R at k can be defined as

$$RI(R)@k = \sum_{P_i \in R} RI(P_i)@k. \tag{25}$$

One can see that the *recommendation intensity* is very similar to *precesion-at-k* in an information retrieval context. If we consider the top-k list R as unordered and

divide $RI_{P_i}@k$ by k, *recommendation intensity* will degenerate to the classical metric *precision-at-k*.

5.3. Baselines

We choose the following methods as baselines.

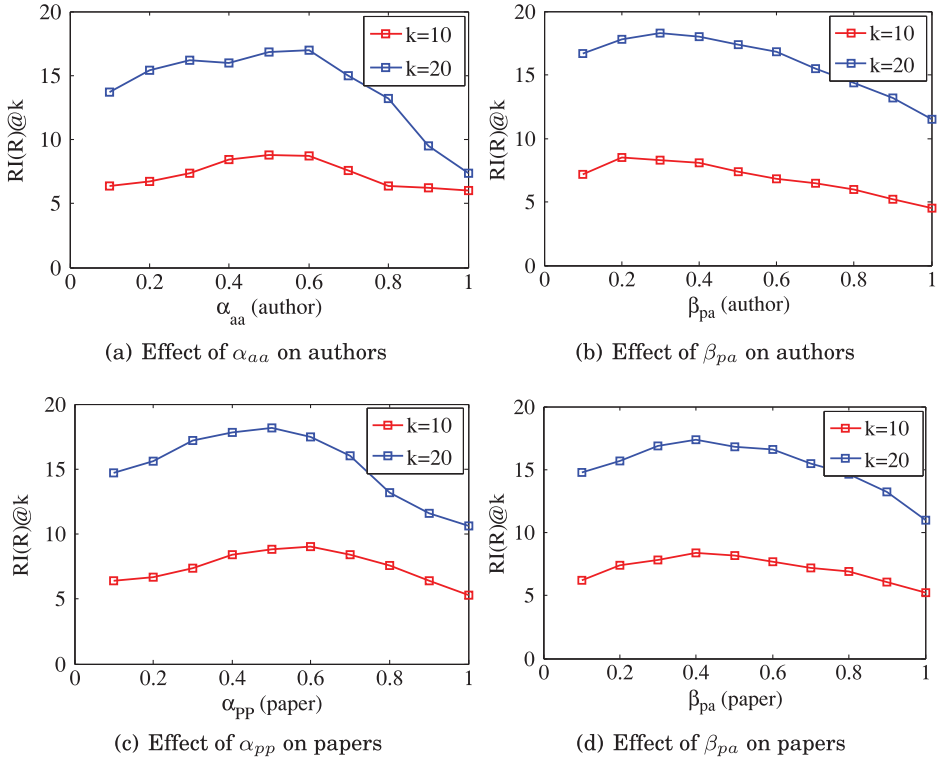
- **Citation Count (CC).** Although very simple and straightforward, citation count can usually achieve reasonable ranking results. In our experiment, we rank the future influence of papers, authors, and venues based on their current citation count.
- **PageRank (PR).** PageRank has achieved great success in ranking web pages and has been widely used in many other applications for ranking the authority of nodes in networks [Page et al. 1999]. Some previous works [Chen et al. 2007; Ding et al. 2009; Bras-Amorós et al. 2010] also tried to utilize PageRank to rank the influence of publications and authors.
- **FutureRank (FR).** FutureRank is a representative method to predict the future important papers proposed recently [Sayyadi and Getoor 2009]. FutureRank combines the authorship network and the publication time of the articles in order to predict future citations. As FutureRank only aims to rank papers, we compare our paper ranking result with it.
- **MutualRank (MR).** MutualRank is the state-of-the-art graph-based method that integrates mutual reinforcement relationships among several graphs to rank papers, authors, and venues simultaneously [Jiang et al. 2012]. However, instead of ranking the future importance, MutualRank aims to rank the current influence of different types of entities.
- **MRFRank (MRFR).** MRFRank is our previously proposed ranking model [Wang et al. 2014]. The main difference between the MRFRank model and the MRCoRank model is that MRFRank only considers and ranks three types of objects, author, paper, and text feature, but ignores the venue information. We compare our new mode MRCoRank with MRFRank to study whether adding the venue information can further improve the performance.

In order to study how much performance can be improved by incorporating the time and text information, we use the following two variations of MRCoRank as baselines: MRCoRank without time information **(MRCoR-T)**, and MRCoRank without text information **(MRCoR-C)**.

5.4. Parameter Sensitivity Analysis

There are seven parameters in the proposed ranking model in all, α_{pp}, α_{aa}, α_f, β_{pa}, γ_{va}, γ_{vp}, and α_v. For simplicity, we assume that paper and author are of the same importance in contributing the future authorities of text features and venues; thus, we set $\alpha_f = \alpha_v = 0.5$. We also set $\gamma_{vp} = \gamma_{va} = \gamma_v$ with the assumption that venue is equally important in contributing the authority of authors and papers. Therefore, there are actually four parameters we need to study: α_{pp}, α_{aa}, β_{pa}, and γ_v.

We first study the parameter sensitivity on the proposed model. Due to space limitation, we only give the results of the parameters α_{aa}, α_{pp}, and β_{pa}. Figure 4 shows the experiment results. For each studied parameter, we first assign it different values from 0.1 to 1 with step length 0.1. Then we obtain the best performance by tuning the other parameters. One can see that the proposed model is more sensitive to the parameters α_{aa} and α_{pp} compared with the parameter β_{pa}. It implies that the coauthor and citation relations play more important roles in ranking both papers and authors than the venue information. For the parameter α_{aa}, the best performance is achieved at around 0.5. For α_{pp}, it shows that $\alpha_{pp} = 0.6$ might be a reasonable choice. One can see that larger or small values for both α_{aa} and α_{pp} will hurt the performance, which



(a) Effect of α_{aa} on authors (b) Effect of β_{pa} on authors

(c) Effect of α_{pp} on papers (d) Effect of β_{pa} on papers

Fig. 4. Parameter sensitivity analysis on MRCoRank model.

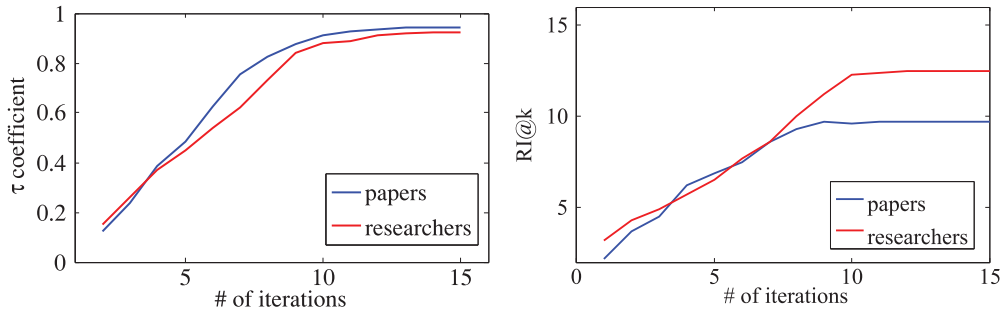
implies that both citation relations and coauthor relations are important to indicate the future influence of papers and authors. The performance will not be desirable if only one type of relations is considered while the others are ignored.

To obtain the group of parameters that can achieve the best ranking performance, we use the exhaustive method to test all the possible combinations of the parameters. Specifically, we first fix parameters α_{pp}, α_{aa}, and β_{pa}, and then set the parameter γ_v from 0 to 1 by step length 0.1. We choose the value of γ_v that achieves the highest *recommendation intensity* as the best parameter we need. Likewise, we find the best parameter values of α_{pp}, α_{aa}, and α_{pa} by fixing the other three parameters, respectively. We find that the proposed MRCoRand achieves the best performance with the parameter settings $\alpha_{pp} = 0.6$, $\alpha_{aa} = 0.5$, $\beta_{pa} = 0.2$, and $\gamma_v = 0.4$. In the following experiments, we use them as the default parameter settings.

5.5. Convergence Analysis

This subsection studies the convergence of the proposed ranking model. Given two ranking lists r_t and r_{t+1}, we calculate their correlation coefficient to measure how similar the two ranking lists are. If the proposed model can converge, the correlation coefficient tends to increase and finally becomes stable during iteration. Here we use the Kendall τ coefficient as the correlation coefficient [Kendall 1938]. The Kendall τ coefficient is defined as

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}. \tag{26}$$



(a) Convergence of τ (b) Convergence of $RI(R)@k$

Fig. 5. Convergence analysis of MRCoRank on papers and researchers.

Given two ranking lists $r_t = \{x_1, x_2, \ldots, x_n\}$ and $r_{t+1} = \{y_1, y_2, \ldots, y_n\}$ in two successive iterations, each pair of elements (x_i, y_i) and (x_j, y_j) are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. In our case, x_i and y_i are the ranking orders of the papers in the ranking lists.

Figure 5(a) shows the increase trends of τ coefficients with the increase of iteration number for papers and researchers. One can see that the proposed MRCoRank algorithm converges very fast: τ coefficients of both paper ranking and researcher ranking tend to become stable after about 10 iterations. To further show the convergence of the algorithm, we also give the convergence curve of the $RI(R)@k$ on papers and researchers in Figure 5(b). One can see that the convergence curve of $RI(R)@k$ is similar to that of τ. For paper ranking, the algorithm converges within nine iterations; for researcher ranking, it converges within about 10 iterations. As the algorithm can converge very fast, it is not time consuming to obtain the ranking results on the large dataset. In our experiments, it only takes less than 2 minutes for the proposed algorithm to converge.

5.6. Case Studies

In this subsection, we give case studies on the ranking results of papers, authors, and venues returned by various ranking approaches in the years of 2001, 2002, and 2005.

Case studies on publications. Tables II through VII list the top 10 potentially influential papers published in the years 2001, 2002, and 2005, respectively, identified by the proposed MRCoRand and baselines. For each table, we list the titles of the top 10 papers returned by the proposed MRCoRand and the published venues in the left two columns. We also list the ground-truth rankings of the top 10 papers returned by different approaches. For clarity, we give the ranking results of papers published in conferences of different research communities separately. Tables II, IV, and VI show the papers published in database-related venues or journals, and Tables III, V, and VII show the papers published in artificial-intelligence-related venues or journals. The figures in bold mean that the rankings of the papers given by these approaches are also in the top 10 rankings of the ground truth.

From Table II, one can see that seven out of the top 10 papers returned by MRCoRank are in the top 10 rankings of the ground truth, compared with five by MRFRank, three by MutualRank and PageRank, and four by FutureRank. Our approach identifies the papers' "Relevant-Based Language Models," which turn out to be very influential, while most baselines fail to give them high rankings. This is because in 2001, the topic of language model this article discussed was relatively new, and our approach captures their novel text features. For the AI papers published in 2001 shown in

Table II. Top 10 Papers in DB Community Published in 2001

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
On Supporting Containment Queries in Relational Database Management Systems	SIGMOD	**7**	**4**	**4**	11	**6**
A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval	SIGIR	**1**	**5**	22	16	12
Optimal Aggregation Algorithms for Middleware	PODS	**3**	**7**	14	23	17
Continuous Queries over Data Streams	SIGMOD	14	12	19	**1**	27
Document Language Models, Query Models, and Risk Minimization for Information Retrieval	SIGIR	**4**	15	**6**	14	**4**
E-Commerce Recommendation Applications	Data Min. Knowl. Discov.	**10**	**1**	37	**8**	18
Mining Time-Changing Data Streams	KDD	11	22	26	**5**	47
Relevance-Based Language Models	SIGIR	**2**	**10**	15	37	35
Flexible Support for Multiple Access Control Policies	ACM Trans. Database Syst.	15	13	**10**	42	**2**
On the Design and Quantification of Privacy Preserving Data Mining Algorithms	PODS	**6**	20	24	**4**	16

Table III. Top 10 Papers in AI Community Published in 2001

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
Semantic Web Services	IEEE Intelligent Systems	13	11	**7**	15	**2**
Soft Margins for AdaBoost	Machine Learning	14	**3**	44	28	17
Agents and the Semantic Web	IEEE Intelligent Systems	**2**	22	35	13	**4**
Latent Dirichlet Allocation	NIPS	**1**	**9**	26	**4**	33
Support Vector Clustering	Journal of Machine Learning Research	30	20	18	26	54
Completely Derandomized Self-Adaptation in Evolution Strategies	Evolutionary Computation	**7**	19	74	32	12
Estimating the Support of a High-Dimensional Distribution	Neural Computation	**6**	**7**	**2**	11	24
Sparse Bayesian Learning and the Relevance Vector Machine	Journal of Machine Learning Research	**5**	25	14	52	17
Random Forests	Machine Learning	**2**	**4**	17	**3**	29
A Machine Learning Approach to Coreference Resolution of Noun Phrases	Computational Linguistics	20	**10**	**6**	**8**	**3**

Table III, MRCoRank also gives better ranking results with six returned papers in the top 10 rankings of the ground truth. The numbers for MRFRank, MutualRank, FutureRank, and PageRank are five, three, three, and three, respectively. Tables IV through VII show the ranking results of papers published in 2002 and 2005. For the papers published in these 2 years, the proposed MRCoRank also performs best. For example, for the database papers published in 2002, five top 10 papers identified by MRCoRank are in the top 10 list of ground truth. The result is much better than MutualRank (two papers), FutureRank (one paper), and PageRank (three papers). For the papers published in 2005, we use their citations from 2005 to 2007 to rank the model, and use the new citations obtained from 2007 to 2011 for evaluation. The results in Tables VI and VII also show that the proposed model MRCoRank outperforms all

Table IV. Top 10 Papers in DB Community Published in 2002

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
Models and Issues in Data Stream Systems	PODS	**2**	**10**	42	11	25
Storing and Querying Ordered XML Using a Relational Database System	SIGMOD	**7**	**4**	34	15	14
Learning to Map Between Ontologies on the Semantic Web	WWW	11	19	**4**	20	64
Middle-Tier Database Caching for e-Business	SIGMOD	107	**2**	72	33	38
Topic-Sensitive PageRank	WWW	**4**	43	18	22	16
Continuously Adaptive Continuous Queries over Streams	SIGMOD	17	22	44	18	12
Optimizing Search Engines Using Clickthrough Data	KDD	**1**	16	35	12	**9**
Containment and Equivalence for an XPath Fragment	PODS	46	17	24	71	21
Accelerating XPath Location Steps	SIGMOD	28	**5**	16	29	**3**
A taxonomy of Web Search	SIGIR	**3**	23	**8**	**5**	**1**

Table V. Top 10 Papers in AI Community Published in 2002

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
Gene Selection for Cancer Classification Using Support Vector Machines	Machine Learning	**2**	**5**	**8**	16	13
Shape Matching and Object Recognition Using Shape Contexts	IEEE Trans. Pattern Anal. Mach. Intell.	**1**	**3**	**10**	23	**7**
Choosing Multiple Parameters for Support Vector Machines	Machine Learning	**10**	14	25	**3**	**1**
Algorithm for Optimal Winner Determination in Combinatorial Auctions	Artif. Intell.	14	19	33	**7**	14
Extending and Implementing the Stable Model Semantics	Artif. Intell.	11	**7**	**9**	18	**8**
Unsupervised Learning of Finite Mixture Models	IEEE Trans. Pattern Anal. Mach. Intell.	**9**	21	45	62	33
Kernel Independent Component Analysis	Journal of Machine Learning Research	17	25	**1**	13	**5**
A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms	International Journal of Computer Vision	**3**	**8**	62	**4**	17
Learning Surface Text Patterns for a Question Answering System	ACL	13	14	11	21	25
A Study of Approaches to Hypertext Categorization	J. Intell. Inf. Syst.	35	**6**	19	**2**	16

the other baselines. For the influential paper "Personalizing Search via Automated Analysis of Interest and Activities," which ranks third in groud truth, all the other methods failed to identify it, while our model gives it a very high ranking. By comparing MRCoRank with our previous model MRFRank, the results show that in most cases, the MRCoRank outperforms MRFRank, which implies that adding the venues' information does help.

Case studies on researchers. Tables VIII, IX, and X show the case studies on the ranking results of the top five researchers who start publishing papers from the years 2000, 2001, and 2004/2005. Likewise, we use boldface figures to denote that the identified top five researchers are also in the top five list of ground truth. For the researchers whose research focus is mainly on databases, one can see that four out of the top five researchers returned by MRCoRank is in the top five list of the ground truth in 2000. The numbers for MRFRank, MutualRank, PageRank, and citation count

Table VI. Top 10 Papers in DB Community Published in 2005

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
Accurately Interpreting Clickthrough Data as Implicit Feedback	SIGIR	**1**	**7**	**4**	26	19
A Markov Random Field Model for Term Dependencies	SIGIR	**8**	**9**	15	74	42
Maximal Vector Computation in Large Data Sets	VLDB	16	26	12	**6**	**1**
Progressive Skyline Computation in Database Systems	TODS	**5**	34	**2**	31	15
Incognito: Efficient Full-Domain K-Anonymity	SIGMOD	**2**	20	**7**	12	42
Personalizing Search via Automated Analysis of Interests and Activities	SIGIR	**3**	**4**	25	**3**	16
Stabbing the Sky: Efficient Skyline Computation over Sliding Windows	ICDE	32	21	37	**1**	**6**
Schema Mappings, Data Exchange, and Metadata Management	PODS	18	16	13	52	44
Top-Down Specialization for Information and Privacy Preservation	ICDE	**10**	**1**	**1**	14	26
Efficient Computation of the Skyline Cube	VLDB	24	37	17	27	**5**

Table VII. Top 10 Papers in AI Community Published in 2005

Titles of Top 10 Papers Returned by MRCoRank	Venue	Rankings in Ground Truth				
		MRCoR	MRFR	MR	FR	PR
Histograms of Oriented Gradients for Human Detection	CVPR	**1**	**1**	**1**	12	**1**
Large Margin Methods for Structured and Interdependent Output Variables	JMLR	**6**	**2**	**6**	**1**	**9**
A Hierarchical Phrase-Based Model for Statistical Machine Translation	ACL	**10**	**7**	23	17	45
Face Recognition Using Laplacianfaces	IEEE Trans. Pattern Anal. Mach. Intell.	**3**	21	15	25	**5**
The HumanID Gait Challenge Problem: Data Sets, Performance, and Analysis	IEEE Trans. Pattern Anal. Mach. Intell.	21	34	**2**	**3**	73
A Support Vector Method for Multivariate Performance Measures	ICML	36	**5**	20	16	22
Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking	ACL	15	**9**	**5**	52	**10**
Working Set Selection Using Second Order Information for Training Support Vector Machines	JMLR	**8**	14	42	**6**	**3**
A Comparison of Affine Region Detectors	Int J Comput Vis	**2**	39	31	**4**	75
A Sparse Texture Representation Using Local Affine Regions	IEEE Trans. Pattern Anal. Mach. Intell.	24	**4**	11	15	26

Table VIII. Top 5 Researchers in DB and AI (Publications Start from 2000)

Database						Artificial Intelligence					
	Rankings in Ground Truth						Rankings in Ground Truth				
Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC	Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC
Jian Pei	**1**	**2**	**5**	9	**3**	Aaron Hertzmann	**2**	**3**	**1**	11	**4**
Ninghui Li	**3**	7	7	11	**2**	Patrick Pantel	**5**	7	7	8	6
Junghoo Cho	**2**	11	**4**	**1**	25	Michael Beetz	**3**	**5**	19	**2**	12
Aristides Gionis	**4**	**4**	15	17	14	Koby Crammer	6	12	16	22	17
Gail-Joon Ahn	12	**3**	23	19	11	Stephen Cranefield	11	6	**5**	14	10

Table IX. Top 5 Researchers in DB and AI (Publications Start from 2001)

Database						Artificial Intelligence					
	Rankings in Ground Truth						Rankings in Ground Truth				
Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC	Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC
Yufei Tao	**2**	**4**	**4**	**2**	**1**	Haixun Wang	**1**	**2**	**4**	**2**	**1**
Haixun Wang	6	**3**	12	10	**4**	Peter McBurney	7	**1**	8	14	8
Shivnath Babu	**5**	**5**	7	**4**	12	David M. Pennock	**4**	10	12	10	**3**
Qiong Luo	16	6	11	17	16	David C. Parkes	6	8	23	16	18
Samuel Madden	**1**	8	9	8	6	Alex Dekhtyar	14	**5**	**9**	11	14

Table X. Top 5 Researchers in DB and AI (Publications Start From 2004/2005)

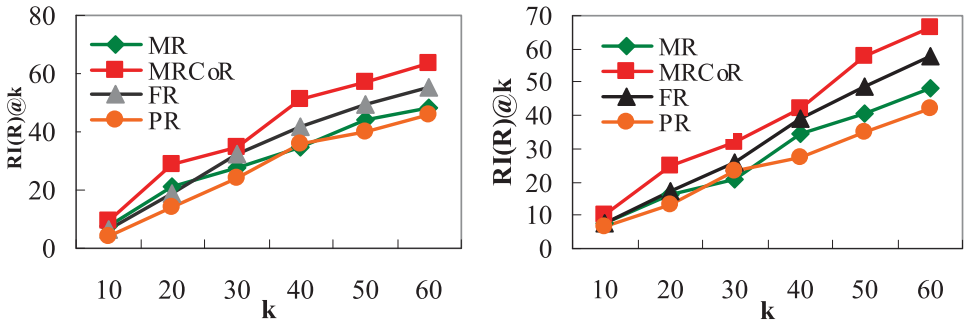
Database						Artificial Intelligence					
	Rankings in Ground Truth						Rankings in Ground Truth				
Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC	Top5 by MRCoR	MRCoR	MRFR	MR	PR	CC
Xiaokui Xiao	**5**	7	21	16	**5**	Chris Callison-Burch	**1**	**4**	**1**	13	8
Bing Pan	**2**	**4**	9	10	10	David Chiang	13	22	16	9	**3**
Benjamin C. M. Fung	**3**	8	**1**	24	31	Dong Xu	**4**	19	42	32	16
Qiaozhu Mei	7	16	**4**	43	7	18Ryan T. McDonald	**2**	7	**3**	**5**	23
Filip Radlinski	12	**2**	15	**2**	6	Vikas Sindhwani	8	**2**	12	6	**4**

Table XI. Top 5 Venues/Journals in DB and AI in 2000/2001

Rank	Database		Artificial Intelligence	
	2000	2001	2000	2001
1	SIGMOD	SIGMOD	IEEE Trans. Pattern Anal. Mach. Intell.	JMLR
2	VLDB	SIGIR	Machine Learning	Machine Learning
3	ICDE	PODS	JMLR	NIPS
4	ACM Trans. Database Syst.	ACM Trans. Database Syst.	ACL	IEEE Trans. Pattern Anal. Mach. Intell.
5	SIGIR	SIGIR	COLING	Computational Linguistics

are three, two, one, and two, respectively. For all the researchers who start to publish papers from 2000, Jian Pei turned out to become the most influential researcher in database, followed by Junghoo Cho, Ninghui Li, and Aristides Gionis. MRCoRank identifies them all and ranks them high. MRCoRank also performs best in the field of artificial intelligence. One can see that three out of the top five researchers in artificial intelligence are also in the top five list of the group truth; while MRFRank and MutualRank have two, citation count and PageRank only have one. The ranking result for the researchers starting to publish papers from 2001 is also desirable. MRCoRank successfully identifies the most potentially influential researchers Samuel Madden and Haixun Wang in database and artificial intelligence, respectively. For the researchers whose publications start from 2004 or 2005, one can see that MRCoRank also performs much better than baselines and identifies more influential younger researchers.

Case studies on venues/journals. The rankings of venues and journals are much more similar for different ranking models as the prestige of venues and journals is rather stable. There are not many new venues and journals appearing every year, so it is hard to evaluate their future prestige. We list the ranking lists of the top five venues and journals of 2000 and 2001 in Table XI in the communities of database and artificial intelligence. One can observe that venues are more influential than journals in database, while journals have higher prestige than venues in artificial intelligence. *SIGMOD* seems to be the most influential conference in the database community, and *ACM Transactions on Database Systems* is the most influential database journal. For



(a) Database (b) Artifitial Intelligence

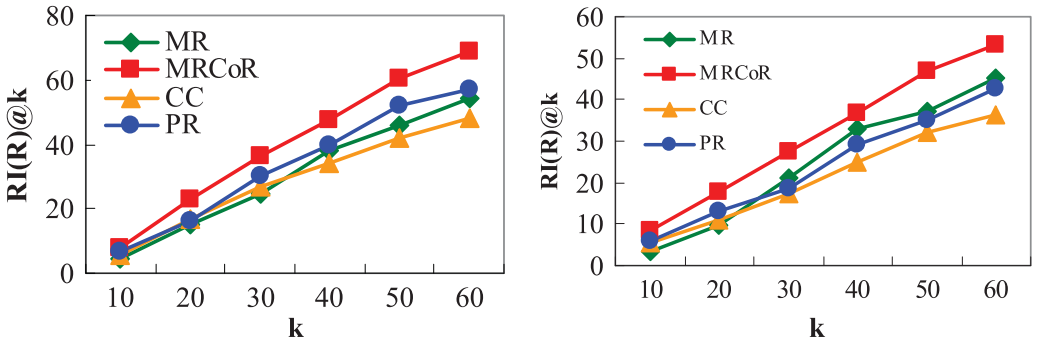
Fig. 6. RI(R)@k of ranked papers in two research fields published in 2000.

the artificial intelligence community, the most influential journals in 2000 and 2001 are *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Machine Learning*, and *Journal of Machine Learning Research*, and influential venues include *ACL*, *NIPS*, and *COLING*.

Comparison with MRFRank To further show the performance improvement achieved by adding the venue information, we give several concrete cases for both papers and researchers. For the top 10 influential papers published in 2005, one can see that there are two papers that are successfully identified by MRCoRank but ignored by MRFRank. The first one is "Progressive Skyline Computation in Large Dat Sets," whose ranking position in ground truth is five. One can see that the ranking positions given by MRCoRank and MRFRank are four and 27, respectively. The second paper is "Incognito: Efficient Full-Domain K-Anonymity" with the ground-truth ranking position two. It is ranked five by MRCoRank and 33 by MRFRank. One can see that both papers were published in top conferences of database (*TODS* and *SIGMOD*). The high prestige of the conferences also increases the authority of the papers. The MRCoRank model takes the conference information into consideration; thus, the two papers are given higher ranking positions. Similar examples can also be found in the top 10 papers published in 2001 ("Sparse Bayesian Learning and the Relevance Vector Machine": ground truth: 5; MRCoRank: 7; MRFRank: 17) and 2002 ("Optimizing Search Engines Using Click Through Data": ground truth: 1; MRCoRank: 4; MRFRank: 43). For the researcher ranking, the performance improvement may not be as significant as that of the paper ranking because a researcher can publish many papers in different venues. However, one can still see that MRCoRank outperforms MRFRank in most cases. Jian Pei is a database-focused researcher whose true ranking is one in 2,000. His ranking position in MRCoRank is one, while MRFRank ranks him eight. From these case studies, one can see that venues do provide us useful information and improve the ranking performance.

5.7. Quantitative Comparison

Next, we quantitatively compare the performance of the proposed approach with baselines. As the results for different years are similar, we only report the results of papers and researchers in 2000 in Figures 6 and 7. Figure 6 shows the $RI(R)@k$ values of ranked papers published in 2000 in the communities of database and artificial intelligence. Figure 7 shows the $RI(R)@k$ values of ranked authors who start to publish papers from 2000. The figures show that for both researcher and paper rankings, the proposed approach MRCoRank outperforms baselines over various k. For the ranking of papers, FutureRank is generally better than MutualRank, but inferior to MRCoRank.



(a) Database (b) Artifitial Intelligence

Fig. 7. RI(R)@k of ranked authors in two research fields whose publications start from 2000.

Table XII. Experiment Results of Comparing MRCoRank with Three Variations

Year	Method	k = 10		k = 20		k = 50	
		Paper	Researcher	Paper	Researcher	Paper	Researcher
2000	MRCoR	**9.4**	**12.3**	**27**	28.4	**60.5**	**58.7**
	MRCoR-T	7.2	8.9	21.7	**29.2**	54	52.6
	MRCoR-C	7.6	11.5	23	25.7	57	56.4
	MRFR	6.8	9.1	18.5	20.5	50.7	52.5
2001	MRCoR	**10.5**	**16.4**	21	**24.8**	**57**	**64.3**
	MRCoR-T	9.0	11.6	18	22.9	50.7	54.4
	MRCoR-C	8.7	10.2	**21.5**	23.4	44	55.3
	MRFR	9.5	13.4	16.8	20.4	44.8	52.4
2002	MRCoR	**7.7**	**9.6**	**12.8**	14.2	**38.7**	**44.6**
	MRCoR-T	6.2	6.7	9.6	12.5	38.2	39.4
	MRCoR-C	6.5	7	11.3	**16.4**	32.4	37.3
	MRFR	5.7	7.4	10.8	12	31.6	36.8
2003	MRCoR	5.4	**7.2**	**10.2**	9.7	**24.6**	**28.6**
	MRCoR-T	5.2	5.6	9.6	9.7	22.6	25.6
	MRCoR-C	5	4.8	9.8	**10.4**	21.7	24
	MRFR	**5.6**	4.8	9.7	10.2	22.5	27.4

MRCoRank outperforms FutureRank by at most 10% on the ranking of papers published in 2000. For author ranking shown in Figure 6, MutualRank is surprisingly no better than simply counting current citations. MRCoRank outperforms the baselines by at most 20% for author ranking.

Comparison with three variations of MRCoRank. To investigate whether and to what extent the time and content information can improve the performance, we conduct experiments to compare MRCoRank with MRCoR-T, MRCoR-C and MRFRank. The result is given in Table XII. The boldface figures denote the best results. One can see that in most cases the time and content information do help us get better rankings. It also shows that the results of 2000 and 2001 are much better that that of 2002 and 2003. This is mainly because we only use the available data before 2005 for ranking. Papers have not obtained sufficient citations, and authors have not published many papers in such a short time. One can also see that the performance of MRCoRank is almost consistently better than MRFRank. It implies that compared with our previous model, adding the venue information to the ranking model can further improve the performance on both paper ranking and author ranking.

Table XIII. Topics Discovered on Database Community

Topics	Spectral Clustering	LDA
1	model, expert, relationship, lda, garch, profile, factor, linear, uml	data, database, system, algorithm, performance, analysis, query
2	engin, stream, cluster, ensemble, gene, tempor, project, hierarchy, consensus	data, mining, experiment, clustering, unsupervised, classification, information, based
3	web, collect, predict, crawler, deep, browser, integry, directory, server	information, retrieval, search, web, semantic, document, text, algorithm, index
4	query, classify, revers, identify, algebra, uncertain, log, session, opinion	knowledge, system, reasoning, logic, representation, model, method, database
5	network, privacy, online, hoc, analysis, friend, publish, evolution, biology	language, knowledge, topic, query, structure, information, order, word, vector
6	keyword, xml, search, advertise, probabilist, popular, blog, behavior, strategy	mining, algorithm, learning, discovery, search, online, efficiency, detection, data
7	skyline, algorithm, space, greedy, set, link, spam, estimate, social, media	social, network, applied, analysis, graph, mining, Twitter, user, prediction
8	database, system, multi, label, scheme, answer, match, inconsistent, entity	efficient, large, dataset, system, analysis, result, show, time

5.8. Potentially Popular Topics Discovery by Spectral Clustering

An additional advantage of the proposed MRCoRank is that MRCoRank can also rank the future importance of text features. Based on these text features, we can further analyze the potentially popular research topics. In this subsection, we will show the effectiveness of MRCoRank in topic discovery by clustering the ranked words based on their co-occurrence.

By running MRCoRank, we can obtain the rankings of words and word-pairs. Based on the two rankings, we can construct a weighted graph as follows. A single word can be considered as a node in the graph. If words $word_i$ and $word_j$ form a word-pair in the word-pair ranking, we consider that there is a link between node $word_i$ and node $word_j$. The authority score of each word is assigned to be the weight of the corresponding node in the graph. Likewise, the authority score of each word-pair is assigned to be the weight of the corresponding edge in the graph. Then we apply spectral clustering [White and Smyth 2005] to cluster the nodes in the constructed graph. Each cluster can be naturally considered as a topic. Traditional topic-modeling-based clustering methods, such as PLSA [Hofmann 1999] and LDA [Blei et al. 2003], focus on static corpus; thus, they are not effective in discovering topics that are becoming popular.

Tables XIII and XIV show the discovered topics of our proposed cluster method and LDA model on two computer communities: database and artificial intelligence. We use the papers published before 2010 to obtain the rankings of words and word-pairs to construct the text feature graph. Tables XIII and XIV give the top eight popular research topics discovered by our method and LDA in each community. One can see that the topics discovered by our proposed spectral clustering method are relatively new and more specific, and the topics discovered by LDA are rather general. It is hard to identify some currently interesting and popular research topic based on the results of LDA. Taking the database community as an example, our method can discover some currently popular research topics, such as “stream cluster,” “network privacy,” “skyline algorithm,” and “lda model.”

6. CONCLUSION

While previous related works focus mainly on ranking the current importance of papers and authors, this article proposes an approach, MRCoRank, to predict the future influence of new publications and young researchers. MRCoRank integrates the available time, graphs, and rich text information into a unified framework to corank four types

Table XIV. Topics Discovered on Artificial Intelligence Community

topic ID	Spectral Clustering	LDA
1	human, leg, parallel, locate, bipe, configure, intelligence, sense, remote	problem, algorithm, optimal, solution, solve, constrain, heuristic, genetic
2	eye, detect, defect, track, anomaly, shadow, watermark, fraud, gaze, outlier	word, language, speech, system, recognition, sentences, grammar, character
3	qa, interact, question, segment, prior, predict, protein, cut, handwritten, number	learning, inference, machine, sampling, robot, move, feature
4	game, vote, rule, weight, equilibrium, sat, real, equilibria	training, learning, algorithm, data, statistics, optimization, model, sparse
5	belief, logic, motion, trajectory, coalit, knowledge, pattern, update, match	vision, image, process, dimension, decision, pattern, object, recognition
6	image, wavelet, compress, secret, share, minutia, authenty, fingerprint, feature	bayesian, reasoning, inference, constrain, logic, expert, system, fuzzy, model
7	synchron, grammar, bam, network, neural, delay, cohen, gross berg, exponential, stabil	speech, language, recognition, synthesis, translate, machine, text, interface
8	texture, classify, string, kernel, fisher, discrimine, multiclasses, binary, reproduce, nonlinear	artificial, intelligence, support, decision, system, knowledge, network, social

of entities, papers, authors, venues, and text features, simultaneously. Via a mutual reinforcement framework, we fuse the rich information of multientities and iteratively rank their future importance. On the ArnetMiner dataset, we empirically evaluate our approach against state-of-the-art methods, and the results show the effectiveness of our approach.

REFERENCES

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning* 3 (2003), 993–1022.

M. Bras-Amorós, J. Domingo-Ferrer, and V. Torra. 2010. A bibliometric index based on the collaboration distance between cited and citing authors. *Journal of Informetrics* 5, 2 (2010), 248–264.

P. Chen, H. Xie, S. Maslov, and S. Redner. 2007. Finding scientific gems with Google. *Journal of Informetrics* 1, 1 (2007), 8–15.

Y. Ding, E. Yan, A. Frazho, and J. Caverlee. 2009. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology* 60, 11 (2009), 2229–2243.

L. Egghe. 2006. Theory and practice of the g-index. In *Scientometrics*, Vol. 69. 131–152.

E. Garfield. 1972. Citation analysis as a tool in journal evaluation. *Science* 178, 4060 (1972), 471–479.

J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences*, Vol. 102. 16569–16572.

T. Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of 21st Annual International SIGIR Conference on Research and Development in Information Retrieval*.

X. R. Jiang, X. P. Sun, and H. Zhuge. 2012. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 714–723.

X. R. Jiang, X. P. Sun, and H. Zhuge. 2013. Graph-based algorithms for ranking researchers: Not all swans are white! *Scientometrics* 96 (2013), 743–759.

M. Kendall. 1938. A new measure of rank correlation. In *Biometrika*, Vol. 30. 81–89.

J. Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 91–101.

M. Li, X. B. Xue, and Z. H. Zhou. 2009. Exploiting multi-modal interactions: A unified framework. In *Proceedings of 21st International Joint Conference on Artificial Intelligence*. 1120–1126.

X. Li, B. Liu, and P. S. Yu. 2008. Time sensitive ranking with application to publication search. In *Proceedings of the 8th IEEE International Conference on Data Mining*. 893–898.

B. Liu, M. Q. Hu, and J. S. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International World Wide Web Conference*.

N. Ma, J. C. Guan, and Y. Zhao. 2008. Bringing pagerank to the citation analysis. *Information Processing and Management: An International Journal* 44, 2 (2008), 800–810.

S. Nerur, R. Sikora, G. Mangalaraj, and V. G. Balijepally. 2005. Assessing the relative influence of journals in a citation network. In *Communications of the ACM*, Vol. 48. 71–74.

M. K. P. Ng, X. T. Li, and Y. M. Ye. 2011. MultiRank: Co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1217–1225.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab* (1999).

H. Sayyadi and L. Getoor. 2009. Future rank: Ranking scientific articles by predicting their future pagerank. In *Proceedings of SIAM International Conference on Data Mining*. 533–544.

J. F. Si, A. Mukherjee, B. Liu, Q. Li, H. Y. Li, and X. T. Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Z. K. Silagadze. 2011. Citation entropy and research impact estimation. *Acta Physica Polonica B* B, 41 (2011), 2325–2333.

J. Tang, J. Zhang, L. M. Yao, J. Z. Li, L. Zhang, and Z. Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 990–998.

D. Walker, H. F. Xie, K. K. Yan, and S. Maslov. 2007. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics* 7 (2007), 6010–6019.

S. Z. Wang, X. Hu, P. S. Yu, and Z. J. Li. 2014. MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1246–1255.

S. Z. Wang, S. H. Xie, X. M. Zhang, P. S. Yu, and Z. J. Li. 2014. Future influence ranking of scientific literature. In *Proceedings of SIAM International Conference on Data Mining*. 749–757.

S. Z. Wang, Z. Yan, X. Hu, P. S. Yu, and Z. J. Li. 2015. Burst time prediction in cascades. In *Proceedings of 29th AAAI Conference on Artificial Intelligence*. 325–331.

Y. J. Wang, Y. H. Tong, and M. Zeng. 2013. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Proceedings of 27th AAAI Conference on Artificial Intelligence*. 933–939.

J. S. Weng, Y. X. Yao, E. Leonardi, and F. Lee. 2011. Event detection in twitter. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*. 401–408.

S. White and P. Smyth. 2005. A spectral clustering approach to finding communities in graphs. In *Proceedings of SIAM International Conference on Data Mining*. 274–285.

J. J. Yao, B. Cui, Y. X. Huang, and Y. H. Zhou. 2010. Detecting bursty events in collaborative tagging systems. In *Proceedings of IEEE 26th International Conference on Data Engineering*. 780–783.

M. Zhang, S. Feng, J. Tang, B. Ojokoh, and G. J. Liu. 2011. Co-ranking multiple entities in a heterogeneous network: Integrating temporal factor and users' bookmarks. In *Proceedings of the 13th International Conference on Asia-Pacific Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*. 202–211.

D. Zhou, S. A. Orshanskiy, H. Y. Zha, and C. L. Giles. 2007. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of IEEE 13th International Conference on Data Mining*. 739–744.

Received December 2014; revised October 2015; accepted December 2015