Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Finding the top conferences using novel ranking algorithm

Muhammad Farooq [1], Hikmat Ullah Khan [2, *], Ajmal Shahzad [3], Saqib Iqbal [4], Abubakker Usman Akram [2]

[1]Department of Computer Science, Government College, Rehmatabad, Rawalpindi, Pakistan
[2]Department of Computer Science, COMSATS Institute of Information Technology, Wah, Pakistan
[3]Department of Computer Science, Centre for Advanced Studies in Engineering (CASE), Islamabad, Pakistan
[4]Department of Software Engineering and Computer Science, College of Engineering and Information Technology, Al Ain University of Science and Technology, Al Ain, United Arab Emirates

ARTICLE INFO

Article history:
Received 4 March 2017
Received in revised form
18 May 2017
Accepted 25 May 2017

Keywords:
Semantic cache
Framework
XML

ABSTRACT

An analysis of an academic network reveals interesting insights into the prevailing research domains. The scientometrics analysis of an academic network mainly focuses to find the top authors and top journals using Bibliometircs. There are few work identify the top conferences to help the scholars to know about the current trends in computer science and helps them to participate in the conferences to extend their interactions with other researchers. In this paper, we aim to find the top conferences in the computer science research domain. We present a novel algorithm, ConferencRank by adapting the state of the art Pagerank algorithm to rank the conferences. We use the various features as weights to rank the conferences as well. The proposed algorithm has been applied on a very large data set of the DBLP and the results confirm that the proposed algorithm is helpful to find the top conferences and ranks the conferences in a proper manner. As a future work, we will like to extend this work by finding topic-sensitive ranking of the conferences.

1. Introduction

The scientometrics is the research domain of analysis and measure to find the impact of the prevailing science and technologies. The recent trends in scientometrics have used algorithms to dig into the ever increasing data of the scholarly network. The number of journals, conferences and other academic events has increased tremendously over the last few decades. With the increase in volume of the research field, the debates emerge about the quality of the research work. So, in addition to the publication count, the concept of citation analysis has emerged. Citation analysis is considered as the measure of quality by counting the number of other works which have cited a research publication, and indirectly in this way, we measure the impact or quality of the authors. This prompted the use of self-citing the research work in which the authors tried to cite their own work to enhance their citation score. Then, the concept of the use of the index was introduced. The first index, h-index (Alonso et al., 2009; Mingers, 2009), was introduced that is still regarded as a good metric to measure the quantity as well as the quality of an author's work. Then, a number of indices (Alonso et al., 2009; Alonso et al., 2010; Jin et al., 2007) were introduced which addresses the various aspects, but mainly the goal of all the indices is same, that is, to measure an author's impact in the research community.

The importance of the journal and conferences where authors publish their work attained the focus of the scientometrics experts. Then, in addition to finding the citation analysis of the author, the citation analysis of the journal in which the research paper published has also been considered to find the impact of research works and their authors indirectly. Then, the idea of impact factor (Amin and Mabe, 2003) also introduced which is still regarded as the best indicator of the quality of a paper. The impact factor is measured by calculating the ratio of the publication count of a journal by the citation count in a calendar year. The comparison of impact factors of two journals from different domains still raises questions (Bollen and Sompel, 2008) but within a domain, the impact factor is still being used as an important and perhaps the most important metric to find the significance of a journal.

Now let us shift our focus to the conferences is interesting. The conferences are usually neither

* Corresponding Author.
Email Address: Hikmat.Ullah@ciitwah.edu.pk (H. U. Khan)
https://doi.org/10.21833/ijaas.2017.06.021

analyzed by any index like authors nor measured like impact factor like journals (Kulczycki and Rozkosz, 2017). There is no specific metric to measure the importance of a research conference. Usually, the conferences which are more specific to a topic or small set of topics are regarded as better as compared to those conferences which calls for submission regarding a number of research domains (Khan and Daud, 2016). For instance, a conference related to computer vision, internet of things, cognitive science, and web semantics will attain more attention as compared to a conference of computer science, information technology or communication (Khan and Daud, 2017). The broader the perspective, the lesser will be the expertise level. Also, the quality or rank of the member of the program committee or the editorial board is also an indirect measure of the level of its significance.

In this paper, we take a novel approach by adapting the state of the art ranking algorithm of Pagerank (Page et al., 1999) to rank the top conferences. It is notable to mention here that we first extracted a large data from the DBLP (http://dblp.uni-trier.de/) sources. DBLP is regarded as one of the most highly ranked forums for managing the research publication. Then, we created a directed graph based network of conferences. The direction of the edge is taken from citing conference publication to the cited conference publication. The factors have been assigned weights according to their significance and compare our results. The related work section shares that no such conference ranking algorithm has been proposed earlier so it is a novel contribution in this regard.

This paper has been presented in the following layout. After introduction, the related work reviews research work regarding finding the research conferences. Then, the details are provided how the dataset was prepared. At last, before concluding the paper, the results are discussed evaluating the significance of the proposed algorithm.

2. Related works

We find a number of research works to find the top authors in academic network and ranking journals as well. But, here our only focus is to discuss work regarding finding the top conferences. Cai and Card (2008) analyzed research publications from conferences as well as from journals related to software engineering and found that merely 20 percentage of the conference publications were related to the core concepts of software engineering such as software testing, software quality and software verification. It only analyzed the topics of the papers are related to the core subject or not. Work has also been done to mine influencial bloggers (Khan et al., 2017). The use of Bibliographic tools has also been adopted in the field of expertise mining (Akram et al., 2016). Another study considered as many as 70, 000 research publications for study and found that the authors of the conferences are increasing by 40% per decade. The various research focus to study the trend of conferences in certain part of the world, such as Turkey (Garousi, 2015) and Canada (Garousi and Varma, 2010). They analyzed the role of industry in research, but did not find the importance of the conferences. Fernandes (2014) studied the citation based study to find the top cited papers in the conferences. Vasulescu et al. (2014) analyzed the characteristics of the conferences with respect to authors such as openness to new authors, level of program committee members and prestige within the research community. Faisal et al. (2017) used the co-existence to find the expertise of user in online forums same work has been done to evaluate authors in academic social networks (Yu et al., 2017). Ontology based search is little helpful for retrieval and ranking of journals (Khan et al., 2013).

The computer science conferences have been ranked using the self-organizing maps with dynamic nodes splitting method (Da Silva Almendra et al., 2015). It concludes that the conferences are the best measure of dissemination of the recent research trends. The conferences have been ranking according to their significance using the social network metrics and on the basis of these ranking, the research institutions have been found (Orouskhani and Tavabi, 2016).

The information retrieval domain is based on finding the relevant information and then ranking the content as well. The link based algorithms of HITS (Nomura et al., 2004) is one of the first algorithm to provide link based ranking. One of the limitations of the HITS algorithm is that it gives importance to both outlinks and inlinks (Khan et al., 2013). PageRank algorithm is regarded as the standard algorithm as it is based on relation between Inlinks and Outlinks, More inlinks will mean more importance of the document. The incoming links having high rank contribute more accordingly. The formula for PageRank is given below (Eq. 1):

$$PR(k) = (1 - d) + d \sum_{y \in Q(k)} \frac{PR(y)}{N(y)} \quad (1)$$

In the above equation, d is used to denote damping factor. PR (k) is the Pagerank of node k. PR(Y) is the Pagerank of the node y which is linking the node k. Q(Y) is the out-degree count of the node y. Its downsides are that at startup all the incoming links and outgoing links are treated equally and thus distributing the rank among them in an equal way, while in reality all links are not of same importance, different links have different importance. Another drawback is that it is aimed for a random surfer but every user is not a random surfer, for example it is not always going to provide good results to a researcher. In PageRank the older pages get higher rank than the new one which is also a drawback of PageRank.

In addition to this work, we have also ranked the authors (Farooq et al., 2016) using the Pagerank algorithm. This work uses the DBLP dataset to find the top authors in the field of the computer science.

In addition to these, we have focused on reviewing the feature and network based methods to find top users (Khan et al., 2017), and modelling to find the top users in the blogging community (Khan et al., 2015). A recent study of the conferences, shares the topic sensitive trends in the field of computer architecture only (Martin and Sorin, 2016). It shares that the authors who are also part of the program committee of the conferences are more related to the core topics of the conferences. Another study shares the citation based network analysis and ranks the academic publications and venues. It is notable to mention here that the publication venue refers the type of publisher such as conferences and journal. The ranking of conferences based on state of the art information retrieval algorithms has not been studied so in this paper, we propose a novel algorithm to find the top conferences (Peiris and Weerasinghe, 2015).

3. The proposed algorithm

The proposed algorithm adapts the PageRank algorithm. We first create a social network using DBLP algorithm. The social network is created using a directed graph where a node represents a conference where as an edge represents the link between the two conference. If a conference A cites a research publication of conference B then a link is created from A to B which shows the significance of B. The proposed algorithm is presented as follows (Eq. 2):

$$CR(k) = (1 - \mathrm{d}) + d \sum_{c \in G(k)} \frac{\mathrm{CR(c)}}{\mathrm{N(c)}} \quad (2)$$

Where JR and CR represents the Journal Rank and Conference Rank respectively, and other symbols are similar as defined above.

In addition, we propose the weighted algorithm as well. The proposed algorithm is presented in Eq. 3.

$$CR_{w(c_i)} = (1 - d) * + d \sum_{cx \in M(c_i)} \frac{CRw(c_x)}{L(c_x)} \quad (3)$$

Where $CR_{w(c_i)}$ represents the weighted Conference Rank, $w(c_i)$ represents the weight of ith conference, c_x represents the set of conferences having inlinks to publications published in ith conference, L(c_x) represents the number of outlinks of c_x, c_k represents the set of all the conferences in the network and d is the damping factor having value 0.85. In all the above mentioned equations, h-index, g-index and R-index have been used as weight.

4. DBLP dataset

In this research, DBLP[1] data is used. DBLP is a widely used bibliography portal of computer science listing more than 2.3 million articles (http://dblp.uni-trier.de/xml/ Retrieved March 01, 2016) in October 2013. The downloaded data set is an eXtensible Markup Language (XML) file which has size 1.23GB and has 3818185 publications, 6598 conferences, and 1403 journals. The dataset contains data latest by December 2013. The XML file of the data set is imported in Oracle database by developing an application to convert XML data into database format. It is notable that the DBLP data provided in XML format has been modelled in the BibTex format, which is defined in the Document Type Definition (DTD) files in the same directory as well. For the researchers who want to extract and prepare the data for research, DBLP does not present any restriction on the access of data to element level and also the element order does not matter. Thus the nonsensical child elements in tags such as <editor>, <article>, <author> and other tags can be accessed and data can be extracted. Another easiness provided in the structure of the dataset is that even there are over three million records in DBLP but the no element in the tag hierarchy is deeper than level three. A sample of data set records is given as follows:

```
<article mdate="2002-01-03" key="persons/Codd69">
        <author>E. F. Codd</author>
        <title>Derivability, Redundancy and Consistency of
Relations Stored in Large Data
            Banks.</title>
        <journal>IBM Research Report, San Jose,
California</journal>
        <volume>RJ599</volume>
        <month>August</month>
        <year>1969</year>
        <cdrom>ibmTR/rj599.pdf</cdrom>
        <ee>db/labs/ibm/RJ599.html</ee>
</article>
```

The dataset statistics are given in Table 1.

Table 1: DBLP data set characteristics

Publications Years	1936-2013
# Publications	38,18,185
# Conferences	6,598
# Authors	13,51,586
Avg # Publ. per year	49,587
Avg # Publ. per author	3
Avg # Publ. per conference	579

5. Results discussion

The results of the proposed methods for ranking of conferences are discussed. ConferenceRank is an adapted form of PageRank, in which, both in-links and out-links are used to calculate the rank of a Conference. It is important to mention here that the DBLP covers more six thousand and five hundred conferences worldwide. The conferences are mainly from computer science and software engineering. The core topics of these conferences are related to databases such as data warehouse, data mining, social network analysis and information retrieval. Our study does not focus on the topic sensitive ranking of the conferences. The weights used for weighted ConferenceRank are h-index and g-index. The top ten results of the ConferenceRank and

weighted ConferenceRank are mentioned in Table 2 showing the variations in the results using different weights.

Table 2: The top ten conferences using ConferenceRank and various weights

Conferences	CR	hWCR	gWCR	Variance
SIGMOD	1	1	1	0.00
VLDB	2	2	2	0.00
PODS	3	3	3	0.00
ICDE	4	4	4	0.00
EDBT	5	5	5	0.00
ICDT	6	6	6	0.00
STOC	7	10	7	2.25
OODBS	8	8	10	1.33
BW	9	12	11	1.58
ER	10	7	9	1.58

The top 6 conferences are ranked similar by all algorithms. There is also very small variation in the ranking order of other conferences. It is worth mentioning that the new approach for ranking the top authors in the scholarly network is usually evaluated by comparing the ranking with the h-index or other measures of ranking the authors. Similarly, the approach to find the top journals compares the result with their impact factors, which is a standard measure of the importance of a journal. In case of conferences, no such standard exist, so we cannot compare our results to any parameter. As a result, take the novel approach by finding the top conferences using the state of the art algorithm and apply various weights as well.

To compare the results, we take the variations in the results to depict the overall similarities in the results. The zero variations in the top five results reveal that the top results are accurate and even the use of weights does not alter the ranking order much. Similarly, the variations in the other top k results are not high thus the results are good. Also, it is notable that the top conferences are very prestigious conferences in the world as far as the compute conferences are concerned. For instance, the SIGMOD and VLDB are regarded as the top conferences.

Scatter charts does not only show the trend but also the each element is marked and it helps in overall analysis. Let us now focus on the comparative scattering on the ranking orders of the proposed ConferenceRank (short as CR in charts) algorithm against the use of weights of h-index, represented as hWCR (h-index weighted Conference Rank) and g-index, represented as gWCR (g-index weighted conference rank).

The Fig. 1 represents the scatter plot of Conference Rank versus hWCR and it shows that majority of the ranking are common and the similar results are evident in the Fig. 2 which shows the comparison of ranking orders of the proposed algorithm of ConferenceRank with the hWCR.

Let us now consider another perspective, that is, to compare the significance of the use of weights of h-index and g-index with each other and then compare the results. It is evident from the Fig. 3 that the overall results are similar and we find about the linear growth of the ranking order.

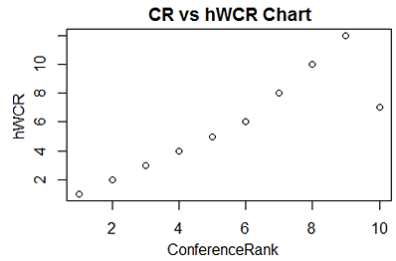

Fig. 1: Top results comparison of the ConferenceRank versus using the h-index as a weight for the proposed algorithm

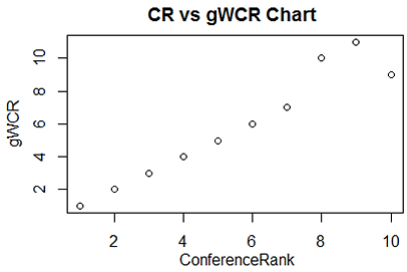

Fig. 2: Top results comparison of the ConferenceRank versus using the g-index as a weight for the proposed algorithm

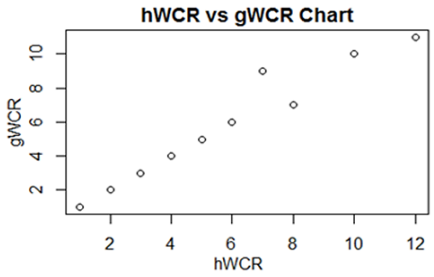

Fig. 3: The comparison of h-index and g-index based eights for ranking the conferences

6. Conclusion

In this paper, we have prepared very large dataset of DBLP, which is world famous computer science bibliography extracting all the records in the bibliography from year 1936 to 2013. The Pagerank, one of the most widely used ranking algorithms, has been adopted to rank the conferences. In addition, the weighted version of Pagerank using the novel weights of h-index and g-index has also been proposed. The results of top k have been discussed. The variations within the results have also been discussed. The top authors are similar, but there are a lot of variations among the lower ranking authors, but there are fewer variations in the results of journals and conferences which shows the significance of our work. In the future, we would like to compare the proposed methods with the difference index schemes and analyze the variations with the ranking orders.

References

Akram AU, Iqbal K, Faisal CMS, and Ishfaq U (2016). An effective experts mining technique in online discussion forums. In the

International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), IEEE, Quetta, Pakistan: 95-99. https://doi.org/10.1109/ICECUBE.2016.7495204

Alonso S, Cabrerizo FJ, Herrera-Viedma E, and Herrera F (2009). H-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics, 3(4): 273-289.

Alonso S, Cabrerizo FJ, Herrera-Viedma E, and Herrera F (2010). Hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. Scientometrics, 82(2): 391-400.

Amin M and Mabe MA (2003). Impact factors: use and abuse. Medicina (Buenos Aires), 63(4), pp.347-354.

Bollen J and Sompel HVD (2008). Usage impact factor: the effects of sample characteristics on usage-based impact metrics. Journal of the American Society for Information Science and technology, 59(1): 136-149.

Cai KY and Card D (2008). An analysis of research topics in software engineering–2006. Journal of Systems and Software, 81(6): 1051-1058.

Da Silva Almendra V, Enăchescu D, and Enăchescu C (2015). Ranking computer science conferences using self-organizing maps with dynamic node splitting. Scientometrics, 102(1): 267-283.

Faisal M, Daud A, and Akram A (2017). Expert ranking using reputation and answer quality of co-existing users. International Arab Journal of Information Technology (IAJIT), 14(1): 118-126

Farooq M, Khan HU, Malik TA, and Shah SMS (2016). A novel approach to rank authors in an academic network. International Journal of Computer Science and Information Security, 14(7): 617-623.

Fernandes JM (2014). Authorship trends in software engineering. Scientometrics, 101(1): 257-271.

Garousi V (2015). A bibliometric analysis of the Turkish software engineering research community. Scientometrics, 105(1): 23-49.

Garousi V and Varma T (2010). A bibliometric assessment of canadian software engineering scholars and institutions (1996-2006). Computer and Information Science, 3(2): 19-29.

Jin B, Liang L, Rousseau R, and Egghe L (2007). The R-& AR-indices: Complementing the h-index. Chinese Science Bulletin, 52(6): 855-863.

Khan HU and Daud A (2016). Finding the top influential bloggers based on productivity and popularity features. New Review of Hypermedia and Multimedia, 22(5) 1-18. https://doi.org/10.1080/13614568.2016.1236151

Khan HU and Daud A (2017). Using machine learning techniques for subjectivity analysis based on lexical and non-lexical features. International Arab Journal of Information Technology, 14(4). Available online at: http://ccis2k.org/iajit/PDF/ Vol 14, No. 4/10038.pdf

Khan HU, Daud A, and Malik TA (2015). MIIB: A metric to identify top influential bloggers in a community. Plos One, 10(9): 1-15.

Khan HU, Daud A, Ishfaq U, Amjad T, Aljohani N, Abbasi RA, and Alowibdi JS (2017). Modelling to identify influential bloggers in the blogosphere: A survey. Computers in Human Behavior, 68: 64-82.

Khan HU, Saqlain SM, Shoaib M, and Sher M (2013). Ontology based semantic search in Holy Quran. International Journal of Future Computer and Communication, 2(6): 570-575.

Kulczycki E and Rozkosz EA (2017). Does an expert-based evaluation allow us to go beyond the Impact Factor? Experiences from building a ranking of national journals in Poland. Scientometrics, 111(1): 417-442.

Martin MM and Sorin DJ (2016). Top picks from the 2015 computer architecture conferences. IEEE Micro, 36(3): 6-9.

Mingers J (2009). Measuring the research contribution of management academics using the Hirsch-index. Journal of the Operational Research Society, 60(9): 1143-1153.

Nomura S, Oyama S, Hayamizu T, and Ishida T (2004). Analysis and improvement of HITS algorithm for detecting Web communities. Systems and Computers in Japan, 35(13): 32-42.

Orouskhani Yand Tavabi L (2016). Ranking research institutions based on related academic conferences. arXiv preprint arXiv:1611.08839. Available online at: https://arxiv.org/pdf/1611.08839

Page L, Brin S, Motwani R, and Winograd T (1999). The PageRank citation ranking: Bringing order to the Web. Technical Report No. SIDL-WP-1999-0120. Stanford University Info Lab, Stanford, USA. Available online at: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Peiris D and Weerasinghe R (2015). Citation network based framework for ranking academic Publications and venues. In the 15th International Conference on Advances in ICT for Emerging Regions, IEEE, Colombo, Sri Lanka: 146-151. https://doi.org/10.1109/ICTER.2015.7377681

Vasilescu B, Serebrenik A, Mens T, van den Brand MG, and Pek E (2014). How healthy are software engineering conferences?. Science of Computer Programming, 89: 251-272.

Yu D, Wang W, Zhang S, Zhang W, and Liu R (2017). A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals. Scientometrics, 111(1): 521-542.