Scientometrics (2022) 127:4045–4074
https://doi.org/10.1007/s11192-022-04429-z

Analysing academic paper ranking algorithms using test data and benchmarks: an investigation

Yu Zhang[1] · **Min Wang**[2] · **Morteza Saberi**[3] · **Elizabeth Chang**[1]

Received: 29 September 2021 / Accepted: 31 May 2022 / Published online: 21 June 2022

Abstract
Research on academic paper ranking has received great attention in recent years, and many algorithms have been proposed to automatically assess a large number of papers for this purpose. How to evaluate or analyse the performance of these ranking algorithms becomes an open research question. Theoretically, evaluation of an algorithm requires to compare its ranking result against a ground truth paper list. However, such ground truth does not exist in the field of scholarly ranking due to the fact that there does not and will not exist an absolutely unbiased, objective, and unified standard to formulate the impact of papers. Therefore, in practice researchers evaluate or analyse their proposed ranking algorithms by different methods, such as using domain expert decisions (test data) and comparing against predefined ranking benchmarks. The question is whether using different methods leads to different analysis results, and if so, how should we analyse the performance of the ranking algorithms? To answer these questions, this study compares among test data and different citation-based benchmarks by examining their relationships and assessing the effect of the method choices on their analysis results. The results of our experiments show that there does exist difference in analysis results when employing test data and different benchmarks, and relying exclusively on one benchmark or test data may bring inadequate analysis results. In addition, a guideline on how to conduct a comprehensive analysis using multiple benchmarks from different perspectives is summarised, which can help provide a systematic understanding and profile of the analysed algorithms.

Keywords Academic paper ranking · Ranking algorithms · Bibliometric analysis · Citation analysis

Yu Zhang and Min Wang have contributed equally to this work.

✉ Yu Zhang
yu.zhang@adfa.edu.au

Extended author information available on the last page of the article

Introduction

The academic paper rankings have been used for decision making in many areas including university ranking, grant funding, academic hiring to name a few. Biased or unreliable ranking results could mislead these important decisions and should be alleviated by the community of scientometrics. While we have seen advanced ranking methods in the literature and new ones are being proposed, there is an urgent need for a generalised oversight mechanism collectively, otherwise the proposed ones cannot lead to reliable ranking at the end but to bad practices. The society of AI has recognised this issue in general, and some progress, including developing a collective standard, has been undertaken (Ristoski et al., 2016). In this paper, we present to what extent different paper ranking methods are performing differently when there is no gold standard available.

The digitisation has greatly facilitated the academic publications, but an ever-growing number of articles bring great challenges for researchers to identify the important ones (Zhang et al., 2019a). To address this problem, many algorithms have been proposed for ranking academic papers. For example, the classic PageRank algorithm (Page et al., 1999) measured the importance of papers based on article citation networks. Its variants, including CoRank (Zhou et al., 2007), P-Rank (Yan et al., 2011a), FutureRank (Sayyadi & Getoor, 2009) and W-Rank (Zhang et al., 2019b), extended the homogeneous citation network into heterogeneous networks that further integrated the author, venue, and publication time information. Given a collection of papers and their bibliometric information, these algorithms calculate scores for the papers and translate the scores into paper ranking lists where the top papers can be recommended to users (Zhang et al., 2019c). However, how to evaluate or analyse the performance of these ranking algorithms is an open research question (Xia et al., 2017; Cai et al., 2019).

Theoretically, evaluation of a ranking algorithm requires to compare its results against a ground truth paper list. However, such ground truth does not exist in the field of scholarly ranking due to the fact that there does not and will not exist an absolutely unbiased, objective, and unified standard to formulate the quality or impact of academic papers. Similar situation also exists in the field of Learning to Rank (L2R), where comparing different L2R models based on their evaluation results is hindered by the absence of a standard set of benchmark collections (Tax et al., 2015). In practice, researchers evaluate or analyse their proposed ranking algorithms by different approaches, such as using test data collected from domain expert decisions (Dunaiski et al., 2016; Mariani et al., 2016; Dunaiski et al., 2018) and comparing against predefined benchmarks. Examples of such benchmarks include citation count-based benchmarks (Chen et al., 2007; Yan et al., 2011b; Wang et al., 2013; Ma et al., 2018) and citation network-based benchmarks (Sayyadi & Getoor, 2009; Waumans & Bersini, 2017). Employing different benchmarks and test data sets for evaluation can raise concerns about the reliability and comparability of the results, leading to questions such as what the relationships are between these approaches? Whether using these approaches produces different analysis results, and if yes, how the performance of ranking algorithms should be analysed?

To answer these questions, this research focuses on assessing the impact of using different benchmarks and test data on the analysis results of paper ranking algorithms. Firstly, we clarify the definitions of the terms that are used in this paper, and then review the existing approaches that have been used to evaluate and analyse paper ranking algorithms in the literature and summarise them into two major categories, which are citation-based benchmarks and test data. Secondly, we conduct experiments to examine their relationships and

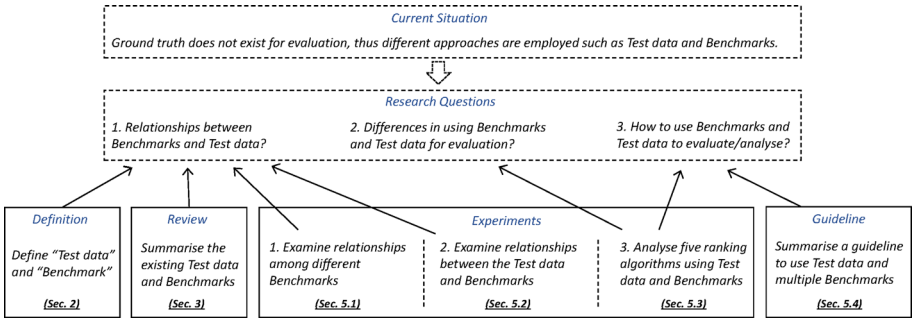


Fig. 1 Overall research structure. The motivations of this study, including the current situation and the research questions, are contained in the dashed boxes, while the studies of this research are enclosed in the solid line boxes. The arrows directing from the study actions to the research questions denote the design of this research in terms of how the actions are assigned to discuss and answer each question

perform a comparative analysis of these benchmarks as well as the test data. In addition, an investigation is carried out to evaluate or analyse nine ranking algorithms using these different benchmarks and test data. The Spearman's rank correlation coefficient, receiver operating characteristic (ROC) curve, and normalised discounted cumulative gain (nDCG) are used as evaluation measures to obtain insights into their scoring behaviours from three different perspectives, namely statistical relationship, classification performance, and ranking effectiveness. Finally, a guideline on how to analyse ranking algorithms using multiple benchmarks and test data is summarised based on our findings. The overall structure of this paper is demonstrated in Fig. 1 where the motivation is briefed in the dashed boxes and the actions undertaken in this study are enclosed by the solid lines.

This study discusses at length on the review and comparison of different benchmarks and test data and touches upon the practice suggestions of algorithm evaluation and analysis, leading three contributions to the field of academic paper ranking. The first contribution lies in the systematic review of the existing approaches that have been employed in literature to evaluate and analyse paper ranking algorithms with data. The second contribution is the investigation on the relationships amongst the benchmarks and test data, and the assessment of the impact of using different benchmarks and test data on the results when evaluating paper ranking algorithms. In light of the findings from the investigation, a guideline for analysing paper ranking algorithms using multiple benchmarks is summarised, which may provide researchers with a holistic vision and understanding of the analysed algorithms.

Definitions

As mentioned earlier, the evaluation of an ranking algorithm requires to examine its ranking result against a ground truth ranking list which does not exist especially in the case of academic paper ranking. Hence, the problem is how to evaluate or analyse paper ranking algorithms if such ground truth is unavailable. The existing studies form two genres: one is evaluating the algorithms using test data which contains a special collection of papers that integrates domain expert decisions; and the other is analysing the algorithms using predefined

ranking metrics which are referred to as benchmarks in this study. For clarity, we define the following terms.

Ground truth refers to the underlying objective ranking over a given collection of papers. In the field of scholarly rankings, there is usually no absolutely unbiased, objective and unified ground truth ranking list that can reflect the true impact, popularity and quality of papers.

Test data refers to a collection of bibliometric entities with certain ground truth that can be used to evaluate ranking algorithms. For academic paper rankings, a test data set usually contains papers assessed by domain experts (Bornmann & Marx, 2015a) or won certain awards, such as the best paper awards and high-impact awards (Dunaiski, 2019) in conferences and journals of a research field. Using test data is an effective way to evaluate ranking algorithms but one should be aware of the availability issue, such as the potential difficulties in obtaining appropriate test data and the limitations of choices in research fields and publication venues (Dunaiski et al., 2018).

Benchmarks refer to the predefined ranking metrics used for examining the ranking results of an algorithm. Examples of such benchmarks used in existing studies include future citation count (FCC) (Wang et al., 2013) and future PageRank (FPR) (Sayyadi & Getoor, 2009). Since ground truth refers to the information provided by direct observation as opposed to that provided by inference, benchmarks are not desirable candidates to form ground truth because they can only infer rather than observe the impact or quality of papers. Although benchmarks cannot be used directly as ground truth, analysing the ranking algorithms based on benchmarks is still a useful way to gain insights into the algorithms. For example, the FCC was adopted as benchmark to evaluate, or more precisely speaking, analyse a proposed paper ranking algorithm which was able to rank papers using a heterogeneous network that integrates citation, author, and journal/conference information; and the publication time information is also involved in the network propagation (Wang et al., 2013). Since the motivation underlying the algorithm design is to capture time information in the evolving network to obtain predictions on the future impact of papers, it is reasonable to compare its ranking result with the FCC ranking list as it can infer the future impact of papers and is an interpretable benchmark. Another study adopted FPR as benchmark to analyse a proposed paper ranking algorithm named FutureRank (Sayyadi & Getoor, 2009). This algorithm inherited the basic assumption of PageRank, that is, important articles are likely to receive more citations from other articles, and extended this definition from historical citation networks (data available at the user query time) to future citation networks (future data after the user query time). The proposed FutureRank exploits the citation, author, and publication time information in order to predict the FPR of query papers, where the FPR, in the authors' opinion, infers the future impact of papers.

In addition, we define *evaluation measures* as the measures used to judge the performance of an algorithm and to compare the performance of different algorithms. Three evaluation measures are adopted in our analysis, including the Spearman's rank correlation coefficient, ROC curves and nDCG, which consider the statistical relationship, classification performance, and ranking effectiveness, respectively.

Test data and benchmarks

This section reviews the two practical approaches for evaluating and analysing paper ranking algorithms. The first approach is using test data, and the second one is using benchmarks. Since a benchmark itself is a ranking metric that infers the impact or quality of papers from a specific perspective, using different benchmarks can reflect the

characteristics and the scoring behaviours of the analysed ranking algorithms. In this study, we group the benchmarks into two categories, one is based on citation counts (where the paper scores are directly computed by counting citation numbers), and the other is based on citation networks (where the paper scores are computed by iterative calculations on the future citation networks).

Test data

Peer assessment by domain experts is generally considered as a method that is more appropriate to obtain paper rankings than citation-based metrics. Human expert decisions are based on human intelligence and their domain knowledge which involves a comprehensive evaluation of all kinds of information. Hence, the scoring or ranking results given by human experts are often considered more reliable and convincing because they examine the quality and impact of each paper using complex but interpretable criteria that do not exclusively rely on citations (Ahlgren & Waltman, 2014; Saarela et al., 2016). However, hiring domain experts to evaluate large-scale paper sets is impractical since it is expensive and time-consuming. Therefore, this approach is only suitable for the application scenarios where the number of scholarly entities to be evaluated are within the human capacity.

A similar but practical approach is to collect test data sets based on existing assessments provided by domain experts. For example, Bornmann and Marx (2015a) collected a test data set from a post-publication peer review system of the biomedical literature. However, such data resources only exist in few cases. An alternative approach of collecting test data sets is to use certain reputable awards, such as the best paper awards and high-impact paper awards in conferences and journals of a research field (Dunaiski et al., 2016).

Test data has been widely used to evaluate rankings of entities in scholarly citation networks. For example, a test data set comprising papers awarded the prize of being Most Influential Papers (MIP) by the ICSE program committee was used to evaluate the ranking results of NewRank (Dunaiski & Visser, 2012). According to the ICES MIP awarding rules[1], each year the current program committee for ICSE'N reviews the papers from ICSE'(N-10) and awards the ones which are consider to be the most influential paper during the previous 10 years. SARank (Ma et al., 2018) proposed the RECOM approach which assumed academic papers with larger number of recommendations indicated higher importance of the papers. Following this assumptions, the RECOM adopted the number of recommendations of 93 articles on ACL Anthology Network (AAN) dataset (Radev et al., 2013) as ground truth. Sidiropoulos and Manolopoulos (2005) used the 'VLDB 10 Year Award' and the 'SIGMOD Test of Time Award' as test data to evaluate ranking algorithms for scientific publications. Mariani et al. (2016) used the list of Milestone Letters (MLs) selected by editors of Physical Review Letters from American Physical Society (APS) dataset. Dunaiski et al. (2016) collected a list of 207 academic papers which received high-impact awards from 14 conferences from the domain of computer science and used it to evaluate paper ranking algorithms. As for predicting high-impact papers, they used 464 papers which were awarded best papers by 32 venues for evaluation. Later, Dunaiski et al. (2018) extended the test data to 1155 best awarded papers collected from 36 venues and 849 high-impact awarded paper from 30 venues to evaluate rankings of academic papers.

[1] http://www.sigsoft.org/awards/icseMIPAward.html.

Using test data is an effective way to evaluate ranking algorithms. However, collecting appropriate large-scale test data sets is a challenge due to the limited resources which are only available in specific research fields, publication venues (journals and conferences) and publication time. In addition, the difficulty in ensuring the obtained test data is representative, unbiased, and sufficient is another obstacle to the collection and broader application of test data.

Citation count-based benchmarks

Citation count is an intuitive and interpretable indicator to infer the impact of a paper, and it has been widely used to analyse paper ranking algorithms. Given a collection of query papers and a predefined historical time point (the user query time), citation count can be further refined to historical citation count (HCC) and future citation count (FCC).

Historical citation count

The HCC infers the impact of the query papers based on their citation status at the time of the user query. The assumption is that high-impact papers should have received more citations at the time of the user query. Therefore, for each query paper, HCC is the number of citations received before a historical time point.

HCC has a long history of being used as benchmark to analyse ranking algorithms (Lawani & Bayer, 1983). For instance, in the study where PageRank was introduced in citation analysis for the first time, citation count was adopted as the benchmark to demonstrate the effectiveness of PageRank (Chen et al., 2007). Yan et al. (2011b) used the paper citations as a measurement for the popularity among researchers and focused on the problem of citation count prediction to examine the characteristics for popularity, and citation counts were used to assess the popularity predictions. Later, another algorithm, P-Rank (Yan et al., 2011a), was proposed which took into account the heterogeneous network containing articles, authors, and journals. In this study, the citation count was used to generate a benchmark list in which the top 20 papers were compared with the top 20 papers ranked by the P-Rank algorithm. Additionally, a recently proposed ranking algorithm, PePSI (Zhang et al., 2018), investigated personalised prediction of scholars' scientific impact by classifying scholars into different types according to their citation dynamics, and the citation count was adopted as the benchmark.

Future citation count

The FCC infers the impact of the query papers based on their future citation trends after the user query time with the assumption that high-impact papers should be able to obtain more citations in the future. Therefore, for each query paper, FCC is the number of citations received after a historical time point.

The relationship between HCC and FCC is illustrated in Fig. 2. It is worth noting that the use of FCC requires to define a historical time point to divide a database into two parts: a historical part and a future part. The historical time point is an imaginary time point predefined by researchers to simulate the time of the user query. They view the system from the historical time point, run the ranking algorithms using historical data before this time point, and calculate FCC on the future data as benchmark to evaluate the ranking algorithms.

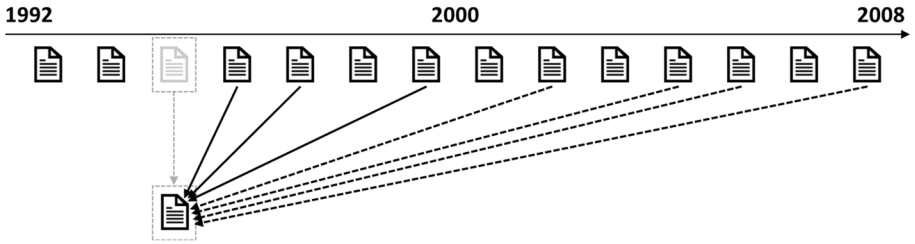


Fig. 2 Illustration of HCC and FCC. In this example, the user query time is the year 2000 and for a query paper *P*, its historical citations and future citations are denoted as solid lines and dashed lines, respectively

FCC was used as a benchmark to assess a heterogeneous network-based ranking algorithm that integrated PageRank and HITS (Wang et al., 2013). To analyse the performance of this algorithm, the FCC of the query papers was generated and compared with the ranking result of the proposed algorithm. Focusing on ranking the future popularity of new publications and young researchers, MRFRank (Wang et al., 2014) was proposed which combined various available information, such as time and text features. The number of papers' future citations was employed as benchmark by sorting them in the descending order. CocaRank (Zhang et al., 2016) and MRCoRank (Wang et al., 2016) were proposed to measure the future influence of academic rising stars for both researchers and publications, and both algorithms used the FCC as benchmarks for analysis. In addition, a recent query independent academic paper ranking algorithm, SARank (Ma et al., 2018), was proposed which further improved traditional PageRank with a time decaying factor, and it adopted two benchmarks, namely RECOM (short for recommendation) and PFCTN (short for past and future citations). The PFCTN proposed to use both past and future citations of a paper with the same period of time to represent the significance of the paper at the concerned time, while the RECOM assumed more recommendations indicated higher importance.

The advantages of using citation counts as benchmarks lie in the interpretable results, low computational cost, and high applicability. In addition, HCC and FCC formulate the impact of papers differently as they respectively consider the current citation status and future trends. However, they also show limitations. Firstly, HCC tends to favour older papers because they have existed longer for more citations. Some old articles become unfashionable, but they may have received many citations in the past. Secondly, as discovered by Bornmann and Daniel (2008), the motivations for citing articles are different, thereby the impact represented by the citations could be different. Finally, citation count is partially subject to research fields and publication time because the citation densities are uneven in different fields (Radicchi et al., 2008) and are dynamic over time (Bornmann & Mutz, 2015b).

Weighted future citation count

The weighted FCC (wFCC) further improves the FCC by considering the quality of each citation. Since the reference papers are usually cited for various reasons and motivations, these citations should not be treated equally (Garfield, 1979). The concept of citation relevance was proposed accordingly (Zhang et al., 2019b). It is an attribute of the citation link, with a value between 0 and 1. Specifically, a larger value indicates a higher relevance between the two papers of a citation link. The calculation of citation relevance consists of

two components, one is the semantic similarity based on the contextual information of the two papers and the other is the structural similarity calculated from the citation network. The interpretation is that the semantic similarity of two papers is higher when the issues or methods they are addressing are similar or related, meanwhile, their structural similarity is higher when they are simultaneously linked by more common article nodes in the citation network. A citation is considered highly relevant when the two articles are semantically similar or share many mutual links in the citation network.

The definition of the wFCC is built on top of the FCC, and includes w, a weight indicating the citation relevance. It is given by:

$$w_{i,j} = \alpha \cdot S_{\text{semantic}}(p_i, p_j) + \beta \cdot S_{\text{structural}}(p_i, p_j) \tag{1}$$

where $w_{i,j}$ denotes the relevance weight for citation from paper i to paper j, α and β are two hyper-parameters that adjust the ratio of the two components, S_{semantic} and $S_{\text{structural}}$ denote the semantic similarity and structural similarity between the two papers respectively. Calculation of the $S_{\text{semantic}}(p_i, p_j)$ is based on the 'align, disambiguate and walk' algorithm (Pilehvar et al., 2013) and the computation of $S_{\text{structural}}(p_i, p_j)$ is based on the cosine similarity in the citation network, as follows:

$$S_{\text{structural}}(p_i, p_j) = \frac{| L_{p_i} \cap L_{p_j} |}{\sqrt{| L_{p_i} | \times | L_{p_j} |}} \tag{2}$$

where L_{p_i} denotes the set of nodes linked to and from paper p_i, $L_{p_i} \cap L_{p_i}$ denotes the set of common nodes that connect to both p_i and p_j, and $| \cdot |$ denotes the number of nodes in the set.

The wFCC embeds the semantic and structural similarities between papers into the FCC, which potentially provides a benchmark that is less biased since wFCC is not exclusively based on citation count. However, the involvement of citation relevance reduces the interpretability of its ranking results, and the computational cost increases dramatically due to the complexity of semantic analysis.

Other variants of citation count include those considering the citation functions and conflict of interest (COI). In the first variant, only the functional citations were counted while the perfunctory citations were filtered out (Xu et al., 2014). The second variant only considered citations without COI, and it was used to evaluate the positive and negative COI-distinguished objective ranking algorithm proposed in the study (Bai et al., 2017). However, these variants share a similar limitation as the wFCC, that is, they require additional calculation procedures or modules to identify the function of each citation, or recognise the relationship of interest between the authors of the papers.

Time normalisation

Benchmarks based on citation counts tend to be biased towards old papers and underestimate new papers as they have less time to accumulate citations. For example, a paper published earlier with more citations is not necessarily more impactful than a paper published later with fewer citations. In other words, citations should not be treated equally and the publication time should be considered when comparing papers using citation count (Waltman, 2016). Therefore, in this study, we apply normalisation to citation count metrics to correct the bias introduced by publication time. The normalisation factor is based on an exponential function, as follows:

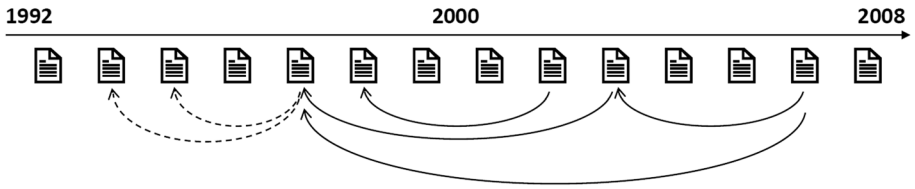


Fig. 3 Illustration of future citation network. In this example, the historical time point (user query time) is the year 2000, and the historical citation network and future citation network are denoted by the dashed lines and solid lines, respectively

$$Z_t = e^{-\rho \cdot (T_{\text{historical}} - t)} \tag{3}$$

where ρ is a constant set to 0.3, $T_{\text{historical}}$ is the historical time point (evaluation time point), and t denotes paper's publication year ($t \leq T_{\text{historical}}$). The normalised versions of HCC, FCC and wFCC metrics are denoted as HCC_t, FCC_t and wFCC_t, respectively, and they are calculated by multiplying by the normalisation factor as follows:

$$\text{HCC\_}t = Z_t \cdot \text{HCC} \tag{4}$$

$$\text{FCC\_}t = Z_t \cdot \text{FCC} \tag{5}$$

$$\text{wFCC\_}t = Z_t \cdot \text{wFCC}. \tag{6}$$

Citation network-based benchmarks

Compared with citation count-based benchmarks that only focus on the number of citations, citation network-based benchmarks further consider the citation relationships through iterative calculations on the citation networks and take into account the influences of other bibliometric entities such as authors and publication venues. Citation network-based benchmarks are recursively defined on the citation graph or extended heterogeneous graphs which integrate paper-author and paper-venue graphs.

In this study, we examine four citation network-based benchmarks, including the future PageRank (FPR), future PageRank with author entity considered (FPR+A), with publication venue entity considered (FPR+V), and with publication time entity considered (FPR+T). These benchmarks are selected based on the gradual expansion and utilisation of bibliographic information. Specifically, the FPR is based on the citation graph, FPR+A and FPR+V further incorporate the paper-author and paper-journal graphs respectively, and FPR+T considers citation graph and publication time. It is worth noting that these benchmarks are calculated based on future citation networks as illustrated in Fig. 3. The use of FPR-based benchmarks requires to set a historical time point, as in the case of FCC, to divide a database into two parts. One part is the historical citation network which contains query papers and their citations before the historical time point; and the other part is the future citation network which contains future citations received by the query papers and citations among papers published after the historical time point. The historical citation network is used for testing ranking algorithms, and the future citation network is used for generating benchmarks to analyse the behaviour of the ranking algorithms.

Future PageRank

The FPR refers to the PageRank algorithm running on the future citation networks. PageRank is an algorithm used by Google Search, one of the pioneers of Internet search engines, to rank web pages in their searching results (Page et al., 1999). PageRank determines a rough estimate of the importance of web pages based on the inbound and outbound connections to the pages. Let $S(p_i)$ denote the score estimated for paper p_i, PageRank is computed by the follow iteration:

$$S(p_i) \longleftarrow (1-d)\cdot\frac{1}{N} + d\cdot\sum_{p_j\in In(p_i)}\frac{S(p_j)}{\mid Out(p_j)\mid} \tag{7}$$

where $In(p_i)$ refers to the set of papers that cite p_i, $\mid Out(p_j)\mid$ is the number of outbound papers that p_j cites, N is the total number of papers under consideration, and d is a damping factor that usually set as 0.85 to allow a random jump in the citation network.

In PageRank, a paper with higher citations and is cited by papers with higher PageRank scores will have a higher ranking. However, it has the following limitations. Firstly, the random walking model does not conform to the paper evaluation and ranking behaviour in reality. Secondly, it does not filter internal and low-value links. Finally, it is biased to older papers and not friendly to high-quality new papers (Walker et al., 2007; Wang et al., 2014). Despite its drawbacks, PageRank is still an important algorithm in academic paper ranking and citation analysis.

The use of FPR as benchmark for algorithms analysis was proposed in the FutureRank algorithm (Sayyadi & Getoor, 2009). It took the top 50 papers sorted by FPR from a dataset as the true paper ranking and measured the precision by comparing this ranking list with the top 50 returned by the proposed FutureRank algorithm. Waumans and Bersini (2017) analysed the citation growth trend and leveraged this information to predict the significance of academic papers. To validate the proposed ranking algorithm, its ranking result was analysed using the FPR as benchmark.

Future PageRank and author

The FPR+A extends the citation network into a heterogeneous network that integrates the impact of authors through the author-paper network. In each iteration, two components are required, one from the citation network using PageRank and the other from the author-paper network using hyperlink-induced topic search (HITS) algorithm. Below we will briefly introduce the HITS algorithm and give the calculation formula of the PR+A based on it.

HITS is a well-known link analysis algorithm proposed to rate web pages (Kleinberg, 1999). Unlike PageRank, HITS defines hubs and authorities in a heterogeneous network, in which authoritative pages refer to the important high-quality pages, while hub pages refer to directory pages that link to authoritative pages. Two assumptions are applied, namely 1) a high-quality authoritative page will be pointed to by many high-quality hub pages; and 2) a high-quality hub page will point to many high-quality authoritative pages. In the field of academic paper ranking, these two assumptions are revised accordingly, namely 1) a high-quality authoritative paper will be cited by many high-quality hub papers; and 2) a high-quality hub paper will cite many high-quality authoritative

papers. Let $S(p_i, t)$ and $H(p_i, t)$ denote the authority score and hub score for paper p_i, the computation of HITS is based on recursive iterations between the authority update and hub update, as follow:

$$H(p_i) \longleftarrow \sum_{p_j \in Out(p_i)} S(p_j) \tag{8}$$

$$S(p_i) \longleftarrow \sum_{p_j \in In(p_i)} H(p_j) \tag{9}$$

where $In(p_j)$ and $Out(p_j)$ denote the set of papers that cite p_j and the set of papers that p_j cites, respectively. Let A denote the adjacency matrix of the citation network, the above equations can be rewritten as:

$$\vec{H} = \lambda A \vec{S} \tag{10}$$

$$\vec{S} = \mu A^T \vec{H} \tag{11}$$

where λ and μ are scale factors for normalisation, and $\vec{H}$ and $\vec{S}$ denote hub and authority scores for all papers. Hence, the authority score of papers can be computed by the following equation.

$$\vec{S} = \lambda \mu A^T A \vec{S} \tag{12}$$

Comparing with a paper's PageRank score which is only related to the papers that cite the paper, its HITS score is also related to papers that were cited by the paper. Therefore, HITS algorithm is easy to be manipulated by cheaters (Kleinberg, 1999). A cheater can turn a low-quality paper into a high-quality hub by citing many high-authority papers, and meanwhile, cites a cheating paper to improve its authority score. Despite being less stable than PageRank, HITS offers a neat solution to integrate the impact of other bibliometric entities in paper ranking. Specifically, a heterogeneous network can be defined by integrating the citation network with author-paper network and venue-paper network. Then the idea of hubs and authorities can be extended to the heterogeneous network.

For an author-paper network, the authors and papers are considered as hubs and authorities respectively, and they are updated by the following steps:

$$H(a_i) \longleftarrow \sum_{p_j \in Out(a_i)} \frac{S(p_j)}{\mid Out(a_i) \mid} \tag{13}$$

$$S_A(p_i) \longleftarrow \mu_A \sum_{a_j \in In(p_i)} H(a_j) \tag{14}$$

where $H(a_i)$ denote the hub scores of author a_i computed by collecting the authority scores from corresponding papers; $S_A(p_i)$ denote the authority score of paper p_i propagated from the corresponding authors in the paper-author network; $In(\cdot)$ and $Out(\cdot)$ denote the set of nodes link to and link out from the entity, respectively; and μ_A is a scaling factor used for normalisation. Finally, the PR+A score of paper p_j is a weighted summation of the two components, as follows:

$$S(p_i) = \alpha \cdot PageRank(p_i) + \beta \cdot S_A(p_i) + (1 - \alpha - \beta) \cdot \frac{1}{N} \tag{15}$$

where N is the total number of query papers, and α and β are two hyper-parameters that adjust the weight of each component. A similar approach was used by the CoRank algorithm (Zhou et al., 2007). This method couples two random walks in the citation network and the authors' social network by the paper author network to jointly rank authors and papers.

Future PageRank and venue

Similar to the PR+A, FPR+V is defined on a heterogeneous network that integrates the citation network and paper-venue network considering the influence of publication venues (journals and conferences). The same idea was used by the P-Rank algorithm for measuring prestige in heterogeneous scholarly networks (Yan et al., 2011a).

For the paper-venue network, the venues and papers are considered as hubs and authorities respectively, and they are updated by the following steps:

$$H(v_i) \longleftarrow \sum_{p_j \in Out(v_i)} \frac{S(p_j)}{\mid Out(v_i) \mid} \tag{16}$$

$$S_V(p_i) \longleftarrow \mu_V \sum_{v_j \in In(p_i)} H(v_j) \tag{17}$$

where $H(v_i)$ denote the hub scores of venue v_i computed by collecting the authority scores from corresponding papers; $S_V(p_i)$ denote the authority score of paper p_i propagated from the corresponding venues in the paper-venue network; $In(\cdot)$ and $Out(\cdot)$ denote the set of nodes link to and link out from the entity, respectively; and μ_V is a scaling factor used for normalisation. Finally, the PR+V score of paper p_j is a weighted summation of the two components, as follows:

$$S(p_i) = \alpha \cdot PageRank(p_i) + \beta \cdot S_V(p_i) + (1 - \alpha - \beta) \cdot \frac{1}{N} \tag{18}$$

where N is the total number of query papers, and α and β are two hyper-parameters that adjust the weight of each component.

Future PageRank and time

The FPR+T benchmark takes into account papers' publication time by adding a time-related term to compensate for the smaller number of citations of newly published papers. Similar to the normalisation factor used in direct citation metrics, the time term S_T is defined on the following exponential function:

$$S_T(p_i) = e^{-\rho(T_{historical} - t(p_i))} \tag{19}$$

where ρ is a constant set to 0.3, $T_{historical}$ and t denote the historical time point and the publication year of paper p_i respectively, therefore, $T_{historical} - t(p_i)$ represents the time past since publication. Then PR+T is calculated by the following equation:

$$S(p_i) = \alpha \cdot PageRank(p_i) + \beta \cdot S_T(p_i) + (1 - \alpha - \beta) \cdot \frac{1}{N} \tag{20}$$

where N is the total number of query papers, and α and β are two hyper-parameters that adjust the weight of each component.

Summary

Table 1 summarises the benchmarks/test data and evaluation measures used in a collection of existing studies.

From the table, we observe that HCC used to be an important benchmark widely employed to analyse scholarly ranking algorithms. Despite its advantages, it is not friendly to high-quality new papers. FCC and some variants of the HCC were used in more recent studies as they considered the factors which could address the issue of the HCC. In addition, using test data to evaluate ranking algorithms has received increasing attention in recent studies since the awards and recommendations are decided by domain expertise. Besides, it is worth noting that the evaluation measures corresponding to different types of benchmarks are different. Specifically, when comparing ranking results to those generated by citation-based metrics, the Spearman's rank correlation and classification-based measures such as accuracy and ROC curves are often used as evaluation measures. On the other hand, when using test data for evaluation, distribution-based measures such as average and median, and alternative evaluation measures such as normalised discounted cumulative gain and average precision are more popular.

Experiment design

Datasets

Two academic data sources are used in this study, namely the arXiv and Microsoft Academic Graph (MAG). The arXiv has been popular for its use in evaluating paper ranking algorithms. It contains rich bibliometric information collected mainly from the domains of mathematics and physics. The MAG covers a wider range in terms of publication time and research fields of the papers included, which provides an opportunity to analyse recent data in multiple research topics. The use of multiple datasets alleviates potential bias brought by dataset itself and contributes to reliable results and interpretations. Four datasets are collected from these two sources and described as follows.

- *The arXiv* - It contains 29,555 academic papers published from 1992 to 2003 with 352,808 associated citations. We further extracted the author, venue (journal or conference) and publication time information of these papers, involving 14,909 authors and 428 venues.
- *The MAG (ID)* - This dataset was extracted from the MAG, containing 6428 papers published in the field of Intrusion Detection (ID), a secondary subject of Cyber Security. The corresponding bibliometric information was also extracted, including 94,887 citations, 18,890 authors and 720 venues.
- *Test data (CS)* - This test data set was collected from the MAG, containing the papers in the field of computer science from 2000 to 2011. We manually labelled gold stand-

Table 1 A summary of existing ranking algorithms and their associated benchmarks/test data and evaluation measures

Ranking algorithms	Benchmarks/test data	Measures
PageRank (Chen et al., 2007)	HCC	Corr
(Ma et al., 2008)	HCC	Corr
(Yan & Ding, 2010)	HCC	Intermedium
TS-Rank (Li et al., 2010)	FCC	SFD
(Yan et al., 2011b)	FCC	R^2
P-Rank (Yan et al., 2011a)	HCC	PCA, Corr
AAAI13 (Wang et al., 2013)	FCC	Corr
MRFRank (Wang et al., 2014)	FCC	RI
(Xu et al., 2014)	Functional HCC	AP, SFD
CocaRank (Zhang et al., 2016)	FCC	Corr
MRCoRank (Wang et al., 2016)	FCC	RI
PANDORA (Bai et al., 2017)	HCC w/o CoI	Corr, RI
PePSI (Zhang et al., 2018)	HCC	ROC
SARank (Ma et al., 2018)	HCC, FCC	Pairwise accuracy
W-Rank (Zhang et al., 2019a, 2019b, 2019c)	FCC, wFCC	Corr, ROC
(Sidiropoulos & Manolopoulos, 2005)	Awarded papers	Position sum
CoRank (Zhou et al., 2007)	Recommendations	DCG
NewRank (Dunaiski & Visser, 2012)	Awarded papers	% In Top10, Avg.Dist
(Mariani et al., 2016)	Awarded papers	Standard deviation
(Dunaiski et al., 2016)	Awarded papers	MAP
Rescaled PageRank (Mariani et al., 2016)	Awarded papers	ARR, NIR
SARank (Ma et al., 2018)	Recommendations	Pairwise accuracy
MutualRank (Jiang et al., 2016)	Recommendations	nDCG, GAP
(Dunaiski et al., 2018)	Awarded papers	AP, ROC, DCG, nDCG
FutureRank (Sayyadi & Getoor, 2009)	FPR	Corr, DET
(Waumans & Bersini, 2017)	FPR	In-degree

CoI: conflict of interest, *Corr:* Spearman's rank correlation, *DET:* detection error tradeoff,
SFD: Spearman's Footrule distance, R^2: coefficient of determination,
PCA: principal component analysis, *RI:* recommendation intensity, *AP:* average precision,
ARR: average ranking ratio, *NIR:* normalised identification rate,
ROC: receiver operating characteristic curve, *GAP:* graded average precision,
MAP: mean average precision, *nDCG:* normalised discounted cumulative gain

ard papers (ground truth) in this dataset based on their long lasting impact and significant contribution in their respective fields for a certain period of time after publishing. Many conferences in computer science issue awards, such as Test of Time award or Most Influential award, to those high impact papers published 10 years prior to the award year (or longer depending on the conference). The awarding decisions are generally made through nomination and peer-assessment by domain experts, and finally agreed by conference committee panels. Hence we selected 33 reputed computer science conferences, covering artificial intelligence, machine learning, natural language processing, computer vision, software engineering, data mining, programming lan-

guages, databases, information retrieval, etc., and searched the test data (CS) set against their award lists. In this way, we labelled 143 papers which were published from 2000 to 2005 and awarded mainly after 2010. The collection of conferences and the number of papers awarded by each conference are listed in Table A1. Note that the papers which were honourably mentioned or shortlisted in these awards were also labelled to cater for the non-linear process of scientific development (Mariani et al., 2016; Jiang et al., 2016). Covering more fields of study in the gold standard collection was also for this purpose, so that the time distribution of these papers is not uniform over years, which better represents the real development of the domain. The distribution of the yearly number of gold standard papers in test data (CS) is demonstrated in Fig. A1(a).

- *Test data (PH)* - This test data set is composed of two parts. The first part comprises Nobel Prize papers (Li et al., 2019) and Milestone papers in the journal of Physical Review Letters[2]. These papers were awarded due to their long-lived contribution to the field, by either announcing significant discoveries or initiating new research areas, and the awarding decisions were made after years of validating and proving the significance of these papers. The second part contains the references of the Nobel Prize papers. This is inspired by the study where papers cited by popular textbooks and survey papers for a domain or certain topics were considered as gold standard (Jiang & Zhuge, 2019). Being cited by the Nobel Prize papers shows that these papers have gained recognition from the Nobel Prize laureates, which can be considered as being recommended by the domain experts. By taking these recommended papers into account, the gold standard collection in this test data set is assured to conform with the non-linear scientific development as aforementioned. A total of 246 papers were collected, and related information is listed in Table A2. The distribution of the yearly number of gold standard papers in this test data is demonstrated in Fig. A1(b). It is noted the existence of many "influential" papers which may not receive a high number of citations but give direct birth to these gold standard papers (Hu & Rousseau, 2016), however they are not included in this study due to the complexity of accurately identifying and verifying them.

A historical time point, T, was set to simulate the user query time, which divided a dataset into two parts, namely a historical part for running the ranking algorithms and a future part for computing benchmarks defined on the future citation. Considering the publication time range of the datasets, we set the T of arXiv, MAG, test data CS, and test data PH to 1998, 2008, 2005, and 2001, respectively. Details of the four datasets are summarised in Table 2.

Evaluation measures

Three evaluation measures, including the Spearman's rank correlation coefficient, receiver operating characteristic (ROC) curve and the area under the curve, and normalised discounted cumulative gain (nDCG), are adopted in the experiments. These measures evaluate ranking results from the perspectives of mathematical statistics, classification accuracy, and ranking effectiveness, respectively, and have been commonly used to evaluate ranking algorithms. A synergistic use of multiple measures can help obtain more reliable analysis results.

[2] https://journals.aps.org/prl/50years/milestones.

Table 2 Statistics of the four collected datasets

Dataset	ArXiv	MAG (ID)	Test data (CS)	Test data (PH)
#Papers	29,555	6428	647,180	1,483,924
#Gold papers	N/A	N/A	143	246
#Citations	352,808	94,887	2,862,492	6,892,638
#Authors	14,909	18,890	306,343	908,421
#Venues	428	720	5059	2938
Publication time	1992–2003	2000-2017	2000–2011	1981–2011
Evaluation time T	1998	2008	2005	2001
#Papers before T	16,142	3418	262,845	240,356
#Papers after T	13,413	3,010	384,335	1,243,568

Spearman's rank correlation

Spearman's rank correlation coefficient, ρ, measures the strength and direction of association between two paper ranking lists (Myers et al., 2013). It is defined by the following equation:

$$\rho = \frac{\sum_i (R_1(P_i) - \overline{R_1})(R_2(P_i) - \overline{R_2})}{\sqrt{\sum_i (R_1(P_i) - \overline{R_1})^2 (R_2(P_i) - \overline{R_2})^2}} \tag{21}$$

where $\overline{R_1}$ and $\overline{R_2}$ denote the average ranking positions of all papers in these two ranking lists. $R_1(P_i)$ and $R_2(P_i)$ are the ranking positions of paper P_i in the first ranking list of the first period P_1 and the second ranking list of the second period P_2 respectively. In addition, we calculate the confidence interval (CI) corresponding to each coefficient at the significance level of 0.05. The CI is defined on Fisher transformation (Fisher, 1915) as follows:

$$CI = \tanh(\operatorname{arctanh} \rho \pm z_{\alpha/2} / \sqrt{N-3}) \tag{22}$$

where N is the number of papers in the ranking list, and $z_{\alpha/2} = 1.96$ is the two-tailed critical value of the standard normal distribution with a significance level of 0.05 ($\alpha = 0.05$).

Receiver operating characteristic curve

Correlation analysis has its own limitations (West et al., 2010; Thelwall, 2016), therefore we also perform the ROC curves as a supplementary measure. For academic paper ranking, the ROC curve is defined on a classification task that aims to discriminate good-quality or high-impact papers from low-quality or low-impact papers. It visualises the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings (Fawcett, 2006). The calculation of TPR and FPR are based on comparing the ranking results to a selected baseline.

Normalised discounted cumulative gain

Discounted Cumulative Gain (DCG) is an effective measure to evaluate ranking quality in the field of information retrieval (Järvelin & Kekäläinen, 2002). It measures the gain of a paper based on its position in the result rank list using a graded relevance scale of papers in the list. The DCG at a certain position (cut-off value), denoted as DCG@p, is accumulated from the top of the result list to the position p with the gain of each result discounted at lower ranks, as follows:

$$\mathrm{DCG@}p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} \tag{23}$$

where rel_i refers to the graded relevance of the result at position i. We set the relevance of awarded paper to 1 and 0 otherwise. The normalised DCG is computed as:

$$\mathrm{nDCG@}p = \frac{\mathrm{DCG@p}}{\mathrm{IDCG@p}} \tag{24}$$

where $IDCG@p$ denotes the ideal DCG at position p that is calculated as follows:

$$\mathrm{IDCG@p} = \sum_{i=1}^{|REL_p|} \frac{rel_i}{log_2(i+1)} \tag{25}$$

where REL_p represents the list of relevant papers (ordered by their relevance) in the corpus up to position p. The nDCG@p is an effective measure when using test data to evaluate paper ranking results. In this experiment, we calculate the nDCG at different cut-off values from 1 to 300. Papers ranked top will contribute greater to the nDCG@p compared to those behind in the rank, and the decrease in the gain of papers is proportional to the logarithm of their position. A greater nDCG@p value indicates better performance in terms of recommendation correctness.

Experiments

Three experiments are designed. Experiment 1 aims to examine the relationships between different benchmarks. Specifically, six citation count-based benchmarks are analysed, including the HCC, FCC, wFCC, and their time-normalised versions, HCC_t, FCC_t, and wFCC_t. For citation network-based benchmarks, we examine the FPR, FPR+A, FPR+V, and FPR+T. Analysing the relationships between different benchmarks can bring valuable insights into the similarities and differences in their scoring behaviour. Spearman's rank correlation coefficients and ROC curves are used in the analysis.

Experiment 2 aims to examine the relationships between benchmarks and domain expert recommendations in the test data. An important task of assessing academic papers is to identify top quality and high impact papers from large scale paper collections. This task is considerably different from ranking all the papers by granting scores to each of them because top paper recommendation focuses on the few gem publications. Therefore, this experiment simulates the scenario of paper recommendation and is designed to compare the benchmarks against domain expert decisions using test data. The assumption is that decisions made by domain experts are not exclusively based on bibliometric information of

Table 3 Paper ranking algorithms and their features

Algorithm	Features		
	PageRank(PR)- or HITS-based	Time-sensitive	Factors concerned
PageRank (Chen et al., 2007)	PR	×	C
TS-Rank (Li et al., 2010)	PR	√	C
Rescaled PageRank (RS-PageRank) (Mariani et al., 2016)	PR	√	C
CoRank (Zhou et al., 2007)	PR	×	C, A
P-Rank (Zhao et al., 2009)	PR	×	C, A, V
HITS (Ng et al., 2001)	HITS	×	C
FutureRank (Sayyadi & Getoor, 2009)	HITS	√	C, A
MutualRank (Jiang et al., 2016)	HITS	×	C, A, V
AAAI13 (Wang et al., 2013)	HITS	√	C, A, V

C: Citations amongst papers, *A:* Authors of papers, *V:* Venues of publications

the papers, yet these papers should be recognised as high impact measured by some benchmarks, e.g., higher future citation count. The results of this experiment can contribute to bridging between domain expert recommendations and the objective benchmarks. The two test data sets and nDCG are used in this experiment.

Finally, in experiment 3, we analyse nine existing ranking algorithms using test data and different benchmarks to show the impact of evaluation method choices on evaluation results. The nine ranking algorithms and their features are summarised in Table 3.

These algorithms are selected due to their distinguishing features, including whether they are based on PageRank or HITS algorithm, whether they are time-sensitive, and the bibliometric factors considered. Specifically, grouping them into PageRank- and HITS-based algorithms will help reveal whether the evaluation methods favour random walk based or mutual reinforcement based algorithms. Note that we categorised the algorithms based on the core methodology employed in each algorithm, although a few algorithms mixed PageRank scores into the HITS algorithm, such as FutureRank. In addition, comparing time-sensitive and -insensitive algorithms against different benchmarks and test data can uncover the differences between the benchmarks and their time-sensitive variants, as well as the relationship between test data and the two types of benchmarks. Thirdly, evaluating the algorithms which involve diverse combinations of bibliometric factors can demonstrate the differences between citation count-based benchmarks and citation network-based benchmarks. It will also reveal the distinctive paper recommending decisions made by experts compared to the score-based benchmarks. The Spearman's rank correlation coefficients, ROC curves and nDCG are used in these comparative analysis.

Results

Comparison of different benchmarks

The correlation analysis results are summarised in six tables. Specifically, Table 4 lists the Spearman's rank correlation coefficients between citation count-based benchmarks

Table 4 Results of Spearman's rank correlation analysis between citation count-based and citation network-based benchmarks

Results on the arXiv dataset

	FPR	wFPR	FPR+A	FPR+V	FPR+T
HCC	0.28 (0.27,0.3)	0.27 (0.25,0.29)	0.3 (0.28,0.32)	0.34 (0.32,0.36)	0.1 (0.08,0.12)
FCC	**0.83** (0.82,0.83)	**0.82** (0.81,0.83)	0.79 (0.78,0.8)	0.69 (0.68,0.7)	0.77 (0.76,0.78)
wFCC	**0.81** (0.8,0.82)	**0.84** (0.83,0.84)	0.77 (0.77,0.78)	0.67 (0.66,0.68)	0.77 (0.76,0.77)
HCC_t	0.35 (0.33,0.36)	0.34 (0.32,0.35)	0.36 (0.34,0.38)	0.41 (0.38,0.41)	0.27 (0.26,0.29)
FCC_t	0.75 (0.75,0.76)	0.76 (0.75,0.78)	0.72 (0.71,0.73)	0.63 (0.62,0.64)	**0.85** (0.85,0.86)
wFCC_t	0.74 (0.73,0.75)	0.77 (0.76,0.78)	0.7 (0.69,0.71)	0.61 (0.6,0.63)	**0.84** (0.84,0.85)

Results on the MAG (ID) dataset

	FPR	wFPR	FPR+A	FPR+V	FPR+T
HCC	0.45 (0.41,0.5)	0.43 (0.38,0.47)	0.47 (0.42,0.51)	0.41 (0.37,0.46)	0.3 (0.25,0.35)
FCC	**0.84** (0.82,0.86)	**0.82** (0.81,0.84)	0.79 (0.77,0.81)	0.79 (0.77,0.81)	**0.8** (0.78,0.82)
wFCC	0.79 (0.77,0.81)	**0.82** (0.8,0.84)	0.74 (0.71,0.76)	0.75 (0.72,0.77)	0.76 (0.74,0.78)
HCC_t	0.5 (0.46,0.54)	0.47 (0.43,0.52)	0.5 (0.46,0.54)	0.46 (0.41,0.5)	0.38 (0.34,0.43)
FCC_t	0.69 (0.66,0.72)	0.69 (0.66,0.72)	0.64 (0.61,0.67)	0.65 (0.61,0.68)	0.78 (0.76,0.8)
wFCC_t	0.63 (0.6,0.66)	0.67 (0.64,0.7)	0.57 (0.54,0.61)	0.59 (0.55,0.63)	0.72 (0.69,0.74)

Results on test data (CS)

	FPR	FPR+A	FPR+V	FPR+T
HCC	0.64 (0.63,0.64)	0.58 (0.58,0.58)	0.57 (0.57,0.57)	0.33 (0.32,0.33)
FCC	**0.86** (0.86,0.86)	**0.84** (0.84,0.84)	0.78 (0.78,0.78)	0.78 (0.78,0.78)
HCC_t	0.63 (0.62,0.63)	0.58 (0.58,0.59)	0.57 (0.56,0.57)	0.41 (0.4,0.41)
FCC_t	0.74 (0.74,0.74)	0.74 (0.74,0.74)	0.68 (0.67,0.68)	**0.86** (0.86,0.87)

Results on test data (PH)

	FPR	FPR+A	FPR+V	FPR+T
HCC	0.25 (0.25,0.26)	0.23 (0.23,0.23)	0.31 (0.31,0.32)	0.18 (0.18,0.19)
FCC	**0.87** (0.87,0.87)	**0.8** (0.8,0.81)	**0.87** (0.87,0.87)	**0.83** (0.83,0.83)
HCC_t	0.23 (0.23,0.24)	0.21 (0.21,0.22)	0.3 (0.29,0.3)	0.18 (0.18,0.18)
FCC_t	0.51 (0.51,0.51)	0.48 (0.48,0.48)	0.52 (0.52,0.52)	0.69 (0.69,0.69)

The 95% confidence intervals are reported in the brackets, and values greater than 0.8 are highlighted. The header line contains the citation network-based benchmarks, and the citation count-based benchmarks compose the header row

HCC historical citation count, *FCC* future citation count *wFCC* weight future citation count, _t: normalised by the factor of publication time, *FPR* future PageRank, *wFPR* weighted future PageRank, FPR+A, FPR+V, FPR+T future PageRank considering author, venue and time factors

(HCC, FCC, wFCC and their time-normalised versions) and citation network-based benchmarks (FPR, FPR+A, FPR+V, and FPR+T) on the arXiv and MAG datasets as well as two test data sets, respectively. Moreover, Table A3 shows the results of correlation analysis between benchmarks in the citation count-based group, and Table A4 shows that for the citation network-based group.

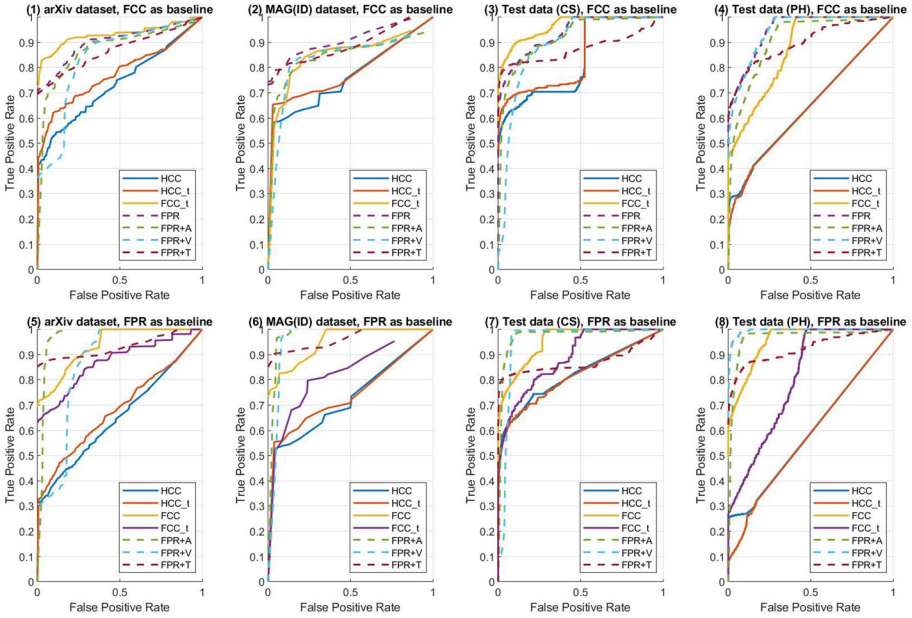


Fig. 4 ROC results on four datasets. The figures in the first row report the ROC of the benchmarks using FCC as baseline, and those in the second row report the ROC of the benchmarks using FPR as baseline. The solid and dashed lines represent citation count-based and citation network-based benchmarks, respectively

In addition, Fig. 4 shows the ROC curves of all the benchmarks on the four datasets. The generation of ROC curves required to assign a baseline to calculate the TPR and FPR, hence the FCC and FPR were selected because they were representatives of the citation count-based benchmarks and citation network-based benchmarks, respectively. Note that the objective of analysing ROC curves is to compare benchmarks, thus the baselines serve the purpose of comparative study, and selection of baseline is not the focus of this analysis.

Based on the results of correlation analysis and ROC curves, the following observations are obtained.

- Comparing between citation count-based and citation network-based benchmarks, as shown in Table 4, we notice that benchmarks defined under the same assumption tend to achieve similar ranking results. For instance, FCC, FCC_t, FPR and FPR+T share the same assumption, that is, future citations matter for paper evaluation. Although they score papers using different approaches, their ranking results are similar in terms of correlation (e.g., greater correlation coefficients as highlighted in the table). This also explains the better performance of the time-awared benchmarks in the ROC results in Fig. 4 where FCC and FPR are used as baselines.
- Comparing between HCC and FCC, it is observed that high HCC does not necessarily imply high FCC. This is reflected by the comparatively lower correlation between the HCC and FCC, also their time normalised versions, as shown in Table A3. Although both of them are based on citation count, they infer the impact of papers

from two different perspectives. HCC considers the current citation status of the papers at the time of the user query, while FCC takes into account new citations received by the papers after the time of user query. Our results confirm that the differences in the underlying assumptions do lead to different ranking results. In addition, the relationship between HCC and FCC is also dependent on the research fields or datasets. In some areas with low timeliness (e.g., theoretical physics and cryptography), high-impact papers published in the past may continue to influence the future and gain high recognition in the future. In contrast, in some other areas, many papers were popular for a period but soon obliterated by time in the river of history.

- Comparing the benchmarks and their weighted versions, the effect of citation relevance is observed. Accumulating citations weighted by relevance takes into account the quality of each citation, and this would be different from directly counting citations, however this difference is small. Our results demonstrate that wFCC achieved different but close ranking results to FCC as shown in Table A3. Similar pattern can be obtained by comparing FCC_t and wFCC_t, and by comparing FPR and wFPR (in Table A4). Considering the extra computational cost brought by citation relevance, we only analysed the wFCC on the arXiv and MAG datasets.
- Comparing the citation count-based benchmarks (HCC, FCC, wFCC) and their time-normalised versions in Table A3, we can find that time normalisation has an impact on the ranking results, and the degree of the impact is related to research fields and datasets (time span of the data and the defined query time). Time normalisation aims to correct the bias introduced by publication time by penalising old papers, which tends to improve the ranking of newly published papers. For some datasets, e.g., test data PH, this could lead to a big difference in the ranking results. This result is also supported by the ROC analysis as shown in Fig. 4(5) to Fig. 4(8) where FPR is considered as baseline. In these four figures, the area under curve (AUC) of FCC is larger than that of FCC_t but the margin in-between their AUC varies in different datasets.
- Comparing FPR+A, FPR+V and FPR+T against FPR, we observe that integrating the paper-author network, paper-venue network, and publication time in the citation network would change the scoring behaviour and lead to different ranking results. The underlying assumption of FPR is that more important papers are likely to receive more citations from future published papers. On this basis, FPR+A and FPR+V further assume that important papers are more likely to be cited by prestigious authors and publication venues, while FPR+T believes newly published papers should be promoted. The difference in their assumptions can explain their different ranking results. In addition, the size of the impact of considering the author, venue, and time factors on ranking results varies according to datasets and research fields. For example, on the arXiv dataset, there is a huge difference between the ranking results of FPR and FPR+V, but this difference is smaller on the MAG dataset. Moreover, the integration of author, venue, and publication time has varying degrees of impact on the ranking results. For example, on the arXiv data set, compared with only using citation information, further integrating venue information leads to a greater impact on the final paper ranks than considering author information.

Our analysis shows similarities and differences between different benchmarks. Since each of them represents a unique perspective, we can collectively use multiple benchmarks to analyse ranking algorithms.

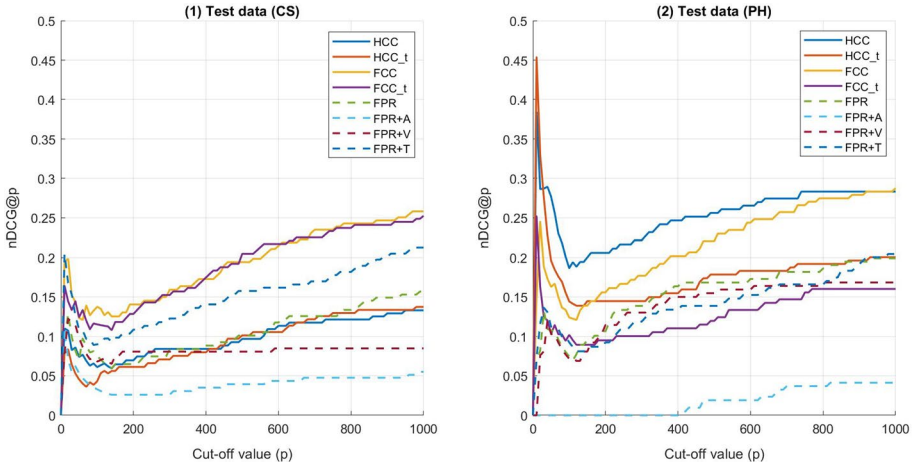


Fig. 5 The nDCG curves of the benchmarks on the two test data sets. Different cut-off values are set to simulate the scenarios where different numbers of papers are recommended. The solid and dashed lines represent citation count-based and citation network-based benchmarks, respectively

Comparison of test data and benchmarks

The nDCG results of the benchmarks on the two test data sets are reported in Fig. 5. We examine each benchmark at different cut-off values (x-axis), which simulates the scenarios where different numbers of papers are recommended. Our results confirmed that the gold standard papers in the test data exhibit the property that is not exclusively based on citations, citation networks and the information of author and venue.

Among all the tested benchmarks, no one is able to identify all the gold standard papers in the test data with high effectiveness. This observation implies that experts do not simply rely on bibliometric information (i.e., citations and authors) to judge the quality or impact of the papers. This is reasonable since domain expert decisions are usually based on peer-assessment of the papers' content. However, the citation status, as well as the author, venue, and publication time information, are still factors that experts may have taken into consideration when selecting awarded papers. This is why using benchmarks can still identify some of the gold standard papers. For example, FCC achieved above 0.2 nDCG at the cut-off value of 600 on both test data sets.

Another finding observed from Fig. 5 is that the extent to which these citation-based benchmarks explain domain expert decisions varies for different fields of research. Specifically, in the test data (CS) the FCC and FCC_t perform better than HCC and HCC_t, however HCC (HCC_t) dominates FCC (FCC_t) in the test data (PH). This result points out how different the fields of computer science and physics develop in terms of paper citation accumulation, which also emphasises the significance of employing multiple test data sets for algorithm evaluation. In addition, the different performance of the citation network-based benchmarks confirms this point as well. The time factor considered in the FPR plays a more significant role in collecting gold papers in the test data (CS) while the venue information does not, however both FPR+T and FPR+V show similar performance in the test data (PH). It is also interesting that the integration of author information does not help this case.

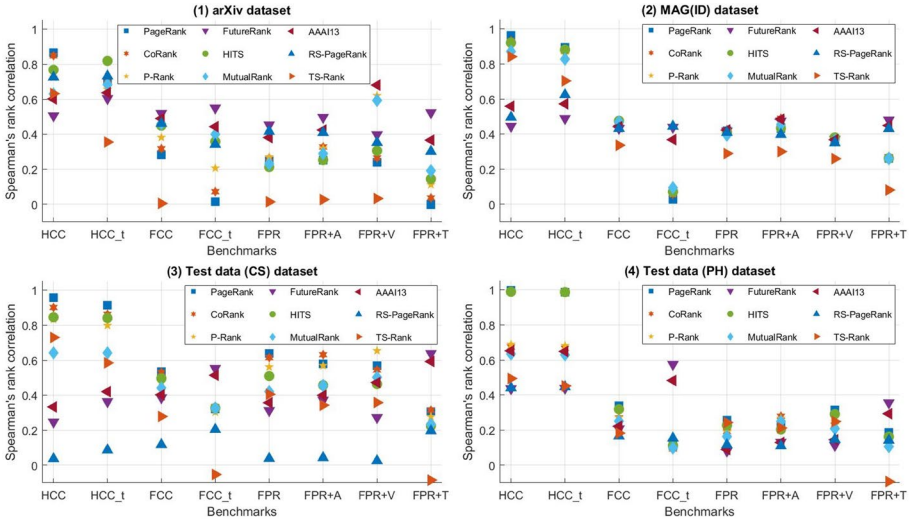


Fig. 6 Spearman's rank correlation coefficients between algorithms and benchmarks on four datasets (time-sensitive algorithms are represented using triangular markers). Different benchmarks are used to analyse the nine selected algorithms. The varying ranks of the algorithms in the columns indicate the difference in evaluation results when using different benchmarks

In summary, citation-based benchmarks cannot fully explain the domain expert decisions in the test data. However, they may share certain properties, or there may be causal relationships. For example, high-quality or high-impact papers recommended by domain experts usually get wide recognition and receive a large number of citations after being awarded, therefore, FCC can identify more awarded papers successfully. Moreover, the nDCG analysis indicates that it is important to employ multiple test data sets in different research areas for the sake of fairness and robustness of algorithm evaluation.

Analysis of ranking algorithms using test data and benchmarks

Results of the Spearman's rank correlation coefficients are summarised in Fig. 6. In addition, Fig. 7 shows the ROC curves of the nine ranking algorithms using each benchmark on the arXiv dataset. Since the calculation of ROC curves of algorithms requires comparing to a baseline, we show the performance of these algorithms using each benchmark as a baseline. The corresponding ROC results on the MAG dataset, test data (CS) and test data (PH) are reported in Online Appendix Fig. A2, Fig. A3 and Fig. A4, respectively. Finally, the nDCG results of the nine algorithms are reported in Fig. 8. Specifically, the calculation of nDCG is based on the relevance of papers. In the test data, the relevance of a paper is defined by whether the paper was awarded (awarded - 1, not awarded - 0).

Results from correlation, ROC and nDCG analysis indicate that using different test data and benchmarks to evaluate and analyse ranking algorithms will lead to different results. From Fig. 6, Fig. 7 and Fig. 8, there is no algorithm that always dominates the others in these three types of analysis. This confirms that assessing paper ranking algorithms from different perspectives can generate diverse results. For example, when using HCC as the benchmark, PageRank shows higher correlation and classification accuracy

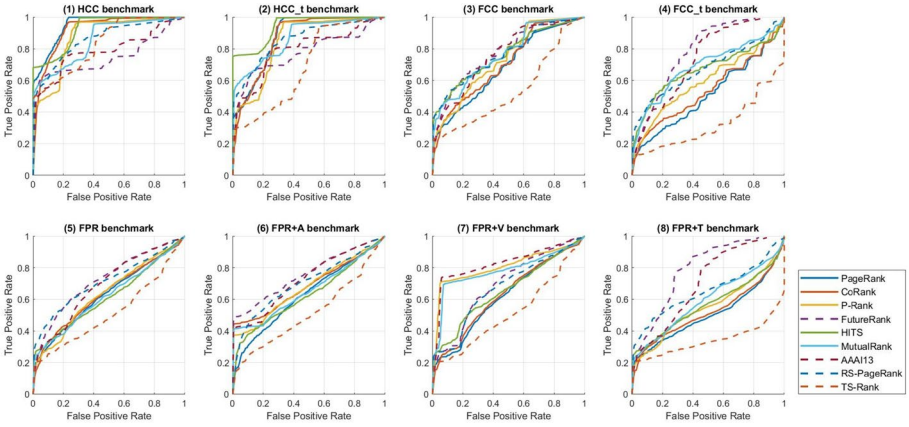


Fig. 7 ROC curves of the nine ranking algorithms using different benchmarks on the arXiv dataset (time-sensitive algorithms are represented using dashed lines). The ROC curves are in different shapes, which indicates the difference in evaluation results when using different benchmarks

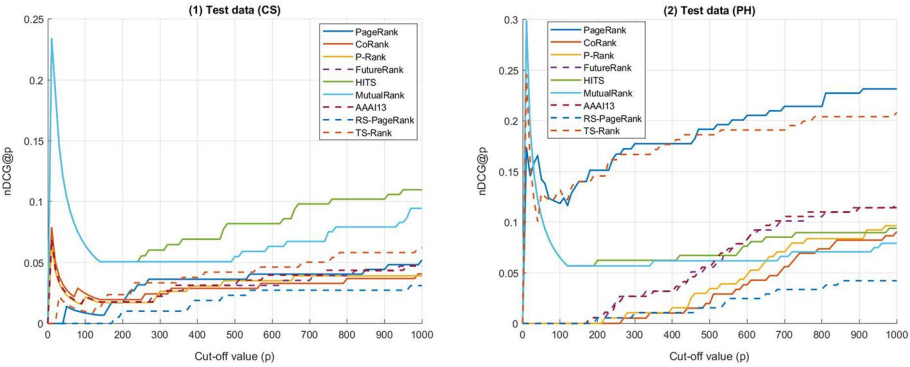


Fig. 8 nDCG curves of the nine ranking algorithms on the two test data sets (time-sensitive algorithms are represented using dashed lines). Different test data is used to evaluate the nine selected algorithms. The nDCG curves are in different shapes, which indicates the difference in evaluation results when using different test data

than FutureRank on four datasets, as shown in the correlation analysis in Fig. 6 and ROC curves in Fig. 7. In contrast, when using FCC as the benchmark, FutureRank demonstrates its advantages over PageRank. These different analysis results are drawn from different perspectives. They are not contradictory, but complementary.

In addition, by comparing between the PageRank-based and HITS-based algorithms, especially between HITS, MutualRank and PageRank, we find no obvious bias from any benchmarks towards these two types of algorithms. However, all of these three algorithms tend to drop performance when time factor is considered in the benchmarks as shown in Fig. 6 and Fig. 7. This makes sense as these algorithms do not take into account the impact of publication time. Interestingly, such drop is smaller for MutualRank compared to HITS and PageRank, which indicates a higher degree of robustness of the MutualRank to perform on different benchmarks. In addition, it can be observed that the time involved benchmarks (HCC_t, FCC_t and FPR+T) favour the time-sensitive algorithms compared to their

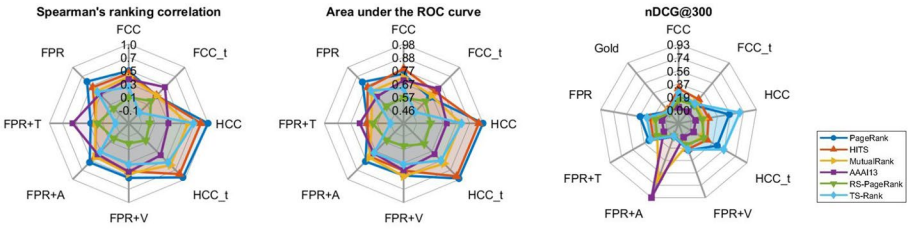
original version (HCC, FCC and FPR), and such case is more obvious when the evaluated algorithms consider multiple bibliometric factors. Specifically, Fig. 6 and Fig. 7 show that FutureRank and AAAI13 perform better on HCC_t, FCC_t and FPR+T benchmarks compared to that on HCC, FCC and FPR benchmarks, yet the time considered benchmarks do not show bias towards RS-PageRank and TS-Rank. Moreover, the algorithms taking into account multiple bibliometric factors tend to obtain better analysis results when these factors are also involved in the benchmarks. For instance, P-Rank shows higher correlation and AUC on the FPR+V benchmark compared to the other PageRank-based algorithms because it integrates the venue information while the others do not.

Another finding is that, the field of research from which the test data set is collected can largely influence the evaluation results of the ranking algorithms. The same ranking algorithm could receive dramatically different nDCG values on different test data sets. This point is clearly evidenced in Fig. 8 where HITS and MutualRank collect a higher number of gold standard papers in the test data (CS) while more gold papers in test data (PH) are found by PageRank and TS-Rank. Furthermore, comparing the nDCG values between the algorithms in Fig. 8 and the benchmarks in Fig. 5, the benchmarks obtain higher nDCG values than the algorithms when cut-off value is smaller than 1000, which means that the benchmarks (e.g., FCC) are able to better explain domain expert decisions in gold paper recommendation. Therefore, analysing algorithms on benchmarks can help build a more comprehensive understanding of the algorithms when using test data alone does not lead to consistent results.

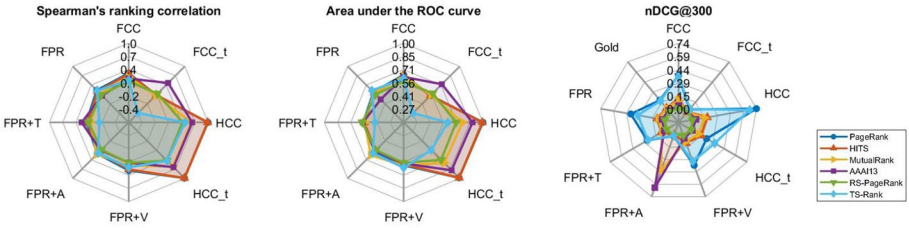
Guideline

Based on results of the three experiments, a guideline for evaluating and analysing paper ranking algorithms is summarised as follows.

- Using appropriate test data is a preferred method to evaluate paper ranking algorithms since the collection of test data is based on peer-assessment and domain expert decisions, thereby it can provide results that are more reliable in terms of the effectiveness of the ranking results. However, collecting appropriate large-scale test data sets is a challenge due to the limitations of relevant data resources. In addition, how to obtain representative, unbiased and sufficient opinions is still an open research question. As a result, the applications of test data in algorithm evaluation are limited by the research fields, publication venues and time.
- Algorithms are usually proposed based on different assumptions and objectives. When the goal of a ranking algorithm is to approximate a certain benchmark, that benchmark should be used to analyse the level of fulfillment for the proposed algorithm. For example, the FutureRank is proposed to predict the future PageRank scores of papers based on their existing data, then it is reasonable to use FPR as benchmark in the analysis. However, it would be not fair to compare with other algorithms on this benchmark since it may favour the algorithm which is proposed under the same assumption. Hence we recommend to employ multiple benchmarks for a fair comparison. For example, if an algorithm is designed to predict the future impact of papers, FCC, FCC_t, as well as FPR and FPR+T can be used together to validate the scoring behaviour of the algorithm.
- Using benchmarks for analysis can shed light on the properties of a ranking algorithm. This is particularly useful when test data is not available. According to the results of



(a) Test data (CS)



(b) Test data (PH)

Fig. 9 Performance profiles of six ranking algorithms on eight benchmarks and two test data sets

the comparative analysis on benchmarks, we recommend to use multiple benchmarks which reflect different characteristics. This is an improvement over the existing common practice where either test data or only one type of benchmark is employed for algorithm evaluation or analysis. A more comprehensive analysis can potentially reduce the bias introduced by applying one single benchmark. The prior knowledge of the benchmarks helps provide a deep understanding of the scoring behaviour of the ranking algorithms and gain valuable insights into their underlying properties.

The following is a walk-through example of how to employ multiple benchmarks for a comparative analysis. Fig. 9 shows the performance profile of the six paper ranking algorithms (PageRank, HITS, MutualRank, AAAI13, RS-PageRank and TS-Rank) in radar charts. Specifically, we analyse the performance of each algorithm on the eight benchmarks (FCC, FCC_t, HCC, HCC_t, FPR, FPR+A, FPR+V, FPR+T) using the two test data sets. Analysis results on the Spearman's rank correlation coefficients, area under the ROC curves, and nDCG@300 are generated separately. For the nDCG values, gold standard papers define the relevance for the Gold indicator in the figures. In addition, we define new relevance by each of the eight benchmarks. To obtain a new relevance, we rank all the papers according to a selected benchmark, and define the relevance of the top K papers as 1 and the rest as 0. This is to simulate the situation that the top K papers are recommended. The K is set to the number of awarded papers in the test data for a fair comparison.

The performance profile of the algorithms on different benchmarks can help identify the properties of the algorithms. Specifically, in each radar chart, the profile of each algorithm forms a unique shape, which displays the performance of this particular algorithm on different benchmarks. If an algorithm covers a larger area toward a benchmark which emphasises on a specific bibliometric factor, it indicates that the scoring behaviour of this algorithm shares similar properties with that of the benchmark. For instance, it is obvious that

PageRank and HITS outperform the other algorithms on the HCC and HCC_t benchmarks in the correlation and ROC analysis, meaning these two algorithms tend to favour the paper with many older citations. In contrast, the AAAI13 algorithm always dominates the others on the FCC_t benchmark, showing its focus on the papers with relatively new citations, which makes it more suitable to search for emerging breakthrough research. In addition, algorithms considering more bibliometric factors for ranking papers do not necessarily obtain higher recommendation effectiveness in terms of identifying gold standard papers.

Conclusion

This study carried out a comprehensive review and investigation on the existing methods used for evaluating and analysing paper ranking algorithms, and then grouped these methods into two main categories (test data and citation-based benchmarks) based on how they generate paper rankings. Extensive comparisons were conducted and the characteristics of the benchmarks were analysed to assess the relationships and differences among these methods. Specifically, three experiments were carried out to respectively investigate the relationships between the all the citation-based benchmarks, compare all the benchmarks against test data, and analyse ranking algorithms using test data and benchmarks. Overall, two academic data sources (arXiv and MAG), three research fields (intrusion detection, computer science and physics) and two test data sets were employed in the investigation. The papers awarded for their long lasting contribution to the field were labelled as gold standard in one test data set, and the Milestone papers, Nobel Prize papers and their references were labelled to construct the other test data.

Our findings confirmed the existence of differences in the analysis results when using test data and different benchmarks with data, which means that relying exclusively on one single benchmark could lead to inadequate analysis results. A guideline was summarised to help choose the evaluation method based on the test data availability and objectives of the evaluated algorithms, and lastly a multi-benchmark approach was suggested for algorithm analysis, which can help build a comprehensive profile for the evaluated algorithms to gain more reliable and holistic insights into their performance. Our systematic review and comparative study revealed the relationships and differences among different benchmarks, and confirmed the impact of using different benchmarks and test data on the evaluation results. This is an importance step towards building up an unbiased and unified benchmark and gold standard for evaluation of ranking algorithms in the field of academic paper ranking.

Despise that a guideline was summarised with demonstration of performance profiles of different ranking algorithms, this study suffers one limitation that it did not provide a conclusive solution to the problem. Our future work will put further effort in designing evaluation procedures and frameworks to support a unified and reliable evaluation of, and a fair comparison of, paper ranking algorithms. One way to achieve such evaluation frameworks is to create a collection of comprehensive and representative test data sets with proposed evaluation metrics (Tax et al., 2015; Dunaiski et al., 2018). However, developing test data sets is always challenging due to the difficulty of access to large scale publications, different research fields, and the ever-growing number of new papers. Another way is to propose an advanced evaluation procedure based on a combination of well-designed benchmarks as discussed in this paper. The summarised guideline suggests the practice of ranking algorithm analysis under different conditions, which serves the first step towards this line of research and aims to advance in the next study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04429-z.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. There is no funding related to this article.

Declarations

Conflict of interest The authors declare that there is no conflicts of interest that are relevant to the content of this article.

References

Ahlgren, P., & Waltman, L. (2014). The correlation between citation-based and expert-based assessments of publication channels: SNIP and SJR vs. Norwegian quality assessments. *Journal of Inforetrics, 8*(4), 985–996.

Bai, X., Lee, I., Ning, Z., Tolba, A., & Xia, F. (2017). The role of positive and negative citations in scientific evaluation. *IEEE Access, 5*, 17607–17617.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Bornmann, L., & Marx, W. (2015a). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics, 9*(2), 408–418.

Bornmann, L., & Mutz, R. (2015b). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222.

Cai, L., Tian, J., Liu, J., Bai, X., Lee, I., Kong, X., & Xia, F. (2019). Scholarly impact assessment: A survey of citation weighting solutions. *Scientometrics, 118*(2), 453–478.

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

Dunaiski, M. (2019). *Using test data to evaluate rankings of entities in large scholarly citation networks (Unpublished doctoral dissertation)*. Stellenbosch University.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). How to evaluate rankings of academic entities using test data. *Journal of Informetrics, 12*(3), 631–655.

Dunaiski, M., & Visser, W. (2012). Comparing paper ranking algorithms. In *Proceedings of the South African institute for computer scientists and information technologists conference* (pp. 21–30).

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.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coeficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507–521.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359–375.

Hu, X., & Rousseau, R. (2016). Scientific in uence is not always visible: The phenomenon of under-cited in uential publications. *Journal of Informetrics, 10*(4), 1079–1091.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20*(4), 422–446.

Jiang, X., Sun, X., Yang, Z., Zhuge, H., & Yao, J. (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.

Jiang, X., & Zhuge, H. (2019). Forward search path count as an alternative indirect citation impact indicator. *Journal of Informetrics, 13*(4), 100977.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM, 46*(5), 604–632.

Lawani, S. M., & Bayer, A. E. (1983). Validity of citation criteria for assessing the influence of scientifific publications: New evidence with peer assessment. *Journal of the American Society for Information Science, 34*(1), 59–66.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2019). A dataset of publication records for Nobel laureates. *Scientific Data, 6*(1), 1–10.

Li, X., Liu, B., & Philip, S. Y. (2010). Time sensitive ranking with application to publication search. In *Link mining: Models, algorithms, and applications* (pp. 187–209). Springer.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management, 44*(2), 800–810.

Ma, S., Gong, C., Hu, R., Luo, D., Hu, C., & Huai, J. (2018). Query independent scholarly article ranking. In *2018 IEEE 34th international conference on data engineering (ICDE)* (pp. 953–964).

Mariani, M. S., Medo, M., & Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics, 10*(4), 1207–1223.

Myers, J. L., Well, A. D., & Lorch, R. F., Jr. (2013). *Research design and statistical analysis*. Routledge.

Ng, A. Y., Zheng, A. X., & Jordan, M. I. (2001). Stable algorithms for link analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 258–266).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. (Technical Report No. 1999-66). Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/.

Pilehvar, M. T., Jurgens, D., & Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 1341–1351).

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation, 47*(4), 919–944.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268–17272.

Ristoski, P., De Vries, G. K. D., & Paulheim, H. (2016). A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International semantic web conference* (pp. 186–194).

Saarela, M., Kärkkäinen, T., Lahtonen, T., & Rossi, T. (2016). Expertbased versus citation-based ranking of scholarly and scientific publication channels. *Journal of Informetrics, 10*(3), 693–718.

Sayyadi, H., & Getoor, L. (2009). Futurerank: Ranking scientific articles by predicting their future PageRank. *Proceedings of the 2009 siam international conference on data mining* (pp. 533–544).

Sidiropoulos, A., & Manolopoulos, Y. (2005). A citation-based system to assist prize awarding. *ACM SIGMOD Record, 34*(4), 54–60.

Tax, N., Bockting, S., & Hiemstra, D. (2015). A cross-benchmark comparison of 87 learning to rank methods. *Information Processing & Management, 51*(6), 757–772.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics, 108*(1), 337–347.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment, 2007*(06), P06010.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391.

Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P. S., & He, Y. (2016). Coranking the future influence of multiobjects in bibliographic network through mutual reinforcement. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(4), 64.

Wang, S., Xie, S., Zhang, X., Li, Z., Yu, P.S., & Shu, X. (2014). Future influence ranking of scientific literature. In *Proceedings of the 2014 SIAM international conference on data mining* (pp. 749–757).

Wang, Y., Tong, Y., & Zeng, M. (2013). Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Twenty-seventh AAAI conference on artificial intelligence*.

Waumans, M., & Bersini, H. (2017). Ranking scientific papers on the basis of their citations growing trend. In *International conference and school on network science* (pp. 89–101).
West, J., Bergstrom, T., & Bergstrom, C. T. (2010). Big Macs and Eigenfactor scores: Don't let correlation coefficients fool you. *Journal of the American Society for Information Science and Technology, 61*(9), 1800–1807.
Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. *IEEE Transactions on Big Data, 3*(1), 18–35.
Xu, H., Martin, E., & Mahidadia, A. (2014). Contents and time sensitive document ranking of scientific literature. *Journal of Informetrics, 8*(3), 546–561.
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. (2011a). P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology, 62*(3), 467–477.
Yan, R., Tang, J., Liu, X., Shan, D., & Li, X. (2011b). Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 1247–1252).
Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T. M., & Peng, Z. (2016). Cocarank: A collaboration caliber-based method for finding academic rising stars. In *Proceedings of the 25th international conference companion on world wide web* (pp. 395–400).
Zhang, J., Xu, B., Liu, J., Tolba, A., Al-Makhadmeh, Z., & Xia, F. (2018). PePSI: Personalized prediction of scholars' impact in heterogeneous temporal academic networks. *IEEE Access, 6*, 55661–55672.
Zhang, Y., Saberi, M., Wang, M., & Chang, E. (2019a). K3S: Knowledge-driven solution support system. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence* (Vol. 33, pp. 9873–9874).
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.
Zhang, Y., Wang, M., Saberi, M., & Chang, E. (2019c). From big scholarly data to solution-oriented knowledge repository. *Frontiers in Big Data, 2*, 38.
Zhao, P., Han, J., & Sun, Y. (2009). P-rank: A comprehensive structural similarity measure over information networks. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 553–562). Association for Computing Machinery.
Zhou, D., Orshanskiy, S. A., Zha, H., & Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Seventh IEEE international conference on data mining (ICDM 2007)* (pp. 739–744).

Authors and Affiliations

Yu Zhang[1] · Min Wang[2] · Morteza Saberi[3] · Elizabeth Chang[1]

Min Wang
maggie.wang1@adfa.edu.au

Morteza Saberi
morteza.saberi@uts.edu.au

Elizabeth Chang
e.chang@adfa.edu.au

[1] School of Business, UNSW Canberra, Northcott Dr, Canberra, ACT 2612, Australia

[2] School of Engineering and Information Technology, UNSW Canberra, Northcott Dr, Canberra, ACT 2612, Australia

[3] Faculty of Engineering and Information Technology, University of Technology Sydney, 15 Broadway, Sydney, NSW 2007, Australia