Improving Citation Network Scoring by Incorporating Author and Program Committee Reputation

Dineshi Peiris, Ruvan Weerasinghe

Abstract— **Publication venues play an important role in the scholarly communication process. The number of publication venues has been increasing yearly, making it difficult for researchers to determine the most suitable venue for their publication. Most existing methods use citation count as the metric to measure the reputation of publication venues. However, this does not take into account the quality of citations. Therefore, it is vital to have a publication venue quality estimation mechanism. The ultimate goal of this research project is to develop a novel approach for ranking publication venues by considering publication history. The main aim of this research work is to propose a mechanism to identify the key Computer Science journals and conferences from various fields of research. Our approach is completely based on the citation network represented by publications. A modified version of the PageRank algorithm is used to compute the ranking scores for each publication. In our publication ranking method, there are many aspects that contribute to the importance of a publication, including the number of citations, the rating of the citing publications, the time metric and the authors' reputation. Known publication venue scores have been formulated by using the scores of the publications. New publication venue ranking is taken care by the scores of Program Committee members which derive from their ranking scores as authors. Experimental results show that our publication ranking method reduces the bias against more recent publications, while also providing a more accurate way to determine publication quality.**

Keywords—**Ranking; Citation Network; Publication Venues; Publications; Publication Authors**

I. INTRODUCTION

The Internet has opened up new ways for researchers to demonstrate research results and share their research findings at a rapid pace than the traditional methods. Today, researchers tend to submit their findings to a wide variety of publication venues such as conferences, journals, and seminars. These publication venues play an important role in the scholarly communication process and the visibility that their work receives. Often researchers might be concerned in knowing about the most important publication venues for publishing their research [1]. However, the selection of publication venues is usually based on the researcher's existing knowledge of the field of his/her discipline [2, 3]. As a result, researchers may not be aware of more appropriate publication venues to which their publications could be submitted.

Manuscript received on 23 Nov 2015. Recommended by Prof. K. P. Hewagamage on 16 June 2016.
This paper is an extended version of the paper "Citation Network Based Framework for Ranking Academic Publications and Venues" presented at the ICTer 2015 Conference.

Dineshi Peiris holds a B.Sc. (Honours) in Computer Science from the University of Colombo School of Computing, Sri Lanka.(e-mail: dineshi.peiris89@gmail.com).
Dr. Ruvan Weerasinghe is a Senior Lecturer at the University of Colombo School of Computing. (e-mail: arw@ucsc.cmb.ac.lk).

On the other hand, Computer Science (CS) is a highly active research area that brings together multiple disciplines such as physics, mathematics, and Life Sciences. The number of publication venues has been increasing continuously,

making it difficult for researchers to be fully aware about the appropriateness of such publication venues [4]. With an abundance of available publication venues, it becomes a very difficult task for new researchers to find exactly what they are looking for or for researchers to keep up to date on all the information [2].

Most of the existing methods to measure the reputation of publication venues use citation count as their chief metric [1]. For journals, among existing methods the most popular citation analysis method is Garfield's Impact Factor (IF) which itself is based on citation counts [5]. The number of citations is not a good individual indicator to measure the quality of publications, since it does not take into account the quality of the citations [6, 7, 8]. In the case of conferences, there are no criteria or consolidated metrics for measuring impact. Unlike some other fields, conferences are essential instruments for the timely dissemination of Computer Science research [9]. As demonstrated in [10], the Computer Science programs follow publication ratio of more than two conference papers per journal paper. In addition, conferences have the precise benefits of giving rapid publication of papers [11]. Therefore, the impact of a publication venue is a key consideration for researchers whether the venue is a journal or a conference [3].

Selecting the most appropriate venue to which to submit a new paper minimizes the risk of publishing in disreputable or fake publication venues. On the other hand, the quality of a publication venue is also important in helping with decisions about awards as well for deciding about scholarships funded by research institutions [12]. If publication venue ranking scores are measured successfully, then researchers can make better decisions about a particular publication venue much quicker based on such a mechanism. There is a significant requirement for an automated process of measuring the publication venue scores to support researchers, so that they can easily recognize the venues in which to publish their research. The findings of this research will definitely be beneficial for the researchers and in return it gives this research a great importance.

In our research, we propose a novel approach for ranking publication venues by considering publication history. We have used a modified version of the PageRank algorithm [13] to generate the scores for publications. We have considered two types of publication venues for which we normally need such information:

1. Known publication venues about which we have historical data
    - For example, publications of previous conference venues in the series with citation and author data

2. New publication venues about which we have little information
    - For such new conferences, we often only have information about the Program Committee (PC). For new journals, we often only have information about the editorial board.

The paper is organized as follows: first, we briefly describe our data sets. Then the major modules of the conceptual approach - citation network construction, out-links and in-links creation, publication score generation, author score estimation, lower citation counts of recent publications smoothing and publication venue ranking are presented in Section II. The results of our experiments on the real datasets obtained from DBLP and EventSeer.net are presented in Section III. Then a survey of the existing approaches which perform academic publication analysis is conducted. The strengths and weaknesses of these approaches are also given in Section IV. Finally, a conclusion is provided in Section V. Some directions for the future research work are also suggested in.

II. Our Approach

Fig. 1 illustrates the architecture of our proposed citation network-based publication venue ranking approach, which consists of an academic database and six major modules: citation network construction, out-links and in-links creation, publication score generation, author score estimation, lower citation counts of recent publications smoothing and publication venue ranking. First of all, data preparation is discussed in detail. Then we explain the design of each module in our proposed approach.

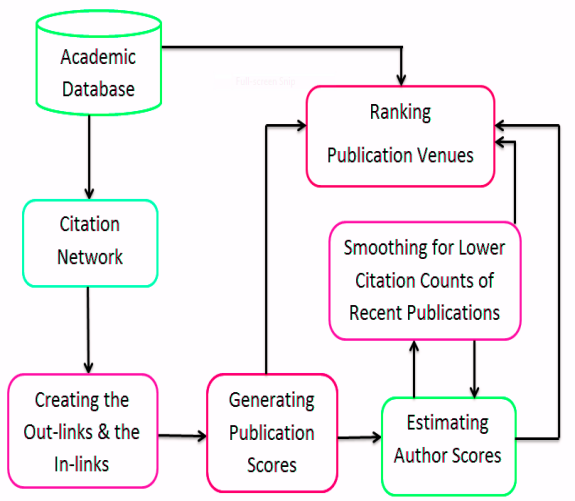

Fig. 1. Architecture of the proposed approach

A. Datasets

We had to work with data that we have access to, which generally are the citation data and the PC/Editorial Board data which may be not easy to get directly through sources such as Google Scholar[1]. Besides, there are diverse digital repositories freely available to the general public. DBLP[2], ACM[3], Microsoft Academic Search[4], and CiteSeer[5] digital libraries are vast collections of citations of past publications. DBWorld[6] and EventSeer.net[7] contain most of CFPs for conferences in Computer Science. From these sources we can collect a list of upcoming and past publication venues with the information about topics and organizations among others.

Our approach used data from two primary sources: DBLP and EventSeer.net. These data sources offer different data services: from DBLP we got XML records, while data from EventSeer.net can only be extracted from its website using a HTML parser. DBLP offers XML records for its dataset which can be download from its website. The DBLP dataset contains information about publications from the numerous fields published over the years. This stores a set of metadata for each publication, including publication title, author(s), type of publication, the year of publication and citations. Each publication is represented by the unique key from DBLP. They did a lot of work into resolving the names problem the same person referenced with many names. Because of that the work in this study relied on the DBLP dataset for author and citation data.

EventSeer.net contains most of the Call for Papers (CFPs) for conferences in Computer Science. Therefore this dataset is essential for our work since it is required in ranking upcoming new publication venues. From EventSeer.net, we collected a list of five new conferences with this information from the listed PC members. Summary statistics of the collected data is shown in Table I.

TABLE I. Summary Statistics of the Collected Data

Data	Quantity
Publications	2645295
Unique Authors	1441289
Conferences	3642
Journals	1345
PC members(within five new conferences)	177

B. Citation Network Construction

Citation networks help in evaluating the impact of publication venues, publications and authors [14]. Citation networks are directed networks in which one publication cites another publication. In most cases, authors cite older publications in order to identify the related body of work or to critically analyze earlier work. Hence citation networks are networks of relatedness on subject matter [15]. On the other hand, publications are well defined units of work and accepted papers play an important part in the success of a publication venue [16]. Our approach is completely based on the citation network represented by publications.

We built the citation network defined as a directed graph, with each publication representing a vertex and the citations representing the edges in the graph; the edges being directed ones, directed from the citing vertex to the cited vertex [1]. Each vertex has several attributes, including publication title, conference/journal of publication, year of publication, author(s) and a unique key from the DBLP dataset.

[1] http://scholar.google.com/
[2] http://www.informatik.uni-trier.de/~ley/db/
[3] http://portal.acm.org/dl.cfm
[4] http://academic.research.microsoft.com/
[5] http://citeseerx.ist.psu.edu/
[6] http://www.cs.wisc.edu/dbworld/
[7] http://eventseer.net/

C. The Ranking Method

The method for ranking publications consists of four phases:

- Creating the publication in-links and out-links
- Using a modified version of the iterative PageRank algorithm to calculate the ranking score for each publication
- Estimating ranking scores for authors using the ranking scores of *stable* publications.
- Smoothing for lower citation counts of recent publications

1) Creating the Publication In-links and Out-links

The method for ranking publications based on the citation network uses the two forms of edges: out-links and in-links.

Definition 1. Out-links: *From a given publication p, link all the publications p_i that the publication p cites.*

Definition 2. In-links: *To a given publication p, link all the publications p_j that cite the publication p.*

2) Generating Publication Scores

According to a class of publication-based ranking methods, the graph vertices represent publications, whereas an edge from node n_i to node n_j represents a citation from publication p_i to publication p_j. Computing the ranking at the publication level has the benefit that only a single procedure is performed to evaluate more than one entity: the publication itself, the publication venue it belongs to, and the author(s) [14]. PageRank offers a computationally simple and effective way to assess the relative importance of publications beyond mere citation counts [6]. Unlike other methods, PageRank constructs a single model that integrates both out-links and in-links [17]. PageRank of a publication is defined as follows [18]:

Definition 3. *Assume publication P has publications R1...Rn which point to it. The parameter d is a damping factor which can be set between 0 and 1. People usually set d to 0.85. C(P) is defined as the number of out-links of publication P. The PageRank of a publication P is given as follows:*

$$PR(P) = (1-d) + d\left(\frac{PR(R_1)}{C(R_1)} + .. + \frac{PR(R_n)}{C(R_n)}\right) \quad (1)$$

PageRank is calculated using an iterative algorithm, and corresponds to the eigenvector of the normalized link matrix [18]. However, damping factor allows for personalization and can make it nearly impossible to deliberately mislead the calculations in order to get a higher ranking score [18]. PageRank extends the idea by not counting citations from all publications equally, and by normalizing its rank by the number of citations on a publication [18]. Another important justification is that a publication can have a high PageRank value if there are many publications that point to it, or if there are publications that point to it which themselves have high PageRank values [18]. PageRank handles both these cases by recursively using the link structure of the citation network.

There are many aspects that contribute to the importance of a publication, such as the number of citations it has received, the rating of the citing publications, the time metric of these citations and its author(s). PageRank only includes the first two factors. The publication environment is not static but changes continuously. PageRank favors older publications because older publications have many citations accumulated over time. Bringing emerging publications to researchers are very important since most of them want the latest valuable information. On the other hand, Aditya Pratap Singh et al. [1] have introduced the timing factor in the PageRank algorithm [13] to reduce the bias against more recent publications which have less time than the older publications to get cited. To make the algorithm time-independent, the metric Aditya Pratap Singh et al. [1] proposed to use is the average of the total number of citations of the publications published in each year. We have also modified the formula for calculation of the PageRank of a publication *P*, to make the algorithm time-independent in this way. This Timed PageRank value of a publication *P* is given by the following:

$$yearY = PY[P]$$

$$TPR(P) = (1-d) + \frac{d * \sum \frac{TPR(R_n)}{C(R_n)}}{AYCC[Y]} \quad (2)$$

where *PY[P]* is the year of publication *P*, *TPR(P)* is the Timed PageRank of *P*, $TPR(R_n)$ is the Timed PageRank of publication R_n that links to publication *P*, $C(R_n)$ is the number of out-links of publication R_n, *AYCC[Y]* is the average number of citations in the year *Y*, and *d* is a damping factor, which is set to 0.85.

3) Alternative Method to Smooth for Lower Citation Counts

Summarizing the weaknesses of the ranking methods we observe that:

- Citation count does not take into account the quality of the citing publications.
- PageRank does not capture the fact that an older publication has more time to be cited in comparison to the recent publications.
- Timed PageRank is able to adjust the rank of emerging quality publications. But it is not sufficient for all the publications since new publications of recent years only have a few or zero citations.

Timed PageRank algorithm is adequate for ranking the publications as it captures the important aspect that an older publication has more time to be cited in comparison to the recent publications. But it is not sufficient for all the publications since new publications of recent years only have a few or no in-links. New publications, which may be of high quality, have a few or no in-links are left behind in this aspect. It is possible the time independent metric of recent publications to become zero.

A study of conferences and journals indicates that many of the references reach back five and more years giving newer publications comparatively little opportunity to get cited [19]. It is possible that the time independent metric of recent publications is zero. Having in mind the above weakness, an alternative method was defined to smooth for lower citation counts of recent publications by modifying the Timed PageRank method.

a) Ranking Scores for Authors

To address the weakness of the Timed PageRank algorithm, we have proposed an alternative metric which uses an author score derived from citations received for publications of that author for previous publications. To assess the quality of a recent publication, its author(s) are useful [20]. It is important to use *stable* publications for calculating scores for authors

since we use these author scores for smoothing the lower citation counts of recent publications.

It is to be noted that the DBLP dataset that we have used only has less citation data after the year 1999 (see Table II). Thus we have taken the year 1999 as the *margin year* to demarcate the *stable* publications and the recent publications. An author score is computed by averaging the Timed PageRank values of all the past publications a given author has written till the year 1999.

The equation for the score of an author A_i is:

$$ARS_{A_i} = \frac{\sum TPRS_{P_{A_i}}}{APC[A_i]} \quad (3)$$

where ARS_{Ai} is the author ranking score, $TPRS_{PAi}$ is the Timed PageRank score of a publication P_{Ai} written by the author A_i and $APC[A_i]$ is the number of publications written by A_i.

TABLE II. AVERAGE NUMBER OF CITATIONS PER PUBLICATION FROM 1999 TO 2014

Year	Average year citation count	Year	Average year citation count
1999	1.547473×10^{-2}	2007	1.231×10^{-5}
2000	2.53011×10^{-3}	2008	5.81×10^{-6}
2001	7.080×10^{-5}	2009	1.073×10^{-5}
2002	0	2010	1.541×10^{-5}
2003	1.034×10^{-5}	2011	2.431×10^{-5}
2004	1.752×10^{-5}	2012	4.68×10^{-6}
2005	1.49×10^{-5}	2013	0
2006	0	2014	0

b) Smoothing for Lower Citation Counts of Recent Publications

Using these *authoritative* scores of authors, we adjust the publication scores after the year 1999. Thus, the score for a new publication is the average score of all the authors of that publication. If this newly calculated publication ranking score is less than the Timed PageRank score of that publication, we will take the Timed PageRank score as the score of the publication.

The equation for the score of a publication P_i is:

$$NPRS_{P_i} = \frac{\sum ARS_{A_{P_i}}}{PAC[P_i]} \quad (4)$$

where $NPRS_{Pi}$ is the new publication ranking score of lower citation count, ARS_{APi} is the author ranking score of an author A_{Pi} who has written the publication P_i and $PAC[P_i]$ is the number of authors who have written the publication P_i.

D. Ranking Publication Venues

In our Adjusted PageRank method, there are many aspects that contribute to the importance of a publication, including the number of citations it has received, the rating of the citing publications, the time metric of these citations and the authors' prior reputation. Besides, computing the ranking at the publication level has the benefit that only a single procedure is performed to evaluate more than one entity: the publication itself, the publication venue it belongs to, as well as the authors of such publications [14]. Hence we can evaluate publication venues based on this Adjusted PageRank scores.

1) Type I: Generating the Scores for Known Publication Venues

The quality of accepted papers plays an important part in determining the success of a publication venue [16]. The ranking score of a publication venue depends on the quality of research papers it publishes [1]. This is the key behind our approach for ranking known publication venues. We have adjusted the publication ranking scores to deal with Computer Science publication venues. Using the Timed PageRank Scores of the publications and the new publication ranking scores of the publications, we formulate scores for publication venues. Known publication venue scores have been formulated by using the scores of the publications.

The equation for the score of a publication venue V_j is:

$$PVRS_{V_j} = \frac{\sum APRS_{P_{V_j}}}{VPC[V_j]} \quad (5)$$

where $PVRS_{Vj}$ is the publication venue ranking score, $APRS_{PVj}$ is the adjusted PageRank score of a publication P_{Vj} in the venue V_j and $V\ PC[V_j\ ]$ is the venue publication count in V_j.

$APRS_{PVj}$ can be either $TPRS_{PVj}$ Timed PageRank score of a stable publication or $NPRS_{PVj}$ New Publication Ranking Score of lower citation count of a publication.

2) Type II: Generating the Scores for New Publication Venues

Adjusted publication ranking scores are not sufficient for all venues because new venues only have PC/Editorial Board data. Research indicates that the quality of a conference is related to that of its PC members [4]. To assess the importance of a new conference, its PC members are useful. As a proof-of-concept, new publication venue scores are generated only for selected conferences. A recent study of PC candidate recommendation shows that the publication history is the strongest indicator for being invited as PC members [16]. New publication venue ranking is taken care by the scores of PC members which derive from their ranking scores as authors.

The score for a PC member is the author score of this person as an author. Earlier we used the publications the author has written till the year 1999 for calculating the author score. Then we adjusted the publication scores. Now we can calculate the author score of the PC using the Adjusted PageRank scores of each of its members as authors. The score for an author is the average score of the Adjusted PageRank values of the publications the author has written. The ranking score for a new conference is the average score of all the PC members of that conference.

The equation for the score of an author A_i is:

$$ARS_{A_i} = \frac{\sum APRS_{P_{A_i}}}{APC[A_i]} \quad (6)$$

where ARS_{Ai} is the author ranking score, $APRS_{PAi}$ is the Adjusted PageRank score of a publication P_{Ai} written by the author A_i and $APC[A_i]$ is the number of publications written by A_i.

The equation for the score of a new conference C_j is:

$$NCRS_{C_j} = \frac{\sum ARS_{C_j}}{CPC[C_j]} \quad (7)$$

where $NCRS_{Cj}$ is the new conference ranking score, ARS_{Cj} is the author ranking score of a PC member in the conference C_j and $CPC[C_j]$ is the conference program committee member count in C_j .

III. Experiments and Results

A. Ranking Publications

We carried out our comparative study mainly based on the studies on academic publication analysis [1, 6, 14, 21]. Most of the existing methods use *Citation Count (CC)* to determine the impact of publications [5, 22, 23]. On the other hand, there has been some work done on academic research using the *PageRank (PR)* algorithm [1, 6, 21], which considers the importance of the citing publication to rank the publication being cited. To integrate the time measurement, we have added a timing factor in the PageRank algorithm named *Timed PageRank (TPR)*. Since our approach has been derived through above mentioned methods, we were able to make a comparison between our *Adjusted PageRank (APR)* method and other mentioned methods.

TABLE III. Ranking Methods

Method	Notation
Citation Count	CC
PageRank	PR
Timed PageRank	TPR
Adjusted PageRank	APR

TABLE IV. Summary of Publication Ranking Methods

Method/Factor	CC	PR	TPR	APR
Number of citations	X	X	X	X
Rating of the citing publications		X	X	X
Time metric			X	X
Smoothing for lower citation counts of recent publications				X

1) Comparison between APR and CC

The following table shows the top 10 publications as determined by our method. Along with the publication APR rank and score, we also show its citation count and its citation rank.

TABLE V. Top 10 Publications in APR Method and Their Citation Ranks

Title	APR		CC	
	Rank	*Score*	*Rank*	*Count*
Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total.	1	1	79	90
Implementing Data Cubes Efficiently.	2	0.93830579	71	95
A Relational Model of Data for Large Shared Data Banks.	3	0.88883433	2	580
Mining Association Rules between Sets of Items in Large Databases.	4	0.77657482	45	111
Fast Algorithms for Mining Association Rules in Large Databases.	5	0.71465106	62	100
Object Exchange Across Heterogeneous Information Sources.	6	0.71062948	108	77
The Entity-Relationship Model - Toward a Unified View of Data.	7	0.59574822	1	604
Relational Completeness of Data Base Sublanguages.	8	0.52836066	18	170
Query Evaluation Techniques for Large Databases.	9	0.51691711	70	95
rganization and Maintenance of Large Ordered Indices.	10	0.50069555	21	153

On analyzing the table, the following key observations were made:

- Citation count, the most common measure of publications, is based on mere citation counts that do not account for the quality of the publications where the citations originate. This table illustrates how accounting for citation origin affects the citation ranking of publications.
- Adjusting for citation origin provided a more refined measure of publication status and changed the publication rankings.

2) Comparison between APR and PR

The following table shows the year-wise contribution in the top 100 publications from both the PageRank and the Adjusted PageRank methods.

TABLE VI. Comparison between PageRank and Adjusted PageRank Method Distributions of the Top 100 Publications

Year	PR	APR	Year	PR	APR
1965	0	1	1992	1	5
1970	1	1	1993	1	9
1971	3	2	1994	1	7
1972	2	2	1995	0	21
1973	0	0	1996	2	13
1974	4	1	1997	0	10
1975	13	0	1998	0	0
1976	9	2	1999	0	1
1977	11	1	2000	0	0
1978	6	1	2001	0	1
1979	9	1	2002	0	1
1980	2	0	2003	0	1

Year	PR	APR	Year	PR	APR
1981	8	0	2004	0	1
1982	4	0	2005	0	2
1983	3	0	2006	0	1
1984	6	2	2007	0	0
1985	1	0	2008	0	0
1986	5	0	2009	0	0
1987	5	0	2010	0	1
1988	0	0	2011	0	4
1989	1	1	2012	0	2
1990	2	2	2013	0	1
1991	0	0	2014	0	2

Fig. 2 shows the variation of the number of publications in the top 100 in both the APR and the PR methods over the years spanning from 1965 to 2014.

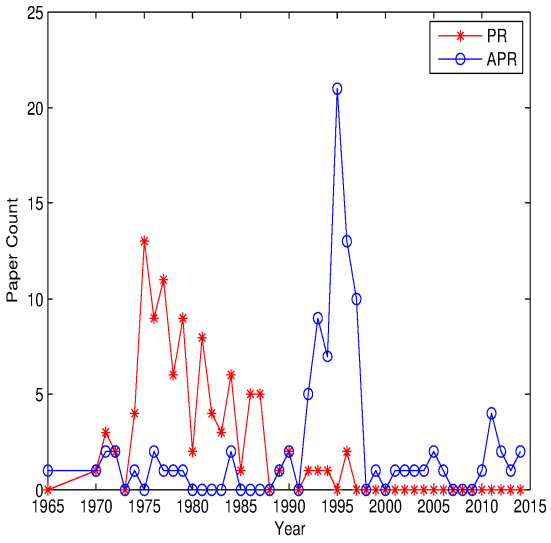

Fig. 2. The number of publications distributed over the years in Adjusted PageRank and PageRank methods

On analyzing the graph, the following key observations were made:

- The top publications in the PR are mostly from 1970s and 1980s whereas in the APR, the top publications are mostly from 1990s and 2000s. This shows that PR favors older publications because older publications have many citations accumulated over time.
- This shows that our method reduces the bias against the recent publications which have less time than older publications to get referenced. Hence it is able to adjust the rank of emerging quality publications.

3) Comparison between APR and TPR

The following table shows the year-wise contribution in the top 100 publications from both the Timed PageRank and the Adjusted PageRank methods.

TABLE VII. COMPARISON BETWEEN TIMED PAGERANK AND ADJUSTED PAGERANK METHOD DISTRIBUTIONS OF THE TOP 100 PUBLICATIONS

Year	TPR	APR	Year	TPR	APR
1965	1	1	1992	7	5
1970	1	1	1993	10	9
1971	2	2	1994	11	7
1972	2	2	1995	24	21
1973	1	0	1996	14	13
1974	2	1	1997	10	10
1975	0	0	1998	0	0
1976	4	2	1999	1	1
1977	1	1	2000	0	0
1978	1	1	2001	0	1
1979	1	1	2002	0	1
1980	0	0	2003	0	1
1981	1	0	2004	0	1
1982	0	0	2005	0	2
1983	0	0	2006	0	1
1984	2	2	2007	0	0
1985	0	0	2008	0	0
1986	0	0	2009	0	0
1987	0	0	2010	0	1
1988	0	0	2011	0	4
1989	1	1	2012	0	2
1990	2	2	2013	0	1
1991	1	0	2014	0	2

Fig. 3 shows the variation of the number of publications in the top 100 in both the APR and the TPR methods over the years spanning from 1965 to 2014.

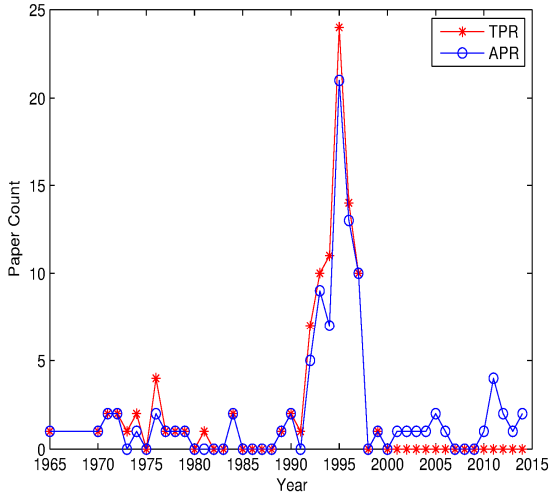

Fig. 3. The number of publications distributed over the years in Adjusted PageRank and Timed PageRank methods

On analyzing the graph, the following key observations were made:

- In the APR method, the publications are distributed over the years as compared to that in the TPR method.
- TPR is able to adjust the rank of emerging quality publications. But it is not sufficient for recent publications which only have a few or zero citations (after 1999). This is clearly visible in the graph as the TPR method is not able to assess the importance of recent publications whereas the APR method is able to assess the importance of recent publications based on their authors.
- To assess the importance of a recent publication, its authors are useful.

For better analysis, we selected few young publications for which citation statistics are not readily available in our dataset, and analyzed them by using their normalized ranking scores in the Time PageRank and the Adjusted PageRank methods over the recent years as shown in Table VIII.

TABLE VIII. NORMALIZED PUBLICATION SCORES IN THE ADJUSTED PAGERANK AND THE TIMED PAGERANK METHODS

Year	Title	Ranking Score	
		TPR Score	APR Score
2006	A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations.	0.06421025	0.34819944
2007	Distributed Resource Management and Admission Control of Stream Processing Systems with Max Utility.	0.06421025	0.20661683
2008	A low-power RF front-end of passive UHF RFID transponders.	0.06421025	0.20620484
2009	A revised r*-tree in comparison with related index structures.	0.06421025	0.24453887
2010	Fair power control for wireless ad hoc networks using game theory with pricing scheme	0.06421025	0.24960894
2011	Permission Re-Delegation: Attacks and Defenses.	0.06421025	0.19908869
2012	Anatomy of a gift recommendation engine powered by social media.	0.06421025	0.24091758
2012	Clickjacking: Attacks and Defenses.	0.06421025	0.19908869
2013	Optimizing budget constrained spend in search advertising.	0.06421025	0.17896617
2014	Thermal design and simulation of automotive headlamps using white LEDs.	0.06421025	0.20620484

On analyzing the table, the following key observations were made:

- Every publication has a PageRank of 0.15 value even though no-one is referencing for it. In the Timed PageRank method, every publication has a ranking score of 0.06420859 when normalizing the 0.15 to scale down the value within the range (0, 1).
- Recent publications, which may be of high quality, have no in-links climbed up their ranking scores when switched from Timed PageRank to Adjusted PageRank method.
- This table shows that our method reduces the bias against the recent publications, which have no in-links.

B. *Ranking Publication Venues*

1) Type I: Known Publication Venues

We relied on our publication ranking method to compute publication venue ranking scores. One approach could be to compute the average score of all their publications. The following table shows the top 10 publication venues by averaging of all their publications.

TABLE IX. TOP 10 PUBLICATION VENUES BY AVERAGING OF ALL THEIR PUBLICATIONS. **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score
IPSJ	C	0.28796447
Electronic Networking: Research, Applications and Policy	J	0.16220377
VLDB workshop on Management of Uncertain Data (MUD)	C	0.08146009
Foundations and Trends in Databases (FTDB)	J	0.08097285
ACM Trans. Database Syst. (TODS)	J	0.08017983
Conference on Very Large Data Bases (VLDB)	C	0.07990088
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.07961964
Science	J	0.07911258

Publication Venue	Type	Score
VIEWS	C	0.07818149
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.07746338

For instance, publication venue A has 30 publications with only 20 being top ranking publications. Assume that these high quality publications have a score of 10 points each, where the remaining ones have a score of 1 point. Publication venue B has in total 5 publications, with 4 publications of them being top ranking publications. It is reasonable to consider that publication venue A should be ranked higher than publication venue B for their scientific contribution, because A has 5 times the number of top ranking publications than publication venue B. If we compute the average of all publication scores, then publication venues A and B would have 7 and 8.2 points respectively. It is not fair to take that approach to compute venue scores.

In order to deal with this problem, we have taken into account the top *n%* of publications to calculate publication venue score. Therefore, our problem was to choose the *n%* of publications of each publication venue that should be considered in the ranking. We performed the following experiment to determine the number *n*. We computed the average score for each publication venue by using their top *n%* publications, $\forall\, n \in \{25, 50, 75\}$. Thus, we produced 3 ranking lists for our publication venue ranking task. As a test bed we used the CORE 2013 Conference Ranking list[8]. In CORE conference ranking, conferences are allocated a rank of A*[9], A[10], B[11] or C[12]. The ratios of A* and A conferences within the top 10 publication venues were calculated, the better the evaluation was considered as the publication venue ranking list. It is to be noted that we have only considered the conferences within the top 10 publication venues to compute the ratio.

The following tables show the top 10 publication venues by averaging the top 25%, 50%, and 75% of publications respectively. Along with our publication venue rank and score, we also show its CORE 2013 Ranking.

TABLE X. TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 25% OF PUBLICATIONS. **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
Foundations and Trends in Databases (FTDB)	J	0.12188608	-
VIEWS	C	0.11511190	-
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.11356147	A*
VLDB workshop on Management of Uncertain Data (MUD)	C	0.11322808	-
Conference on Very Large Data Bases (VLDB)	C	0.11307256	A*
ACM Trans. Database Syst. (TODS)	J	0.11122050	-
ACM SIGMOD Digital Symposium Collection (DISC)	J	0.10751429	-
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.10126516	-
Conference on Parallel and Distributed Information Systems (PDIS)	C	0.10032141	C
Conference on Innovative Data Systems Research (CIDR)	C	0.09893309	A

TABLE XI. TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 50% OF PUBLICATIONS. **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
Foundations and Trends in Databases (FTDB)	J	0.09897220	-
VLDB workshop on Management of Uncertain Data (MUD)	C	0.09815774	-
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.09391697	A*
Conference on Very Large Data Bases (VLDB)	C	09386330	A*
ACM Trans. Database Syst. (TODS)	J	0.09375841	-
VIEWS	C	0.09076796	-
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.09055314	-
Conference on Innovative Data Systems Research (CIDR)	C	0.08719047	A
ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)	C	0.08603544	A*
ACM SIGMOD Digital Symposium Collection (DISC)	J	0.08586227	-

TABLE XII. TOP 10 PUBLICATION VENUES BY AVERAGING THE TOP 75% OF PUBLICATIONS. **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
VLDB workshop on Management of Uncertain Data (MUD)	C	0.08721005	-
Foundations and Trends in Databases (FTDB)	J	0.08706834	-
ACM Trans. Database Syst. (TODS)	J	0.08532985	-
Conference on Very Large Data Bases (VLDB)	C	0.08508535	A*

[8] http://www.core.edu.au/
[9] flagship conference
[10] excellent conference
[11] good conference
[12] other ranked conference venues

Publication Venue	Type	Score	CORE Ranking
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.08475611	A*
Performance and Evaluation of Data Management Systems (ExpDB)	C	0.08391501	-
VIEWS	C	0.08360480	-
Conference on Innovative Data Systems Research (CIDR)	C	0.08026346	A
Workshop on Data Management on New Hardware (DaMoN)	C	0.07970309	-
ACM SIGACT-SIGMOD Symposium on Principles of Database Systems (PODS)	C	0.07965752	A*

According to the tables, Table XIII shows the ratios of A* and A conferences within the top 10 venues by averaging the top 25%, 50% and 75% of publications respectively. Based on this experiment, we concluded that the average of top 50% publications is the most appropriate publication venue ranking list.

The equation for the ratio is:

$$Ratio = \frac{X}{Y} \tag{8}$$

where X is the number of A* and A conferences within the top 10 publication venues, and Y is the number of conferences within the top 10 publication venues

TABLE XIII. THE RATIO OF A* AND A CONFERENCES WITHIN THE TOP 10 PUBLICATION VENUES

	25%	50%	75%
Number of conferences within the top 10 publication venues	7	7	8
Number of A* and A conferences within the top 10 publication venues	3	4	4
Ratio	0.4286	0.5714	0.5

Furthermore, we produce a venue ranking list by using a cut-off of 50 publications to indicate statistical significance. The following table shows the top 10 publication venues which have higher than 50 publications.

TABLE XIV. TOP 10 PUBLICATION VENUES WHICH HAVE HIGHER THAN 50 PUBLICATIONS **TYPE**: WHETHER THE VENUE IS A JOURNAL (J) OR A CONFERENCE (C)

Publication Venue	Type	Score	CORE Ranking
ACM Trans. Database Syst.	J	0.08017982	-
Conference on Very Large Data Bases (VLDB)	C	0.07990088	A*
ACM SIGMOD International Conference on Management of Data (SIGMOD)	C	0.07961964	A*
Conference on Innovative Data Systems Research (CIDR)	C	0.07623364	A
ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)	C	0.07582618	A*
Parallel and Distributed Information Systems (PDIS)	C	0.07480646	C
Journal on Very Large Data Bases (VLDB J.)	J	0.07449637	-
International Workshop on the Web and Databases (WebDB)	C	0.07401554	C
International Conference on Database Theory (ICDT)	C	0.07394727	A
International Conference on Data Engineering (ICDE)	C	0.07238653	A*

2) Type II: New Publication Venues

As a proof-of-concept, new publication venue scores were generated only for following five conferences. Among recent CFPs we have only taken conferences which were stated as their first conference. The following table shows the selected new conference venue details along with their ranking scores.

TABLE XV. NEW CONFERENCE VENUES AND CORRESPONDING RANKING SCORES

Conference Venue	Year	Score
1st International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM)	2015	0.15460470
1st IEEE International Conference on Multimedia Big Data (BIGMM)	2015	0.15156164
1st Biomedical Linked Annotation Hackathon (BLAH)	2015	0.15059667
1st International Conference on Fundamentals and Advances in Software Systems Integration (FASSI)	2015	0.15053273
1st International Conference on Decision Support System Technology (ICDSST)	2015	0.15007116

IV. RELATED WORK

There has been considerable work in the field of academic research. Among existing methods, the most widely adopted method for measuring the quality of publication venues is to use Garfield's IF. This metric uses the publication citations from only the last two years, which neglects the importance of older papers that they cite. On the other hand, it has been criticized for its only dependency on citation counts [7]. As a result, many alternative methods, e.g., h-index [22], g-index [23], and PageRank algorithm [13], have been used to rank venues [24].

Most research work on academic publications uses citation count as the metric. However, metrics like IF, h-index and g-index are based on the citation count, and hence would not give accurate results in all scenarios [1]. The number of citations is not a good individual indicator to measure quality of publications, since it does not calculate the importance of the quality of citations [6]. It is important to look at a metric which

considers the importance of the citing publications to rank the publication being cited.

There has been much interest in applying social network-based methods for generating recommendation and measuring conference quality. A recommender system for academic events and scientific communities based on Social Network Analysis (SNA) is presented in [25]. This work regards on co-authorship and citation networks. The system constructs an academic event participating matrix, based on which similarity between any two researchers is computed. To make recommendations to a target researcher, a group of the most similar researchers is first selected and then the rank of upcoming events is determined by their aggregating ratings.

Zhuang et al. [4] have identified a set of heuristics to automatically determine the quality of the conferences based on characteristics of the PC. This research is completely based on a hypothesis, where the quality of a conference is closely correlated to the reputation of its PC members. The study was unique in the way the authors have brought their views. The heuristics both in combination and isolation have been examined under a classification scheme. In [4], when combined under this scheme, these proposed heuristics achieved a satisfying accuracy in differentiating conferences. These heuristics are also used to rank and recommend conferences. The proposed heuristics rely on the completeness of the list of the PC members. One issue is that a small number of CFPs do not have an entire list of PC members.

There has been some work done on academic research using the PageRank algorithm [1, 6, 21]. Ding et al. [6] used PageRank to rank authors based on the co-citation network. The closest to our work is research work in [1] which uses an efficient approach to rank the papers in various conferences. A modified version of PageRank has been used to rank papers as well as conferences. An important metric in the algorithm which takes the time factor in ranking the papers has been introduced to minimize the bias against new papers which get little time for being cited. Using the year of publication of the papers, the year-wise score for each conference venue has been calculated.

However, the timing factor is not sufficient for all the publications since new publications only have a few or zero citations. Another issue is how to estimate scores for new venues for which citation data are not available using this method. Our work is motivated by this work and takes two further steps. To address the weakness of this method, we have proposed an alternative metric which uses an author score derived from citations received for previous publications of the author. We have also introduced a new way to assess the importance of old and new publication venues.

V. Conclusion and Future Work

We proposed a novel approach for ranking publication venues by considering publication history. The Timed PageRank algorithm is not sufficient for all the publications since new publications of recent years only have a few citations. New publications, which may be of high quality and have a few citations, are left behind in this aspect. To assess the relative importance of recent publications, we have adjusted the Timed PageRank values with its authors' past publication scores. In our approach, there are many aspects that contributed to the importance of a publication, including the number of citations it has received, the rating of the citing publications, the time metric and its authors' reputations. The experimental results indicate that our method reduces the bias against more recent publications, which only have a few citations. The researchers can make better decisions about a particular venue much quicker and easier based on this mechanism.

The DBLP dataset that we have used only have a few or no citation data after the year 1999. Thus we have taken the year 1999 as the margin year to separate the *stable* publications and the new publications. There is definitely room for improvement on the margin year. The proposed margin year relies on the completeness of the citation data. One issue is that our database does not have a complete list of citations. For example, a quality publication may get a lot of citations from scientific domains that are not included in the DBLP dataset. In such cases, it requires further action to harvest citation data before the proposed approach can be applied.

The ranking scores for authors were derived from the publication ranking scores till the year 1999 only. Using the scores of authors, the lower citation counts of recent publications were adjusted by calculating an average score for each publication after the year 1999. The score for a lower citation count publication was taken as the average score of all the authors who have written that research paper. If there was no author score for a particular author, then we would have ignored that author score and take the average score of other authors. On the other hand, if there were no author scores for all the authors of a particular paper then solution would not have been given. Thus we have taken previously measured Timed PageRank value as the score of the publication. Our smoothing method relies on the generated scores of the authors. In such cases, as mentioned earlier, it requires further action to harvest the citation data as well as author data before proposed approach can be applied.

Currently, five CFPs from EventSeer.net were imported into our database. In the future, it would be of interest to add other CFPs for venue ranking problem. EventSeer.net does not offer a structured dataset like that of the DBLP dataset; we have to parse its website to extract the relevant information. Regular expressions could be used to process aspects of the CFPs text. DBLP XML records and EventSeer.net need to be combined in one unique dataset. The problem is to connect these two data sources to provide a unique data repository for publications. Regular expression could be used to match authors' names and join PC members' names in EventSeer.net to DBLP dataset. Various data refining techniques could be applied to make the analysis more precise.

Some other data sources like Google Scholar, CiteSeer and ACM could be integrated into our data repository to make it more complete. Currently, data from DBLP and EventSeer.net is imported into our database. To have better ranking results, we need data from other sources. Publication data gathered from the web by a web crawler is also an interesting development direction.

Acknowledgment

We would like to thank all the staff members of the Network Operating Center of the UCSC for providing a virtual server instance to facilitate our research activities.

References

[1] Singh A. P., Shubhankar K. and Pudi V. (2011). An efficient algorithm for ranking research papers based on citation network. *Data Mining and Optimization (DMO), 2011 3rd Conference on*. IEEE, pp. 88-95.

[2] Luong H., Huynh T., Gauch S., Do P. and Hoang K. (2012). Publication venue recommendation using author network's publication history. *Intelligent Information and Database Systems*. Springer, pp. 426-435.

[3] Luong H. P., Huynh T., Gauch S. and Hoang K. (2012). Exploiting social networks for publication venue recommendations. *KDIR*, pp.239-245.

[4] Zhuang Z., Elmacioglu E., Lee D. and Giles C. L. (2007). Measuring conference quality by mining program committee characteristics. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 225-234.

[5] Garfield E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8): 979-980.

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

[7] Saha S., Saint S. and Christakis D. A. (2003). Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 91(1): 42.

[8] Seglen P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079): 497.

[9] Patterson D. A. (2004). The health of research conferences and the dearth of big idea papers. *Communications of the ACM*, 47(12): 23-24.

[10] Laender A. H. F., de Lucena C. J. P., Maldonado J. C., de Souza e Silva E. and Ziviani N. (2008). Assessing the research and education quality of the top brazilian computer science graduate programs. *ACM SIGCSE Bulletin*, 40(2): 135-145.

[11] Franceschet M. (2010). The role of conference publications in cs. *Communications of the ACM*, 53(12): 129-132.

[12] Martins W. S., Goncalves M. A., Laender A. H. and Pappa G. L. (2009). Learning to assess the quality of scientific conferences: a case study in computer science. *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 193-202.

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

[14] Sidiropoulos A. and Manolopoulos Y. (2006). Generalized comparison of graph-based ranking algorithms for publications and authors. *Journal of Systems and Software*, 79(12): 1679-1700.

[15] Valmarska A. (2014). *Analysis of citation networks*. Diploma Thesis, Faculty of Computer and Information Science, University of Ljubljana.

[16] Han S., Jiang J., Yue Z. and He D. (2013). Recommending program committee candidates for academic conferences. *Proceedings of the 2013 workshop on Computational scientometrics*. ACM, pp. 1-6.

[17] Mihalcea R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 20-23.

[18] Brin S. and Page L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1): 107-117.

[19] Rahm E. and Thor A. (2005). Citation analysis of database publications. *ACM Sigmod Record*, 34(4): 48-53.

[20] Yu P. S., Li X. and Liu B. (2004). On the temporal dimension of search. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, pp. 448-449.

[21] Dellavalle R. P., Schilling L. M., Rodriguez M. A., Van de Sompel H., and Bollen J. (2007). Refining dermatology journal impact factors using pagerank. *Journal of the American Academy of Dermatology*, 57(1): 116-119.

[22] Hirsch J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46): 16 569-16 572.

[23] Egghe L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1): 131-152.

[24] Katerattanakul P., Han B. and Hong S. (2003). Objective quality ranking of computing journals. *Communications of the ACM*, 46(10): 111-114.

[25] Klamma R., Cuong P. M. and Cao Y. (2009). You never walk alone: Recommending academic events based on social network analysis. *Complex Sciences*. Springer, pp. 657-670.