



Measuring academic entities' impact by content-based citation analysis in a heterogeneous academic network

Fang Zhang¹ · Shengli Wu^{1,2}

Received: 18 January 2021 / Accepted: 26 May 2021 / Published online: 23 June 2021
© The Author(s) 2021

Abstract

Evaluating the impact of papers, researchers and venues objectively is of great significance to academia and beyond. This may help researchers, research organizations, and government agencies in various ways, such as helping researchers find valuable papers and authoritative venues and helping research organizations identify good researchers. A few studies find that rather than treating citations equally, differentiating them is a promising way for impact evaluation of academic entities. However, most of those methods are metadata-based only and do not consider contents of cited and citing papers; while a few content-based methods are not sophisticated, and further improvement is possible. In this paper, we study the citation relationships between entities by content-based approaches. Especially, an ensemble learning method is used to classify citations into different strength types, and a word-embedding based method is used to estimate topical similarity of the citing and cited papers. A heterogeneous network is constructed with the weighted citation links and several other features. Based on the heterogeneous network that consists of three types of entities, we apply an iterative PageRank-like method to rank the impact of papers, authors and venues at the same time through mutual reinforcement. Experiments are conducted on an ACL dataset, and the results demonstrate that our method greatly outperforms state-of-the-art competitors in improving ranking effectiveness of papers, authors and venues, as well as in being robust against malicious manipulation of citations.

Keywords Scientific impact evaluation · Heterogeneous network · Content-based citation analysis · Citation strength · Topical similarity

Introduction

Due to the rapid development of science and technology, the total number of papers published in recent years has increased significantly. According to an STM report (Johnson et al., 2018), there were 33,100 peer-reviewed English journals in mid-2018, and over 3 million articles were published per year. The total number of publications and the number

✉ Shengli Wu
s.wu1@ulster.ac.uk

¹ School of Computer Science, Jiangsu University, Zhenjiang, China

² School of Computing, Ulster University, Belfast, UK

of journals have both grown steadily for over two centuries, at the rates of 3% and 3.5% per year, respectively. Facing such a huge number of publications, academia and other sectors of the society have become keen to find answers to the following questions: How can the importance of a research paper be measured? How can the performance of a researcher or a research organization be evaluated? It is necessary to have an objective evaluation system to measure the performance of papers, authors and venues.

For a long time, many researchers have tried various ways to evaluate the academic impact effectively. Citation count plays an important role in evaluating papers and authors. Based on citation count, many metrics, such as the h-index (Hirsch, 2005), the *g*-index (Egghe, 2006), the journal impact factor (Garfield, 2006), and others, have been proposed. These metrics are straightforward, but some factors, such as citation sources and co-authorship, are not considered. Heterogeneous academic networks, which include multiple types of entities including papers, authors, and venues, are very good a platform for academic performance evaluation, because all related information is available for us to exploit. Based on such networks, graph-based methods can be used (Jiang et al., 2016; Simkin & Roychowdhury, 2003; Zhang & Wu, 2020). For example, both SCImago Journal Rank (SJR) (González-Pereira et al., 2010, 2012) and the Eigenfactor score (Bergstrom, 2007) use PageRank-like algorithms (Brin & Page, 1998) to evaluate journals. MutualRank (Jiang et al., 2016) and Tri-Rank (Liu et al., 2014) rank papers, authors and venues simultaneously based on heterogeneous academic networks. These graph-based methods have some advantages for ranking academic entities due to their ability of leveraging structural information in academic networks and the mutual reinforcement relationship among papers, authors and venues.

Many existing graph-based ranking algorithms treat all citations as equally influential (Chakraborty & Narayanam, 2016; Zhu et al., 2015), without distinguishing that some of them may be more important than others. Such an approach may be questionable. Typically, for many papers, a small number of references play an important role (Chakraborty & Narayanam, 2016; Simkin & Roychowdhury, 2003; Wan & Liu, 2014), while most of the others do not have much impact (Teufel et al., 2006). In order to deal with such a problem, various aspects have been considered to weight citation links. For a given paper, we may consider many different aspects such as who cites the paper, where the citing paper is published, the time gap between two papers' publication, if it is a self-citation, and so on. We may also consider the topical similarity of the two papers or how the cited paper is related to the citing paper (referred to as citation strength in this paper). Different rationales are behind those aspects. For example, considering the venue that the citing paper is published, the citation is more valued if it is cited by a paper published in a prestigious venue than in an average venue. If it is a self-citation, it will get less credit than the others.

The primary goal of this paper is to investigate the middle to long-term impact of academic entities through a comprehensive framework (Kanellos et al., 2021). Especially we exploit some content-based features such as citation strength and topical similarity between the cited and citing papers, which are used to define weighted citation links. A heterogeneous network of papers, authors, and venues is built to reflect the relationships among them. Three types of entities are ranked at the same time through a PageRank-like algorithm with mutual reinforcement.

One possible problem with PageRank is it favors older papers than newer papers. This is referred to as the ranking bias (Jiang et al., 2016; Zhang et al., 2019a). It always takes time for a paper to be recognized in the community; a similar situation may also happen to authors. Therefore, a good evaluation system should be able to balance papers published at different time. In the same vein, we apply time-aware weights for all the papers involved.

Moreover, our framework includes a number of good features. In the heterogeneous network generated, seven types of relations are defined and supported. They are paper citation, author citation, venue citation, co-authorship, paper-author, paper-venue, and author-venue relations. For both authors and venues, their performance is evaluated on a yearly basis. Such a fine granularity enables us to catch the dynamics of the entities involved more precisely.

Citation manipulation (e.g., padded, swapped, and coerced citations) usually occurs in citations that do not contribute to the content of an article.¹ Because some government agencies rely heavily on impact factors to evaluate the performance of researchers and research organizations, there is evidence that various types of citation manipulation exist. For example, some scholars add authors to their research papers even those individuals contribute nothing to the research effort (Fong & Wilhite, 2017). Some journal editors suggest or request that authors cite papers in designated journals to inflate their citation counts (Fong & Wilhite, 2017; Foo, 2011). Peer reviewers may deliberately manipulate the peer-review process to boost their own citation counts (Chawla, 2019). Some scientists may self-citing extremely (Noorden & Chawla, 2019). Therefore, it is desirable to take this problem seriously into consideration when ranking academic entities. Citation manipulation (Bai et al., 2016; Chakraborty & Narayanam, 2016; Wan & Liu, 2014) is a problem that needs to be considered for academic entity ranking. As an extra benefit to the measures we apply, we believe that the proposed approach is robust and able to mitigate various kinds of citation manipulation problems (Bai et al., 2016; Chakraborty & Narayanam, 2016; Wan & Liu, 2014).

By consolidating all the measures above-mentioned, in this paper we propose a framework, WCCMR (Weighted Citation Count-based Multi-entity Ranking), to evaluate the impact of multiple entities. There are a number of contributions in this piece of work:

- 1 An ensemble learning method is used with three base classifiers to classify citations into five different categories. The fused results are better than that of all base classifiers, which represent the up-to-date technologies.
- 2 A word embedding-based method is used to measure topical similarity between the citing paper and the cited paper.
- 3 The above two content-based features are combined to define weighted citation links. To the best of our knowledge, we have not seen such a weighing scheme for citation before.
- 4 Apart from the weighted citation scheme, our framework has a number of good features: time-aware weighting, fine granularity for authors and venues, and seven types of relations among the same or different types of entities.
- 5 Experiments with the ACL (Association for Computational Linguistics Anthology Network) dataset (Radev et al., 2013) show that the proposed method outperforms other state-of-the-art methods in evaluating the effectiveness of papers, authors and venues, as well as in robustness against malicious manipulations.

The remainder of this paper is organized as follows: Sect. 2 presents related work on performance evaluation of academic entities, mainly by using various types of academic

¹ https://publicationethics.org/files/COPE_DD_A4_Citation_Manipulation_Jul19_SCREEN_AW2.pdf. Accessed 30 July 2020.

networks. Section 3 describes the framework proposed in this study. Section 4 presents the detailed experimental settings, procedures, and results. Some analysis of the experimental results is also given. Section 5 concludes the paper.

Related work

As an important task to the research community and beyond, evaluating scientific papers, authors and venues has been studied by many researchers for a long time. Citation count has been widely used and many citation-based metrics have been proposed (Jiang et al., 2016; Wang et al., 2016). For example, h-index (Hirsch, 2005) and g-index (Egghe, 2006) are used to measure researchers, the Impact Factor (IF) (Garfield, 1972), 5 year Impact Factor (5 year IF) (Pajić, 2015), and Source Normalized Impact per Paper (SNIP) (Moed, 2010; Waltman et al., 2013) are used to measure venues. These citation-based metrics are easy to understand and calculate. However, they have some crucial shortcomings. Firstly, many related metadata about any paper, such as its author(s) and venue, are ignored. This may have a negative effect on accuracy of the evaluation; Secondly, simple citation count lacks immunity to manipulation of citations. This is also an important issue that needs to be addressed.

As a remedy to some of the problems of using simple citation count, applying PageRank-like algorithms into academic networks has been investigated by quite a few researchers in recent years. For instance, the Eigenfactor score (Bergstrom, 2007) and SJR (González-Pereira et al., 2010, 2012) are used to evaluate journals. According to what type of information is used, we may divide those methods into two categories: metadata-based approach (time-aware weighting is a popular sub-category) and content-based approach.

Metadata-based approach has been investigated in (Yan & Ding, 2010; Zhang & Wu, 2018; Zhang et al., 2019a, b; Zhou et al., 2016) among others. To improve paper ranking performance and robustness against malicious manipulation, Zhou et al. (2016) proposed a weight assignment method for citation based on the ratio of common references between the citing and cited papers. Similar to Zhou et al. (2016), Zhang et al. (2019b) considered the reference similarity between the citing and cited papers. They also considered the topical similarity (calculated using titles and abstracts) between the two papers and combined them for weighting. Believing that immediate citations after publication is an indicator of good quality, some researchers allocated heavy weights to those papers that are cited shortly after publication (Yan & Ding, 2010; Zhang & Wu, 2018; Zhang et al., 2019a). For alleviating the ranking bias towards newly published papers, Walker et al. (2006) and Dunaiski et al. (2016) allocated heavier weights to newer papers, while Wang et al. (2019) considered the citations in the first 10 years of any paper since its publication and ignored the later ones. Self-citation, which is given a lighter weight than a “normal” citation, is investigated in (Bai et al., 2016).

Content-based approach has been investigated in (Chakraborty & Narayanam, 2016; Wan & Liu, 2014; Xu et al., 2014). Wan and Liu (2014) and Chakraborty and Narayanam (2016) classified citations into five categories of strength based on content analysis of the citing papers, and then assigned different weights for those citations accordingly. In Wan and Liu (2014), Support Vector Regression is used to estimate the strength of each citation. While in Chakraborty and Narayanam (2016), a graph-based semi-supervised model, GraLap, is used to estimate citation strength. In both cases, dozens of features, either metadata-based or content-based, are used in their model. Xu

et al. (2014) proposed a variant of PageRank in which a dynamic damping factor is used instead. At each paper node, its damping factor is decided by the topic freshness and publication age of the paper in question. Topic freshness per year is obtained by analyzing contents of all the papers in the dataset investigated.

To make full use of the information in academic networks and/or evaluate multiple entities at the same time, some researchers have proposed some PageRank variants by using various heterogeneous networks (Bai et al., 2020; Jiang et al., 2016; Liu et al., 2014; Meng & Kennedy, 2013; Yan et al., 2011; Yang et al., 2020; Yang et al., 2020; Zhang & Wu, 2018, 2020; Zhang et al., 2018, 2019a; Zhao et al., 2019; Zhou et al., 2021). Yan et al. (2011) proposed an indicator, P-Rank, to score papers. For each citation, the impact of the citing paper, the citing authors and the citing journal are considered at the same time. Differentiating each venue year by year, Zhang and Wu (2018) proposed a ranking method, MR-Rank, to evaluate papers and venues simultaneously. Meng and Kennedy (2013) proposed a method, Co-Ranking, for ranking papers and authors. Tri-Rank, proposed by Liu et al. (2014), can rank authors, papers, and journals simultaneously. Especially, Tri-Rank considers the ordering of authors and self-citation problems. Jiang et al. (2016) proposed a ranking model MutualRank, which is a modified version of randomized HITS for ranking papers, authors and venues simultaneously. Zhang et al. (2018) proposed a classification-based method to predict authors' influence. They firstly classified authors into different types according to their citation dynamics and then applied the modified random walk algorithms in a heterogeneous temporal academic network for prediction. Based on a heterogeneous network that includes both paper citation and paper-author relations, Zhao et al. (2019) measured the influence of authors on two large data sets, and one of which included 500 million citation links. By assigning weight to the links of citation network and authorship network according to the citation relevance and author contribution, Zhang et al. (2019a) ranked scientific papers by integrating the impact of papers, authors, venues and time awareness. By differentiating each venue and researcher on a yearly basis, Zhang and Wu (2020) proposed a framework, WMR-Rank, to predict the future influence of entities including papers, authors, and venues simultaneously. For balanced treatment of old and new papers, they considered both the publication age and recent citations of all the papers involved at the same time. Bai et al. (2020) measured the impact of institutes and papers simultaneously based on the heterogeneous institution-citation network. Based on a heterogeneous network that including co-authorship, author-paper and paper citation relation, Zhou et al. (2021) proposed an improved random walk algorithm to recommend research collaborators. Especially, they considered both time awareness and topic similarity. Similar to Zhou et al. (2021), Yang et al. (2020) recommend researcher collaborators by using an improved walking algorithm. A heterogeneous network by combing co-author network and institution network is used.

It is likely that the work in Wan and Liu (2014) and Chakraborty and Narayanam (2016) are the most relevant to our work in this paper, however, there are considerable differences between our work in this paper and either of them. First, we use an ensemble learning method for citation strength estimation and the results show that it is more effective than the methods used in those two papers. Besides, topic similarity is also included for determining the weighting of citation link. This is not included in either Wan and Liu (2014) and Chakraborty and Narayanam (2016). Lastly, a sophisticated network with multiple types of entities is built and used in this paper to evaluate their impact at the same. As we will see later in the experimental part, it works with other components to achieve very good results.

Table 1 Some symbols used in this paper and their meanings

Symbol	Description
P	Vector indicting the scores of papers for their ranking
A	Vector indicting the scores of authors for their ranking
\bar{A}	Vector indicating the scores of authors in a given year
V	Vector indicting the scores of venues for their ranking
S_p	Set of papers in the entire collection
S_A	Set of authors in the entire collection
S_V	Set of venues in the entire collection
$S_p(a)$	Set of papers of author a
$S_A(p)$	Set of authors of paper p
$S_p(v)$	Set of papers published in venue v
$ S_p $	Number of papers in S_p
$ S_A $	Number of authors in S_A
$ S_V $	Number of venues in S_V
W_{PP}	$A S_p \times S_p $ matrix indicating the paper citation relation (Eq. 1)
$W_{C\bar{A}}$	$A S_A \times S_A $ matrix indicating the author citation relation (Eq. 4)
$W_{CO\bar{A}}$	$A S_A \times S_A $ matrix indicating the coauthor relation (Eq. 7)
W_{VV}	$A S_V \times S_V $ matrix indicating the venue citation relation (Eq. 8)
W_{PA}	$A S_p \times S_A $ matrix indicating the paper-author relation (Eq. 9)
W_{PV}	$A S_p \times S_V $ matrix indicating the paper-venue relation (Eq. 10)
W_{AV}	$A S_A \times S_V $ matrix indicating the author-venue relation (Eq. 11)
W_{RP}	$A S_p \times S_p $ matrix indicating the recent citation bonus of papers (Eq. 14)
$W_{R\bar{A}}$	$A S_A \times S_A $ matrix indicating the recent citation bonus of authors (Eq. 15)
$W_{\bar{A}\bar{A}}$	$A S_A \times S_A $ matrix connecting an author with herself in each year (Eq. 16)
$W_{T\bar{A}}$	$A S_A \times S_A $ matrix indicating the time-awareness weight (Eq. 17)
$W_{\bar{V}V}$	$A S_V \times S_V $ matrix indicating the performance score of venues in past t_v years (Eq. 18)

The proposed method

In this section, we introduce all the components required and then present the multi-entity ranking algorithm. The Symbols used in this paper and their meanings are summarized in Table 1.

Citation strength and topical similarity

When researchers write papers, they usually need to cite other papers for various reasons, such as pointing to a baseline method for comparison, applying a proposed method or making some improvement of it, referring to the definition of an evaluation metric, as evidence of supporting a point of view, and so on. Considering all those different purposes of citation, some of which may be more important than some others. Therefore, in line with the work of Liu (2014) and Chakraborty and Narayanam (2016), we define five levels of citation strength as follows.

1. *Level 1* The cited reference has the lowest importance to the citing paper. It is related to the citing paper casually. It usually follows words like “such as”, “for example”, “note” in the text, and can be removed or replaced without hurting the competence of the references.
2. *Level 2* The cited reference is related to the citing paper to some extent. For example, it is cited to support a point of view or to introduce the development of research fields related to the citing paper. It is usually mentioned together with other references and appears in parts such as “introduction”, “related work”, or “conclusion and future work”.
3. *Level 3* The cited reference is important and related to the citing paper. For example, it may serve as a baseline method. It is usually mentioned several times in the paper with long citation sentences and may appear in more than one part of the paper.
4. *Level 4* The cited reference is very important to the citing paper. It is usually mentioned separately in one or more sentences and appears in the methodology section, such as algorithms or models used in the citing paper. It can be an integral part of the model proposed in the paper.
5. *Level 5* The cited reference is extremely important and highly related to the citing paper. For example, the citing paper makes an improvement based on the cited reference or borrows its main idea from the cited reference. It is usually mentioned multiple times, sometimes following “this method is influenced by”, “we extend”, etc., and very likely appears in multiple parts of the paper such as “introduction”, “related work”, “method”, “experiment”, “discussion”, or “conclusion”.

Citation topical similarity refers to the topical similarity between the cited paper and the citing paper. It is independent from citation strength. A word-embedding based approach is used for this. It is also a good indicator of proper citation. The higher the similarity is between the citing paper and the cited paper, the lower the likelihood that the cited paper is artificially manipulated. A linear combination of them is set to be the weight of the citation. See Eq. (1) later in this paper. Based on that, a heterogeneous network can be built with the desirable properties. We consider that differentiating citations instead of taking simple citation counts may produce more reliable evaluation results.

A heterogeneous academic network

A heterogeneous academic network is composed of nodes and edges. Each node represents an entity and each edge between two nodes represents the relation between the two entities. There are three types of nodes: papers, authors, and venues, and seven types of relations: paper citation, paper-author relation, paper-venue relation, coauthor relation, author citation, author-venue relation and venue citation. A suitable weight needs to be assigned to each of the edges involved. In the following we discuss these seven types of relations one by one, in which weight assignment for each type of edges is the key issue.

Paper citation relation

A paper citation relation exists when one paper cites another paper. If paper p_j cites paper p_i , the weight is defined as

$$W_{PP}(p_i, p_j) = \begin{cases} \text{strength}(p_i, p_j) + \text{sim}(p_i, p_j) & p_i \leftarrow p_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\text{strength}(p_i, p_j)$ and $\text{sim}(p_i, p_j)$ are the citation strength and topical similarity between p_i and p_j , respectively. $p_i \leftarrow p_j$ denotes that paper p_i is cited by paper p_j . It is required that both $\text{strength}(p_i, p_j)$ and $\text{sim}(p_i, p_j)$ are defined in the same range. Otherwise, normalization may be required to make them comparable.

Author citation relation

Through paper citation, we can set up an indirect relation of author citation. paper p_i is cited by paper p_j , \bar{a}_m is the only author or one of the authors of p_i , and \bar{a}_n is the only author or one of the authors of p_j , then \bar{a}_m is cited by \bar{a}_n ($\bar{a}_m \leftarrow \bar{a}_n$). The same as in Zhang and Wu (2020), we differentiate each author year by year and allocate the credit that author \bar{a}_m who published paper p_i in year $t_{\bar{a}_m}$, obtains from \bar{a}_n who published paper p_j in year $t_{\bar{a}_n}$, through paper citation $p_i \leftarrow p_j$ as

$$W_{CA_raw}(\bar{a}_m, \bar{a}_n, p_i, p_j) = \frac{1}{\text{order}(\bar{a}_m, p_i) \times \text{order}(\bar{a}_n, p_j)} \tag{2}$$

where $\text{order}(a, p)$ is the position of author a in paper p . Normalization is required for all the authors involved.

$$W_{CA}(\bar{a}_m, \bar{a}_n, p_i, p_j) = W_{PP}(p_i, p_j) \frac{W_{CA_raw}(\bar{a}_m, \bar{a}_n, p_i, p_j)}{\sum_{\substack{p_i \leftarrow p_j \\ \bar{a}_k \in S_A(p_i) \\ \bar{a}_l \in S_A(p_j)}} W_{CA_raw}(\bar{a}_k, \bar{a}_l, p_i, p_j)} \tag{3}$$

where $S_A(p)$ is the set of all the authors of paper p .

An author \bar{a}_n may cite another author \bar{a}_m multiple times. The total credit that \bar{a}_m in year $t_{\bar{a}_m}$ obtains from \bar{a}_n in year $t_{\bar{a}_n}$ is the summation of all the papers involved.

$$W_{CA}(\bar{a}_m, \bar{a}_n) = \sum_{\substack{p_i \in S_P(\bar{a}_m) \\ p_j \in S_P(\bar{a}_n) \\ p_i \leftarrow p_j}} W_{CA}(\bar{a}_m, \bar{a}_n, p_i, p_j) \tag{4}$$

where $S_P(a)$ is the set of papers written by author a .

Coauthorship relation

A coauthorship relation exists in the network if two or more author nodes connect to the same paper node. Any author obtains certain credit from all other authors if they write a paper together. The credit that \bar{a}_i who has published papers in year $t_{\bar{a}_i}$ obtains from her coauthor \bar{a}_j through paper p is defined as

$$W_{CO\bar{A}_{raw}}(\bar{a}_i, \bar{a}_j, p) = \frac{1}{\text{order}(\bar{a}_i, p) \times \text{order}(\bar{a}_j, p)} \tag{5}$$

which needs to be normalized. We have

$$W_{CO\bar{A}}(\bar{a}_i, \bar{a}_j, p) = \frac{W_{CO\bar{A}_{raw}}(\bar{a}_i, \bar{a}_j, p)}{\sum_{\bar{a}_k, \bar{a}_l \in S_\Lambda(p)} W_{CO\bar{A}_{raw}}(\bar{a}_k, \bar{a}_l, p)} \tag{6}$$

Two authors may co-write more than one paper. Hence, the credit that \bar{a}_i in year $t_{\bar{a}_i}$ obtains from \bar{a}_j over all co-authored papers is

$$W_{CO\bar{A}}(\bar{a}_i, \bar{a}_j) = \sum_{\substack{p \in S_p(\bar{a}_i) \\ p \in S_p(\bar{a}_j)}} W_{CO\bar{A}}(\bar{a}_i, \bar{a}_j, p) \tag{7}$$

where $S_p(\bar{a}_i)$ denotes all the papers written by \bar{a}_i .

Venue citation relation

Similar to author citation, we may define venue citation. For venues v_i and v_j , if $v_i \leftarrow v_j$, the weight between v_i and v_j can be denoted as

$$W_{VV}(v_i, v_j) = \sum_{\substack{p_k \leftarrow p_l \\ p_k \in S_p(v_i) \\ p_l \in S_p(v_j)}} W_{PP}(p_k, p_l) \tag{8}$$

Paper-author relation

Paper coauthorship happens very often. However, for one paper written by a group of coauthors, their contributions to the paper are differentiated by their ordered positions (Abbas, 2011; Du & Tang, 2013; Egghe et al., 2000; Stallings et al., 2013). More specifically, we adopt a geometric counting approach (Egghe et al., 2000) for the paper-author relation. Suppose author a_i is in the R th position among all T coauthors in paper p_j ; then, the amount of credit that author a_i and paper p_j obtain from each other is as follows:

$$W_{AP}(a_i, p_j) = W_{PA}(p_j, a_i) = \frac{2^{T-R}}{2^T - 1} \tag{9}$$

Paper-venue relation

If paper p_i is published in venue v_j , then there is an edge between paper p_i and venue v_j ; thus, paper p_i and venue v_j get credit from each other. We let

$$W_{VP}(v_j, p_i) = W_{PV}(p_i, v_j) = \begin{cases} 1 & p_i \in S_p(v_j) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Author-venue relation

If author a_i publishes more than one paper in venue v_j , then the credit that a_i obtains from v_j is the sum of the credit she obtains from all the papers published in v_j . The same is true for the credit v_j obtains from a_i .

$$W_{AV}(a_i, v_j) = W_{VA}(v_j, a_i) = \sum_{\substack{p_k \in S_P(a_i) \\ p_k \in S_P(v_j)}} W_{AP}(a_i, p_k) \quad (11)$$

Recent citation bonus

An entity (paper or author) obtains a score from a citation and its final score is the sum of these individual scores. In order to mitigate the ranking bias toward old papers (Jiang et al., 2016) and treat all the papers in a balanced way, it is necessary to consider the recent citations of entities including papers and authors. Therefore, besides the normal scores, an entity obtains an extra bonus if the citation is very close to the evaluation year.

For an entity e_i , assume that e_i has been cited in the most recent N years (including the evaluation year), and the evaluation year is t_{evaluate} . A bonus is given to entity e_i as

$$\text{RCB}(e_i) = \sum_{e_j \leftarrow e_i} \text{score}(e_j) \times W(e_i, e_j) \times f(t_j) \quad (12)$$

where $\text{score}(e_j)$ is the score of e_j that is calculated based on some other aspects of the entity, $W(e_i, e_j)$ is the weight between e_i and e_j , $f(t_j)$ is a time-related function.

$$f(t_j) = \begin{cases} \theta^{t_{\text{evaluate}} - t_j} & t_{\text{evaluate}} - t_j \leq N \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where θ is a parameter. In this paper, we set $\theta=0.8$ and $N=5$. $W(e_i, e_j) \times f(t_j)$ is the bonus weight of entities.

For papers, the bonus weight W_{RP} is defined as

$$W_{RP}(p_i, p_j) = W_{PP}(p_i, p_j) \times f(t_j) \quad (14)$$

For authors, the bonus weight W_{RA} is defined as

$$W_{RA}(\bar{a}_i, \bar{a}_j) = W_{CA}(\bar{a}_i, \bar{a}_j) \times f(t_j) \quad (15)$$

Self-connections between same type of entities

In this framework, both authors and venues may be considered as a whole or on a yearly basis. Therefore, we need to connect them in some situations. For example, for an author $a_j \in A$, there are a group of $\bar{a}_i \in \bar{A}$ (for $1 \leq i \leq n$), both a_j and \bar{a}_i refer to the same author. Each \bar{a}_i refers to a_j in a specific year. $W_{AA}(\bar{a}_i, a_j)$ is defined as

$$W_{\bar{A}A}(\bar{a}_i, a_j) = \begin{cases} 1 & \text{if } \bar{a}_i \text{ and } a_j \text{ is the same author} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

The second one is to set different weights for papers published in different years.

$$W_{T\bar{A}}(a_i, \bar{a}_j) = \begin{cases} e^{\mu(t_{\bar{a}_j} - t_{\text{evaluate}})} & \text{if } \bar{a}_i \text{ and } a_j \text{ is the same author} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

where μ is a parameter, $t_{\bar{a}_j}$ is the year at which \bar{a}_j is published.

Venues are considered on a yearly basis. However, there is a need to consider its previous performance for t_v years. Suppose v_i and v_j are the same conference but held in different years, v_i is held later than v_j but within t_v years, the corresponding weight is defined as

$$W_{V_V}(v_i, v_j) = \begin{cases} \frac{1}{t_v + 1} & v_j \text{ and } v_i \text{ satisfy the condition} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The WCCMR method

The proposed method, WCCMR, works with the abovementioned heterogeneous academic network. After setting initial values for all the entities, an iterative process is applied to them, and at each step every entity obtains an updated score. Note that all the entities involved affect each other and all the scores converge after enough iterations. The algorithm stops when a threshold ϵ for the difference between two consecutive iterations is satisfied. Algorithm 1 gives the details of the proposed method.

Initially, the rank vector of papers P , authors A (without considering the time), and venues V are set to $I_P/|V_P|$, $I_A/|V_A|$, and $I_V/|V_P|$. I_P , I_A and I_V are unit vectors, and $|V_P|$, $|V_A|$ and $|V_V|$ are the number of papers, authors and venues.

The main part of the algorithm is included in a while loop. Inside the loop (lines 1–13), the scores for all the nodes involved are updated. All papers’ new scores are calculated in lines 3–4. Four factors are considered: authors (line 3), venues (line 3), citations (line 4), and recent citation bonus (line 4). All authors’ new scores are calculated in lines 5–7. Five factors are considered: published papers (line 5), coauthors to the published papers (line 5), the venues in which the papers are published (line 5), author citations (line 6), and recent citation bonus (line 6). Finally, we sum up all the yearly scores by using a time function to obtain the total score for each author (line 7). All venues’ new scores are calculated in line 8–9. Three factors are considered: published papers (line 8), authors (line 8), and venue citations (line 9). Although multiple types of entities are involved in the algorithm, it still converges quite quickly. For example, with the dataset used in this study and ϵ set to $1e-6$., the algorithm stops after 13 iterations.

Algorithm 1: WCCMR

Input: node sets $V_P, V_A, V_{A'}$ and V_V ; weight matrices $W_{PP}, W_{PA}, W_{PV}, W_{C\bar{A}}, W_{CO\bar{A}}, W_{AV}, W_{VV}, W_{\bar{V}V}, W_{\bar{A}A}, W_{T\bar{A}}, W_{RP},$ and $W_{R\bar{A}}$, and parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \lambda, \theta, \mu, N,$ and ε

Output: \mathbf{P}, \mathbf{A} and \mathbf{V} which store scores for papers, authors and venues, respectively

Initialization: $\mathbf{P} \leftarrow \frac{\mathbf{I}_P}{|V_P|}, \mathbf{A} \leftarrow \frac{\mathbf{I}_A}{|V_A|}, \mathbf{V} \leftarrow \frac{\mathbf{I}_V}{|V_V|}, r=0, \Delta=1000 // \mathbf{I}_P, \mathbf{I}_A,$ and \mathbf{I}_V are unit vectors

01 **While** $\Delta > \varepsilon$ **Do**

02 $\tilde{\mathbf{V}}^{r+1} = W_{\bar{V}V} \mathbf{V}^r$ // Calculate past performance of venues

//calculates scores for papers by considering authors and past performance of their venues

03 $\mathbf{temp} = \alpha_1 W_{PA} \mathbf{A}^r + (1 - \alpha_1) W_{PV} \tilde{\mathbf{V}}^{r+1}$

//update scores for papers by considering both citation and recent citation bonus

04 $\mathbf{P}^{r+1} = [\lambda W_{PP} + (1 - \lambda) W_{RP}] \mathbf{temp}$

// calculates scores for authors by considering their papers, coauthors and venues

05 $\mathbf{temp} = \alpha_2 W_{PA}^T \mathbf{P}^r + \alpha_3 W_{CO\bar{A}} (W_{\bar{A}A} \mathbf{A}^r) + (1 - \alpha_2 - \alpha_3) W_{AV} \tilde{\mathbf{V}}^{r+1}$

//update scores for authors by considering citation and recent citation bonus

06 $\bar{\mathbf{A}}^{r+1} = [\lambda W_{C\bar{A}} + (1 - \lambda) W_{R\bar{A}}] \mathbf{temp}$

07 $\mathbf{A}^{r+1} = W_{T\bar{A}} \bar{\mathbf{A}}^{r+1}$ // sum up the yearly scores by using a time-related function

// calculates scores for venues by considering the papers and authors involved

08 $\mathbf{temp} = [\alpha_4 W_{PV}^T \mathbf{P}^r + (1 - \alpha_4) W_{AV}^T \mathbf{A}^r] / |S_P|$

09 $\mathbf{V}^{r+1} = \lambda W_{VV} \mathbf{temp} + (1 - \lambda) \mathbf{I}_V$ //update scores for venues considering their citations

10 Normalize \mathbf{P}, \mathbf{A} and \mathbf{V}

11 $\Delta = ||\mathbf{P}^{r+1} - \mathbf{P}^r||_1 + ||\mathbf{A}^{r+1} - \mathbf{A}^r||_1 + ||\mathbf{V}^{r+1} - \mathbf{V}^r||_1$

12 $r=r+1$

13 **End Do**

Table 2 Statistical information of experimental data sets

	Number
Number of papers	13, 591
Number of authors (considering year)	23, 161
Number of authors (without considering year)	10, 140
Number of venues (considering year)	437
Number of venues (without considering year)	248
Number of paper citation links	71, 486
Number of author citation links (considering year)	381, 243
Number of author citation links (without considering year)	254, 323
Number of coauthor links (considering year)	60, 503
Number of coauthor links (without considering year)	46, 871
Number of venue citation links (considering year)	18, 118
Number of venue citation links (without considering year)	5 455
Average number of citations of each paper	5.26

Experimental setting

Dataset

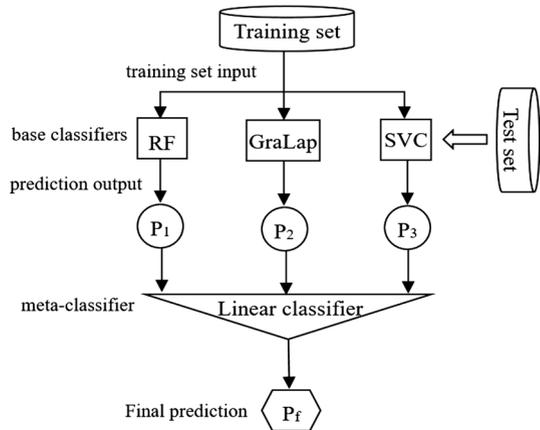
In this experiment, we use the ACL Anthology Network dataset² (AAN) (Radev et al., 2013), which is constructed from papers published in natural language processing venues (including journals, conferences and workshops from 1965 to 2011).³ We choose AAN because it provides both citations and full text for almost all the papers involved.

In order to make it suitable for the experiment, the dataset is pre-processed as follows. First, those papers that neither cite any other papers nor are cited by any other papers are removed, because they have no impact to the investigation in this paper. Those papers that have no full text are also removed, because we need full text for citation strength analysis and estimation. Second, any joint conferences are considered to have dual identity. For example, COLING-ACL'2006 is a joint conference of COLING and ACL. Third, in addition to regular papers, many conferences publish short papers, student papers, demos, posters, tutorials, etc. Usually, the quality of non-regular papers is not as good as that of regular papers. Therefore, we let all regular papers remain in the main conference while putting all non-regular papers into its companion, a separate venue. Finally, for those papers with more than 5 authors, we retained the first five authors and ignored the rest. After above-mentioned pre-processing, 13,591 papers remain with an average of 5.26 references for each of them, 10,140 authors and 248 venues without considering time, or 437 venues if taking each venue per year as a separate entity. Table 2 shows the general statistics of the dataset.

² See <http://clair.eecs.umich.edu/aan/index.php>.

³ Note that the dataset we use does not include papers published in 2011, just as in Jiang et al. (2016).

Fig. 1 The major processes in stacking classification



Calculating citation strength and topical similarity

Machine learning methods are good options for estimating citation strength because they have been very successful in many such applications. Stacking technique can combine classifiers via a meta-classifier to achieve better performance. In this study, we classify the citation strength by using the stacking technique with the features used in Chakraborty and Narayanam (2016). Random Forest (RF), Support Vector Classifier (SVC) and GraLap (Chakraborty & Narayanam, 2016) are selected as base classifiers because they are very good and represent up-to-date technology. Figure 1 shows the major steps involved in a meta-classifier. First a training data set is required to training base models and the meta-model as well. Then the trained model can be used to classify instances in the test set.

First, we select a group of 96 papers from the whole data set randomly. From them we get 2735 valid references whose full texts are available in the data set. By using the Parscit package (Councill et al., 2008) plus a few hand-coded rules, we extracted 4993 citation sentences and sections in which the sentences locate. Such information along with the original papers are provided to a group of 15 annotators, all of which are graduate research students in computer science in our school. Among all 2735 papers, 215 are annotated at level 1, 2046 are at level 2, 287 are at level 3, 142 are at level 4, and 45 are at level 5.

Then as in Chakraborty and Narayanam (2016) and Wan and Liu (2014), we extracted citation features such as the number of occurrences, sections in which it appears, similarity between the citing paper and cited paper, and others for all 2735 citing papers. They are divided into five groups, each of which includes one fifth of the papers at each individual level. This was done by running a random selection process to the papers at each level separately.

A five-fold cross-validation is carried out to validate the performance of the stacking approach. We find that classification of the instances at level 5 are the least accurate, while level 2 instances reaches the highest classification accuracy of more than 0.8. Note that level 2 has the largest number of instances while level 5 has the least number of instances. One possible explanation is: for level 2 instances, we have enough instances for the base classifiers and the stacking method to learn a good model. In contrast for level 5 instances, they are not enough. Table 3 shows its performance with two other approaches, SVR (Support Vector Regression) (Wan & Liu, 2014), and GraLap

Table 3 Performance comparison of three citation strength estimation methods

Method	MSE	F1	Accuracy
SVR	0.586	0.632	0.720
GraLap	0.521	0.662	0.748
Stacking	0.498	0.705	0.776

Table 4 Statistical information of the gold standard papers

Number of recommendations	2	3	4	5	6	7	8	9	10	Total
Number of gold standard papers	63	19	7	1	1	0	0	1	1	93

(Chakraborty & Narayanam, 2016). Note that SVR is slightly different from SVC. Both use support vector machine but treat the same problem as either a classification problem or a regression problem. We can see that the stacking classifier is slightly better than the two other methods when any of the three measures are used for evaluation.

For topical similarity, we extract the title and abstract of each paper and calculate the topic similarity based on word2vec after performing stemming. In the experiment, the dimension of the word vector is set to 200, and the context window is set to 5.

Ranking benchmarks

For papers, rather than calculating citation count of each paper, we consider that experts’ opinion is a more authoritative measure to decide the impact of papers in the scientific community. Therefore, in this article, we use the gold standard papers provided in Jiang et al. (2016). A collection of gold standard papers, named GoldP, is assembled as recommended papers from the reading lists of graduate-level courses in natural language processing or computational linguistics and the reference lists of two best-selling natural language processing textbooks. Only those papers taken from the AAN dataset with at least two recommendations are selected. In total, 93 papers are selected in GoldP. The statistical information of those selected papers is shown in Table 4.

In the same vein as gold standard papers, we use WRT (weighted recommendation times) to measure the influence of authors. The influence score of author a_i is defined as

$$WRT(a_i) = \sum_{p_j \in A_P(a_i) \& p_j \in GoldP} W_{AP}(a_i, p_j) \times RT(p_j) \tag{19}$$

where $RT(p_j)$ is the number of recommendations that paper p_j receives and $W_{AP}(a_i, p_j)$ is related to the ordering position of the author in question. See Eq. (12) in the “Paper-author relation” section for its definition of $W_{AP}(a_i, p_j)$. The final score that a_i obtains, $WRT(a_i)$, is the sum of the scores of all the papers in GoldP written by a_i . We consider this measure to be better than the citation count for authors because the inflationary effect can be mitigated. All the authors are regarded as influential authors (GoldA) if he/she wrote one or more gold standard papers. In this way, we obtain 149 authors in total.

Table 5 Statistical information of the gold standard venue collection

Number of recommended papers	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
Number of gold standard venues	24	5	8	3	8	1	2	0	0	0	0	1	1	1	0	0	1	55

For any venue, if it has two or more recommended papers in GoldP, then we set it as a recommended venue, GoldV. It includes 55 venues in total. The statistical information of GoldV is shown in Table 5.

The influence score of venue v_i is defined as

$$InS(v_i) = \sum_{p_i \in V_p(v_i) \& p_i \in GoldP} RT(p_i) \quad (20)$$

It summarizes the recommendations received by all the papers in the venue.

Evaluation metrics

We use two evaluation metrics: precision at a given ranking level and a modified version of NDCG (Jiang et al., 2016). They are used to evaluate the effectiveness of a ranked list of entities $E = \{e_1, e_2, \dots, e_n\}$.

Precision $P@K$ is defined as

$$P@K = \frac{\sum_{i=1}^K inf(e_i)}{K} \quad (21)$$

where $inf(e_i)$ takes binary values of 0 or 1. If e_i is an influential entity, then $inf(e_i)$ is 1, otherwise, $inf(e_i)$ is 0.

For a number of entities, the best ranking must exist, and it ranks all the entities in descending order of a given metric values. A group of papers can be ranked according to the times of recommendation received. WRT scores and number of recommended papers can be used for author and venue ranking, respectively. For a ranked list of entities $E = \{e_1, e_2, \dots, e_K\}$, assume that its corresponding best ranking list is $E' = \{e'_1, e'_2, \dots, e'_K\}$, we let $credit()$ denote the metric value of entity e_k obtain, and $best_credit()$ the metric value of entity e'_k obtain. NDCG@K is defined as

$$NDCG@K = \frac{\sum_{k=1}^K \frac{credit(e_k)}{\log_2(k+1)}}{\sum_{k=1}^K \frac{best_credit(e'_k)}{\log_2(k+1)}} \quad (22)$$

In Eq. (22), the top-ranked entities are given a weight of 1, then the weights decrease with rank by a factor $1/\log_2(k+1)$.

Methods for comparison

The ranking algorithms used for comparison are as follows:

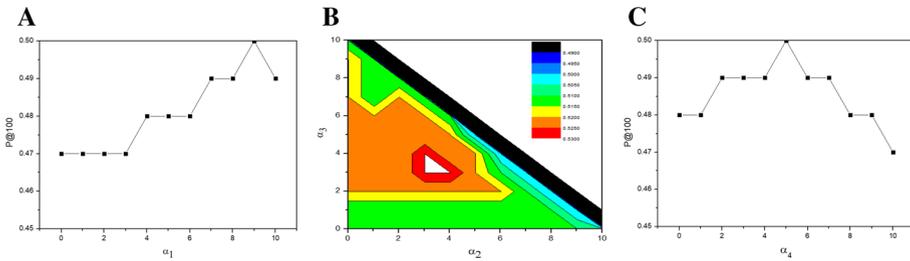


Fig. 2 Effect of different parameter values on ranking performance. **a** Effect of α_1 on papers. **b** Effect of α_2 and α_3 on authors. **c** Effect of α_4 on venues

1. Citation Count (CC). It is widely used to assess the influence of papers because it is single-valued and easy to understand (Zhu et al., 2015).
2. SVR-based Weighted Citation Count (WCC-SVR). It provides each citation with a citation strength value calculated by SVR (Wan & Liu, 2014).
3. GraLap-based Weighted Citation Count (WCC-GraLap). It provides each citation with a citation strength value calculated by GraLap (Chakraborty & Narayanam, 2016).
4. MutualRank (MR). A state-of-the-art method that ranks papers, authors and venues simultaneously in heterogeneous networks (Jiang et al., 2016).
5. Tri-Rank (Tri). Similar to MutualRank, Tri-Rank also ranks papers, authors and venues simultaneously in heterogeneous networks (Liu et al., 2014).
6. PageRank with SVR-based network (PR-SVR). The PageRank algorithm runs over a modified citation network in which each citation has a specific weight calculated by SVR (Wan & Liu, 2014).
7. PageRank with GraLap-based network (PR-GraLap). The PageRank algorithm runs over a modified citation network in which each citation has a specific weight calculated by GraLap (Chakraborty & Narayanam, 2016).
8. WCCMR. The method proposed in this paper (see Algorithm 1).

Parameter setting

There are five parameters in the proposed ranking model: α_1 , α_2 , α_3 , α_4 and ϵ . We set ϵ to $1e-6$. For α_1 , α_2 , α_3 and α_4 , we first set an intuitively reasonable value for each parameter: $\alpha_1=0.50$, $\alpha_2 = \alpha_3=0.33$, and $\alpha_4=0.50$. Then, fix three of them and let the remaining one vary to see its effect, and Fig. 2 shows the results (P@100 is used for performance evaluation).

From Fig. 2a, one can see that paper evaluation performance is quite stable when α_1 is in the range of 0.00 and 1.00. The best performance is achieved when $\alpha_1=0.90$. Similarly, from Fig. 2b, c we can see that $\alpha_2=0.35$, $\alpha_3=0.35$, and $\alpha_4=0.5$ are also good for these parameters.

Note that the parameters of α_1 and $(1-\alpha_1)$ are used to adjust the relative weights of authors and venues. A larger α value does not necessarily mean that authors are more important than venues because these two components are not directly comparable. α_1 partially serves as a normalization measure. We find the same conclusion for the other parameters α_2 , α_3 and α_4 .

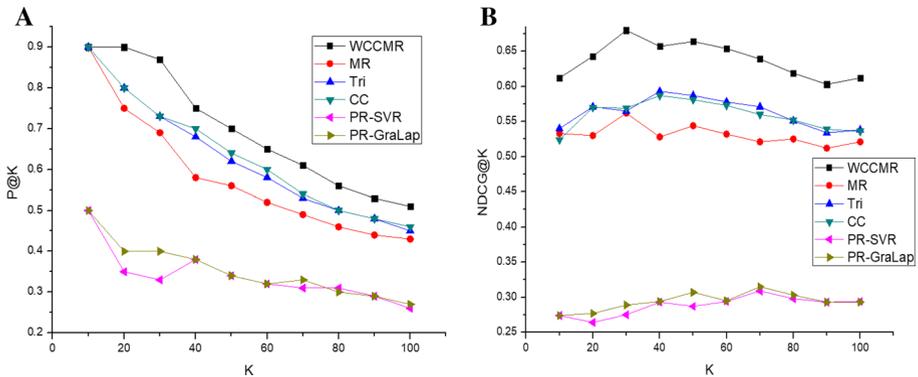


Fig. 3 Effectiveness of different algorithms for ranking papers. **a** Measured by P@K. **b** Measured by NDCG@K

Ranking performance

In this section, we present the evaluation results of the proposed algorithm, along with those of a group of state-of-the-art baseline methods.

Ranking effectiveness for papers

We first study paper ranking effectiveness of the proposed algorithm. Figure 3 shows the effectiveness curves of the different algorithms for ranking papers measured by P@K and NDCG@K. We can see that the proposed method, WCCMR, constantly outperforms all the other methods when either P@K or NDCG@K is used. Tri and CC are close. They are not as good as WCCMR but better than the others. It is also noticeable that the curves of PR-SVR and PR-GraLap are always very close. This is not surprising because both run PageRank. The difference between them is the way of setting citation weights in the heterogeneous network.

To investigate the properties of all the methods involved for top-ranked papers, we list the top 20 papers returned by WCCMR and its competitors in Table 6. We can see that 18 of the top 20 WCCMR papers are influential papers, while the numbers for Citation Count, MutualRank, Tri-Rank, PR-SVR, and PR-GraLap are 16, 15, 16, 7, and 8, respectively. All the methods fail to identify the most influential paper, but all of them successfully identify the second most influential paper in top 20.

Ranking effectiveness for authors

We use both GoldA and WRC for influence evaluation of authors (see Eq. 19 in “Ranking benchmarks” section for its definition). Figure 4 shows the effectiveness curves of the different algorithms for ranking authors measured by precision and NDCG. From Fig. 4, we can see that the proposed method, WCCMR, is better than all the other methods when NDCG is used, MutualRank is the worst, while the other four are very close. However,

Table 6 Top 20 papers ranked by WCCMR and other baseline methods (compared with the Gold standard ranking in descending order of the times of recommendation received, each number indicates the ranking position of that paper in the Gold standard ranking, an interval is given if two or more papers share the same ranking position inside the Gold standard ranking)

Rank	WCCMR	CC	MR	Tri	SVR	GraLap
1	2	31–93	31–93	31–93	5–11	5–11
2	31–93	2	2	2	–	–
3	31–93	5–11	12–30	12–30	–	–
4	5–11	31–93	5–11	5–11	–	–
5	12–30	31–93	31–93	5–11	–	–
6	12–30	–	5–11	–	31–93	31–93
7	–	12–30	12–30	31–93	12–30	12–30
8	31–93	5–11	31–93	12–30	2	2
9	31–93	12–30	31–93	31–93	–	–
10	5–11	31–93	–	31–93	–	12–30
11	31–93	–	12–30	12–30	–	–
12	–	5–11	12–30	–	–	–
13	31–93	31–93	–	12–30	–	–
14	12–30	31–93	31–93	–	12–30	–
15	31–93	12–30	–	5–11	5–11	5–11
16	31–93	5–11	–	31–93	12–30	12–30
17	12–30	3	31–93	3	–	–
18	5–11	–	–	31–93	–	–
19	5–11	31–93	5–11	5–11	–	31–93
20	31–93	–	31–93	–	–	–

CC Citation Count; MR MutualRank; Tri = Tri-Rank; SVR = PR-SVR; GraLap = PR-GraLap

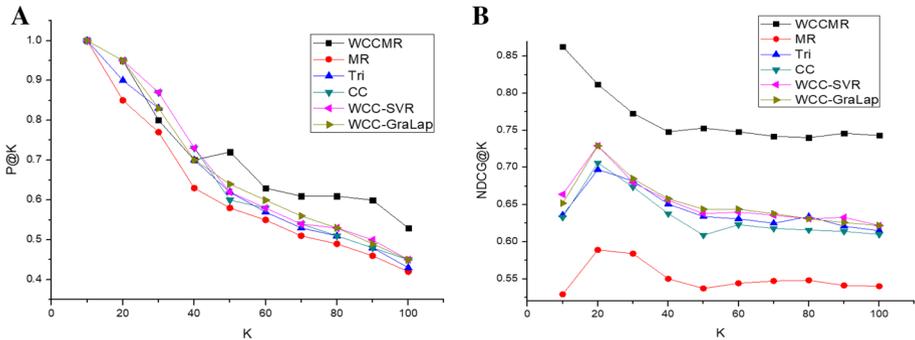


Fig. 4 Effectiveness of different algorithms for ranking authors. **a** Measured by P@K. **b** Measured by NDCG@K

when P@K is used, the performances of all the methods are closer. When K is 50 or more, WCCMR is a little better than the others. MutualRank is the worst in most of the cases, although the difference between it and the others is small.

To have a close look at the top 20 ranked authors by all the methods involved, we list them in Table 7 their corresponding ranking position in GoldA by their WRT scores. MutualRank identifies 17 influential authors, while all other methods reach 19. The results

Table 7 Top 20 authors ranked by WCCMR and other baseline methods (compared with the Gold standard ranking in descending order of WRT scores, each number indicates the ranking position of that paper in the Gold standard ranking, an interval is given if two or more papers share the same ranking position inside the Gold standard ranking)

Rank	WCCMR	CC	MR	Tri	SVR	GraLap
1	4	4	36–44	4	4	4
2	1	27–30	27–30	27–30	27–30	27–30
3	7	75–79	4	1	9	9
4	8	9	9	25	1	1
5	75–79	14	8	7	31	31
6	2	1	25	6	75–79	75–79
7	27–30	31	31	14	23	14
8	6	25	16	9	2	2
9	36–44	16	1	8	14	16
10	3	2	14	16	36–44	36–44
11	19	6	6	31	25	23
12	23	7	2	75–79	16	25
13	36–44	23	–	2	83	6
14	36–44	51	47	23	7	51
15	27–30	36–44	7	51	51	83
16	25	83	51	27–30	6	7
17	83	27–30	–	83	3	–
18	16	8	–	69	8	3
19	–	3	23	36–44	–	8
20	14	–	83	–	27–30	27–30

CC ranks authors by their total citation count; MR MutualRank; Tri = Tri-Rank; SVR = WCC-SVR; GraLap = WCC-GraLap

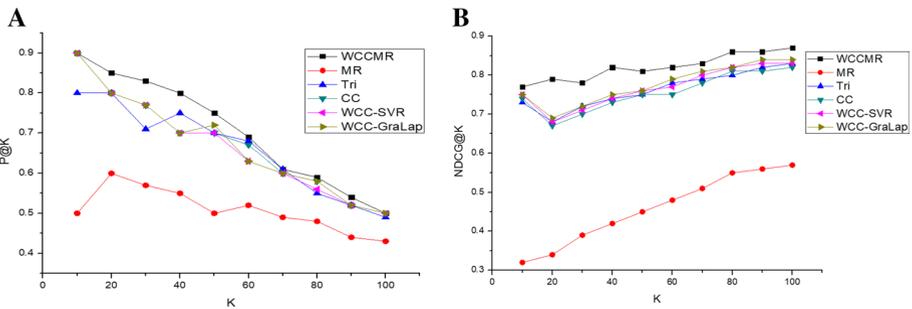


Fig. 5 Effectiveness of different algorithms for ranking venues. **a** Measured by Precision. **b** Measured by NDCG

show that all the algorithms are very good on identifying influential authors. Therefore, P@20 is very good for all the methods involved.

Ranking effectiveness for venues

Figure 5 shows the effectiveness curves of different algorithms for ranking venues measured by precision and NDCG. From Fig. 5, we can see that WCCMR performs better than

Table 8 Top 20 venues ranked by WCCMR and other baseline methods (compared with the Gold standard ranking in descending order of recommended paper numbers, each number in the table indicates the corresponding ranking position of that venue in the Gold standard ranking, an interval is given if two or more venues share the same ranking position inside the Gold standard ranking)

Rank	WCCMR	CC	MR	Tri	SVR	GraLap
1	4	4	18–26	18–26	4	4
2	2	2	–	4	2	18–26
3	1	18–26	–	2	18–26	2
4	8–14	1	4	17	1	1
5	18–26	8–14	32–55	–	8–14	8–14
6	32–55	5–6	–	8–14	5–6	5–6
7	5–6	–	–	8–14	–	32–55
8	32–55	15–16	5–6	–	17	17
9	27–31	32–55	–	32–55	15–16	–
10	–	17	2	27–31	32–55	15–16
11	32–55	32–55	32–55	15–16	32–55	32–55
12	5–6	18–28	32–55	–	–	32–55
13	8–14	–	–	8–14	32–55	18–26
14	–	32–55	32–55	5–6	32–55	32–55
15	15–16	32–55	32–55	32–55	18–26	–
16	–	32–55	32–55	15–16	32–55	32–55
17	3	27–31	8–14	1	27–31	27–31
18	32–55	–	–	7	–	–
19	7	–	17	18–26	–	–
20	8–14	18–26	–	–	18–26	18–26

CC citation count; MR MutualRank; Tri=Tri-Rank; SVR=WCC-SVR; GraLap=WCC-GraLap

the other algorithms when either the precision or NDCG is used. However, the difference between WCCMR and four others besides MutualRank is small. MutualRank is the worst and it is much worse than all the others.

For the top 20 venues returned by WCCMR and all other algorithms, we also list their corresponding ranking positions by the number of recommended papers in Table 8. It shows that all five algorithms besides MutualRank are equally good by identifying the same number of 16 influential venues, while MutualRank is not as good as the others and it secures 12 of them.

Average and median ranking positions of all influential entities

It is generally accepted that a good ranking algorithm should be effective in identifying all the influential entities in a comprehensive style (Wang et al., 2019). For the ranked list from a given ranking method, we find out the ranking positions of all those influential entities (e.g., all the papers in GoldP) and calculate the average rank and median rank of them. In this way, we are able to evaluate the general performance of the algorithm by using a single metric. Figure 6 shows the results.

From Fig. 6, we can see that the average rank and the median rank for WCCMR are the smallest in all the cases. In five out of six cases, the difference between it and the others are significant. However, the difference is very small in the case of average rank for venues. On the other hand, considering performance variance of all the algorithms involved, paper

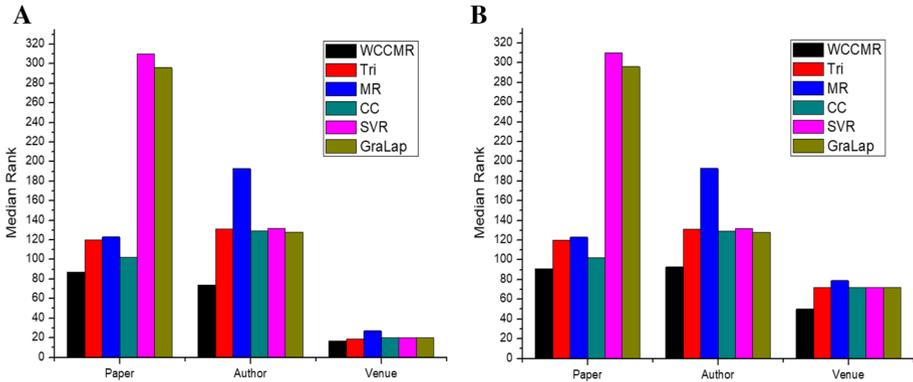


Fig. 6 Performance of different ranking methods by identifying the positions of all influential entities. **a** Measured by average ranking positions. **b** Measured by median ranking positions

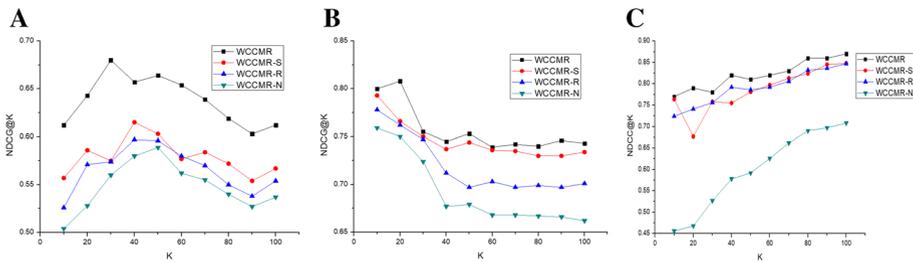


Fig. 7 Comparison of three feature-based variants of WCCMR with the original algorithm. **a** Paper ranking. **b** Author ranking. **c** Venue ranking

ranking is the highest, venue ranking is the lowest, while author ranking is in the middle. Especially when average rank is considered for author ranking, all the algorithms are very close.

Evaluation of several variants of WCCMR

WCCMR incorporates a few factors such as variable citation weights and bonus for recent citations. It is interesting to find how these two factors impact ranking performance. To achieve this goal, we define some variants that implement none or one of the features of WCCMR.

1. WCCMR-R. It is a variant of WCCMR that sets equal weight to all the citations.
2. WCCMR-S. It is a variant of WCCMR that does not implement bonus for recent citations.
3. WCCMR-N. It is a variant of WCCMR. It sets all citation weights equally and does not implement bonus for recent citations.

Now let us have a look at how these variants perform compared with the original algorithm. See Fig. 7 for the results. It is not surprising that WCCMR performs better than all

three variants of WCCMR, while the variant with none of the two components performs the worst in ranking all three types of academic entities. Such a phenomenon demonstrates that both components are useful for entity ranking, either used separately or in combination. However, the usefulness of these two components is not the same. In most cases, WCCMR-S performs better than WCCMR-R, which means that variable citation weights have larger impact than bonus for recent citations.

Robustness

Some types of abnormality may happen in citation networks. It can be caused by citation manipulation. Such a phenomenon certainly impacts the ranking of scientific entities, especially for PageRank-like algorithms. Therefore, robustness is a desirable property for ranking algorithms to fight against inappropriate citations. Of course, if there is no way to distinguish important citations from trivial ones, then we cannot do much to mitigate this problem. Therefore, we assume that it is more likely that citation manipulation happens to those with low to moderate citation strength and/or topical similarity and to those recently published papers.

To investigate the robustness of WCCMR when working with an abnormal network, we need a proper data set. AAN may not be good for this without any moderation. Instead of using some other data sets, we decide to make AAN more suitable for this purpose by adding some fake citations into it. Let us look at the situation for paper, author, and venue ranking separately.

- For paper ranking, we select a target paper p_t from the data set, then generate up to 50 fake papers, and each of which cites p_t and a number of others chosen randomly.
- For author ranking, we select a target author a_t from the data set, then generate up to 50 fake papers, and each of which cites a randomly chosen paper written by a_t and a number of others not written by a_t .
- For venue ranking, we select a target venue v_t from the data set, then generate up to 50 fake papers, and each of which cites a randomly selected paper published in v_t and a number of other papers not published in v_t .

For a target entity, we observe its ranking position change when more fake citations are added into the network. It is obvious that if an entity already has relatively a large number of citations, then adding a few more may not affect much its ranking position, while those entities with very few citations are more sensitive to such changes. In order to investigate the robustness of our algorithm, we choose those entities with very few citations (0 citation for a paper or an author and up to 10 citations for a venue). For all added fake citations, both citation strength and topical similarity are set to small to moderate values. We use rank difference to measure the robustness of any algorithm $\Delta R_h = R_0 - R_h$. Here R_0 is the initial rank of the entity and R_h is the rank position of the entity after h citations are added. Naturally, smaller rank difference indicates better robustness (Zhou et al., 2016).

Figure 8 shows the results of a group of algorithms, which is the average of 50 trials. The curves of WCC-SVR, WCC-GraLap always overlap with each other, because they are implemented in a very similar way with small difference. Not surprisingly, Citation Count is the most sensitive to added citations and WCCMR is the most insensitive, while WCC-SVR, WCC-GraLap, and Tri-Rank are in the middle.

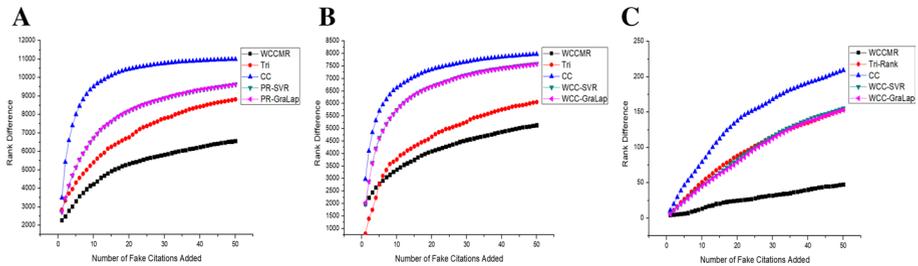


Fig. 8 Robustness of different ranking algorithms against citation manipulation. **a** Paper ranking. **b** Author ranking. **c** Venue ranking

Conclusions

In this paper, we have presented a ranking method for the impact of papers, authors, and venues in a heterogeneous academic network. Its main characteristic is rather than assigning equal weights to all the citations, we assign variable weight to each of them based on its strength and topical similarity between the citing paper and the cited paper. Both of these two values are determined through content analysis of the papers involved. Especially the ensemble learning technique has been used to decide citation strength of two papers. Experiments carried out with a publicly available data set AAN show that the proposed ranking algorithm, WCCMR, outperforms other baseline algorithms including MutualRank, Tri-rank, and GraLap.

Based on the AAN data set with some fake citations added, we demonstrate that WCCMR is more robust than the others. Although the data set used for this purpose is not completely real, the assumptions behind the artificial citations is reasonable.

As our future work, we would go further in a few directions. The first is to study appropriate approaches to deal with the missed citation information in the data set used. For example, for many papers in the AAN data set, their citation information is not complete. Some external resources such as Google scholar and Microsoft Academic may be used to enhance it. How to include such extra information into the academic network and the ranking framework in an efficiently and effectively style is a challenging issue. The second is how to evaluate academic entities across disciplines. For example, Biology and Mathematics are very different. One can expect that on average a Biology research paper can attract more citations than a Mathematics research paper. Even inside one discipline different research areas may have different properties. For example, in computer science, one can expect that on average a machine learning paper may attract more citations than an information retrieval paper. How to balance disparity among different disciplines or areas is also a challenging research problem. The third is to further study machine learning methods for content-based citation strength estimation. Two major subtasks includes detecting useful features and effective machine learning models.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abbas, A. M. (2011). Weighted indices for evaluating the quality of research with multiple authorship. *Scientometrics*, 88(1), 107–131.
- Bai, X., Xia, F., & Lee, I. (2016). Identifying anomalous citations for objective evaluation of scholarly article impact. *PLoS ONE*, 11(9), e0162364.
- Bai, X., Zhang, F., Ni, J., Shi, L., & Lee, I. (2020). Measure the impact of institution and paper via institution-citation network. *IEEE Access*, 8, 17548–17555.
- Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, 68(5), 314–316.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.
- Chakraborty, T. & Narayanam, R. (2016). All Fingers are not Equal: Intensity of References in Scientific Articles. In *Conference on empirical methods in natural language processing* (Pp. 1348–1358).
- Chawla, D. S. (2019). Elsevier investigates hundreds of peer reviewers for manipulating citations. *Nature*, 573, 174.
- Councill I. G., Giles C. L. & Kan M. -Y. (2008). Parscit: an open-source CRF reference string parsing package. In *Proceeding of the Language Resources and Evaluation Conference* (Pp. 661–667).
- Du, J., & Tang, X. (2013). Potential of harmonic counts for encouraging ethical co-authorship practices. *Scientometrics*, 96(1), 277–295.
- Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, 10(2), 392–407.
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131–152.
- Egghe, L., Rousseau, R., & Hooydonk, G. V. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science*, 51(2), 145–157.
- Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS One*. <https://doi.org/10.1371/journal.pone.0187394>
- Foo, J. (2011). Impact of excessive journal self-citations: A case study on the Folia Phoniatrica et Logopaedica journal. *Science and Engineering Ethics*, 17(1), 65–73.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471–479.
- Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1), 90–93.
- González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3), 379–391.
- González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals scientific prestige: The SJR2 indicator. *Journal of Informetrics*, 6(4), 674–688.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572.
- Jiang, X. R., Sun, X. P., Yang, Z., Zhuge, H., & Yao, J. M. (2016). Exploiting heterogeneous scientific literature networks to combat ranking bias: Evidence from the computational linguistics area. *Journal of the Association for Information Science and Technology*, 67(7), 1679–1702.
- Johnson, R., Watkinson, A. & Mabe, M. (2018). The STM report: an overview of scientific and scholarly publishing. https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf. Accessed June 2019.
- Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., & Vassiliou, Y. (2021). Impact-based ranking of scientific publications: A survey and experimental evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 33(4), 1567–1584.
- Liu, Z. R., Huang, H. Y., Wei, X. C. & Mao, X. L. (2014). Tri-Rank: An Authority Ranking Framework in Heterogeneous Academic Networks by Mutual Reinforce. In *26th IEEE international conference on TOOLS with artificial intelligence (ICTAI2014)* (Pp. 493–500).

- Meng, Q. & Kennedy, P. J. (2013). Discovering influential authors in heterogeneous academic networks by a co-ranking method. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (Pp. 1029–1036).
- Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3), 265–277.
- Noorden, R. V., & Chawla, D. S. (2019). Hundreds of extreme self-citing scientists revealed in new database. *Nature*, 572, 578–579.
- Pajić, D. (2015). On the stability of citation-based journal rankings. *Journal of Informetrics*, 9(4), 990–1006.
- Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4), 919–944.
- Simkin, M. V., & Roychowdhury, V. P. (2003). Read before you cite! *Complex System*, 14(2003), 269–274.
- Stallings, J., Vance, E., Yang, J., Vannier, M., Liang, J., Pang, L., Dai, L., Ye, I., & Wang, G. (2013). Determining scientific impact using a collaboration index. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24), 9680–9685.
- Teufel, S., Siddharthan, A. & Tidhar, D. (2006). Automatic classification of citation function. In *Conference on empirical methods in natural language processing* (Pp.103–110).
- Walker, D., Xie, H., Yan, K., & Maslov, S. (2006). Ranking scientific publications using a simple model of network traffic. *Journal of Statistical Mechanics-Theory and Experiment*, 6(6), P06010–P06015.
- Waltman, L., Eck, N. J. V., Leeuwen, T. N. V., & Visser, M. S. (2013). Some modifications to the snip journal impact indicator. *Journal of Informetrics*, 7(2), 272–285.
- Wan, X. J., & Liu, F. (2014). Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, 65(9), 1929–1938.
- Wang, S. Z., Xie, S. H., Zhang, X. M., Li, Z. J., Yu, P. S., & He, Y. Y. (2016). Coranking the future influence of multi-objects in bibliographic network through mutual reinforcement. *ACM Transactions on Intelligent Systems and Technology*, 7(4), 1–28.
- Wang, Y., Zeng, A., Fan, Y., & Di, Z. (2019). Ranking scientific publications considering the aging characteristics of citations. *Scientometrics*, 120(3), 155–166.
- Xu, H., Martin, E., & Mahidadia, A. (2014). Contents and time sensitive document ranking of scientific literature. *Journal of Informatics*, 8(3), 546–561.
- Yang, C., Liu, T., Chen, X., Bian, Y., & Liu, Y. (2020). HNRWalker: Recommending academic collaborators with dynamic transition probabilities in heterogeneous networks. *Scientometrics*, 123(1), 429–449.
- Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8), 1635–1643.
- Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3), 467–477.
- Zhang, F. & Wu, S. (2018). Ranking scientific papers and venues in heterogeneous academic networks by mutual reinforcement. In: *ACM/IEEE joint conference on digital libraries (JCDL)* (Pp.127–130).
- Zhang, F., & Wu, S. (2020). Predicting future influence of papers, researchers, and venues in a dynamic academic network. *Journal of Informatics*, 14(2), 101035.
- Zhang, J., Xu, B., Liu, J., Tobla, A., Al-Makhadmeh, Z., & Xia, F. (2018). PePSI: Personalized prediction of scholars' impact in heterogeneous temporal academic networks. *IEEE Access*, 6, 55661–55672.
- Zhang, L., Fan, Y., Zhang, W., Zhang, S., Yu, D., & Zhang, S. (2019a). Measuring scientific prestige of papers with time-aware mutual reinforcement ranking model. *Journal of Intelligent and Fuzzy Systems*, 36, 1505–1519.
- Zhang, Y., Wang, M., Gottwalt, F., Saberi, M., & Chang, E. (2019b). Ranking scientific articles based on bibliometric networks with a weighting scheme. *Journal of Informetrics*, 13(2), 616–634.
- Zhao, F., Zhang, Y., Lu, J., & Shai, O. (2019). Measuring academic influence using heterogeneous author-citation networks. *Scientometrics*, 118(3), 1119–1140.
- Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2016). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 106(2), 805–816.
- Zhou, X., Liang, W., Wang, K., Huang, R., & Jin, Q. (2021). Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. *IEEE Transactions on Emerging Topics in Computing*, 9(1), 246–257.
- Zhu, X. D., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the American Society for Information Science and Technology*, 66(2), 408–427.