CMPUT 695 - Project Report
Using PageRank to Rank Conferences, Articles, and Authors

Yavar Naddaf
Department of Computing Science,
University of Alberta, Edmonton, Canada

Abstract

This paper describes a method, originally introduced by Bollen et al. [2], to rank conferences, articles, and authors by applying the PageRank algorithm to the citation network. The paper also discusses the methodology used to generate the citation network from Google Scholar. The resulted rankings for a number of well-known conferences in different computer science areas are presnetd for evaluation.

1 Introduction

Many academic and administrative decisions require a measure to rank the quality of research presented in journals, conferences, or individual articles. There is also a need for methods to compare the academic impact of individual or groups of scientists. For instance, a faculty chair, who wants to choose from a number of candidates for a new research position, needs a method to rank the quality of the research done by each candidate. Similarly, a grant committee requires to know the importance and the impact of research done by each group of scientists, in order to objectively divide a funding budget among them. The method used for ranking the quality of research of publications and individuals has a large influence on how researchers are hired and promoted, and how funding is distributed among different research facilities. A more accurate method will result in better scientists being hired and promoted, and influential research projects receiving more funding.

In the field of Computer Science, due to the fast pace of advances in research, and the long publication period of journals, most of the important research is originally published in conferences, rather than journals. Therefore, finding an accurate and subjective ranking method for conferences is of high importance.

This paper presents a short description of the Impact Factor, and a discussion of why it is a measure of popularity, and not prestige, of publications. A method, developed by Bollen *et al.*, is described that ranks journals using a weighted PageRank algorithm and overcomes some of the limitations of the Impact Factor. We demonstrate how the same algorithm can be used to rank conferences. Also, two simple extensions to this method are presented that will allow us to rank articles and authors directly. The paper also describes the methodology used to generate the citation network from Google Scholar, and includes the resulted rankings for some of the well-known conferences in a number of computer science areas.

2 Impact Factor: A measure of popularity

The Impact Factor of a journal in a specific year is the average number of citations that the papers that were published during the previous two years in the journal receive from all the articles published in the given year [2]. For instance, the 2006 Impact Factor of a journal J can be calculated by dividing the number of citations received by the articles published during 2004 and 2005 in J from all the articles published in 2006, over total number of articles published in 2004 and 2005 in J. In other word, the Impact Factor of journal can be interpreted as the average number of citations that each article published in that journal receives in a two years period. The Impact Factor is calculated and published annually by the Institute of Scientific Information (ISI) [1].

One of the criticism of the Impact Factor is that it does not differentiate between citations from highly respected articles and citations from articles with lower status. Intuitively, we would expect that a citation from a ground breaking article that has resulted in a Nobel Prize to carry more importance than a citation from an ordinary article by a graduate student. However, as far as the Impact Factor is concerned, both of these citations contribute equal impact.

Bollen *et al.* compare this characteristic to popularity vs. prestige of books [2]. If we only count the number of readers of a book, we are measuring its popularity. If we also take into account how important the readers of a book are, then we can measure its prestige. For instance, a New York Times bestseller detective book is popular, because many people read it, but it is not necessarily prestigious. On the other hand, an academic book written by an important scholar is read by many other important scholars, and is therefore prestigious. However, it is not necessarily popular.

Bollen *et al.* argue that since the Impact Factor simply counts the number of citations, without taking into account the status of the citing articles, it is really a measure of journal popularity and not journal prestige. Section 3 presents an algorithm, introduced by Bollen *et al.*, that measures the prestige instead of the popularity of publications.

3 PageRank: A measure of prestige

This section will explain how the PageRank algorithm, developed by Page *et al.* to rank the results of a search query [3], can be used to rank the quality of publications. Section 3.1 will provide a short summary of the ideas behind the original PageRank algorithm. Section 3.2 explains how Bollen *et al.* applied the PageRank algorithm to generate a new measure to rank journals. Section 3.3 explains how we can use a similar method to rank conferences, articles, and authors.

3.1 Summary of PageRank algorithm

The PageRank algorithm, developed by Page *et al.*, uses the link structure of the web to rank the results of a search engine. One of the main assumptions behind PageRank is that a webpage that receives many links from other pages is likely to have some material of interest in it. Note that this assumption is similar to what Impact Factor assumes on the academic citation network, *i.e.*, the citation count of an article can be used as a measure of its quality. However, unlike the approach in calculating Impact Factor, the PageRank algorithm does not base the importance of a page only on the number of other pages that link to it, but also how important the linking pages are. This definition results in the following recursive formula for the rank of a page u:

$$Rank(u) = c \sum_{v \in B_u} \frac{Rank(v)}{N_v}$$

where B_u is a set of pages that link to page u, N_v is the number of pages that page v links to, and c is the normalization factor. The idea here is that each page v has some rank, and it distributes its rank uniformly between the pages that it links to. The rank of page u is the sum of all the ranks it receives from pages that links to it[1].

3.2 Journal PageRank

In the paper "Journal Status", Bollen *et al.* demonstrate how the PageRank algorithm can be applied on the journal citation network to generate a new measure for ranking journals [2]. To generate the journal citation network, all articles published in a journal are grouped under a single node. The citations between articles are then transformed to citations between journals. For any two journals v_i and v_j, let $W(v_i, v_j)$ be the number of papers published in v_i that cite a paper published in v_j. The normalized weight of the link from journal v_i to v_j is:

$$w(v_i, v_j) = \frac{W(v_i, v_j)}{\sum_k W(v_i, v_k)}$$

The recursive PageRank formula for journal v_j will be:

$$Rank(v_j) = c \sum_{v_k} Rank(v_k) w(v_k, v_j)$$

[1] Note that a few additional details are required in order to guarantee the convergence of this recursive formula. See the original paper by Page *et al.*[3] for the complete algorithm

Similar to the original PageRank algorithm, we are assuming that each journal has a rank. However, instead of distributing its rank uniformly between the journals that it links to, each journal distributes its rank to the other journals based on the number of papers published in it that cite some paper in the other journal.

3.3 Using PageRank to rank conferences, articles, and authors

This section explains how we can use a similar approach used by Bollen *et al.* to rank conferences, articles, and authors.

3.3.1 Conference PageRank

It is trivially simple to apply an algorithm similar to Journal PageRank to conferences. The only non-trivial requirement is that we need the citation network between the articles published in different conferences. Section 4.1 describes how Google Scholar was used to extract the paper citations. Once we have the citation for n papers, we generate an $n \times n$ matrix A, such that:

$$A[i,j] = \frac{C(i,j)}{\sum_k C(i,k)}$$

were $C(i,j)$ is the number of papers published in conference i that cite a paper published in conference j. We can now use a matrix notation of the PageRank algorithm to find the rank of each conference. Starting from an arbitrary $1 \times n$ ranking vector r, we repeat the following recursive formula, until the change in r is smaller than some predetermined threshold:

$$r = (1-d)rA + dq$$

The first part of the above formula $[(1-d)rA]$ is the recursive rank propagation discussed in section 3.1. The second part $[+dq]$ is a random jump with probability d which guarantees that the recursive formula will converge. For our experiments, we assigned 0.15 to d and a uniform $1 \times n$ vector of $\frac{1}{n}$ to q. With a threshold of $|r_{t+1} - r_t| < 10^{-5}$, the formula always converged within 20 iterations.

Some experts believe that citations from outside a conference should be considered as more important than citations from papers within the same conference. To account for this, we can multiply the diagonal elements of A by a weight $w \leq 1$, and then re-normalize the rows of the matrix. In our experiments, we computed the conference ranks for both $w = 1$, *i.e.,* fully counting self-links, and $w = 0.5$, *i.e.,* a citation from a paper within the same conference has half the importance of a citation from a paper in another conference. The results for both rankings are presented in section 4.2.

3.3.2 Article PageRank

When the Importance Factor or the PageRank of conferences is used to compare the research quality of articles and authors, there is an implicit assumption that the rank of a conference is a fair representative of the quality of research of the papers published in it and their authors. While this assumption is not unreasonable, it may not always be 100% true. it can be argued that neither of the Impact Factor or the Conference PageRank can be used to accurately compare individual articles published in different conferences or their authors. This is because the number of citation received by papers published in the same publication often has a large variance [4, 5]. This high variance invalidates the assumption that just because a conference has a high Impact Factor or PageRank, all papers published in it are heavily cited. To solve this problem, we can use the PageRank algorithm to directly rank articles, and use the resulting ranking system to compare the research impact of different papers. The algorithm is exactly the same as the Conference PageRank algorithm, except that the A matrix holds the citations between papers instead of conferences:

$$A[i,j] = \begin{cases} \frac{1}{N_i} & \text{if article } i \text{ cites article } j \\ 0 & \text{otherwise} \end{cases}$$

where N_i is the total number of articles that i cites.

Aside from allowing us to directly compare different articles, the Article PageRank can also be used to define new ranking methods for conferences. For instance, the rank of a conference can be defined as the average Article PageRank of the papers published in it. Section 4.2 contains the results of ranking a number of well-known conferences using this definition.

3.3.3 Author PageRank

We can generate a direct ranking for authors by applying the same PageRank formula used in Conference PageRank to the author citation network. The author citation network is generated by grouping all papers written by an author under a single node. The citations between articles are then transformed to citations between authors. For m authors, the entries of the $m \times m$ matrix A is defined as:

$$A[i,j] = \frac{C(i,j)}{\sum_k C(i,k)}$$

were $C(i,j)$ is the number of papers written by i that cite some paper written by j. The rest of the algorithm is exactly the same as the Conference PageRank. Section 4.2 contains the resulting rankings for some members of the GAMES group in the University of Alberta.

4 Experimental Results

In order to test our ranking algorithms, we extracted the citation network for 53151 papers published during 2004-2006. These papers were published in 1682 AI, Machine Learning, and Database related conferences. The list of papers and the conference they were published in were taken from DBLP. Section 4.1 explains the methodology used to extract the citation network for these papers from Google Scholar. Section 4.2 demonstrates the resulted rankings for a number of well-known conferences.

4.1 Extracting the citation network from Google Scholar

Google Scholar offers the useful feature of listing all materials that cite a particular article. The problem with this feature is that the list of materials is not restricted to articles only, and includes books, slide shows, and other material. To generate the citation network from Google Scholar, the following procedure was used:

- For every article i that matches our criteria in DBLP:
  - G = List of material that cite i, queried from Google Scholar
  - For every item j in G:
    - Clean up j (e.g. punctuation, Unicode characters, extra spaces)
    - Query DBLP to see if j is a published article.
    - If j is in DBLP, add a citation link from i to j.

In order to get the full citation network, the above procedure was run in parallel on eight machines for about a week.

4.2 Sample Rankings

In order to evaluate the ranking algorithms described in this paper, this section provides the resulted ranks for a number of well-known conferences in various computer science fields (Table 1 - Table 3). The conferences listed in each table are sorted based on the ranking of a human expert. The conference in the top row is the highest ranking conference, and the one in the lowest row is the lowest ranking conference, according to the human expert. The second column of each table contains the Conference PageRank for each conference, in the case that citation within a conference are fully counted for. The third column contains the Conference PageRank, where the self-citations are weighted down by 50%. The fourth column shows the average Article PageRank of the articles published in each conference. The final column contains the 2003 Impact Factor of each conference, as presented on the website CiteSeer.

From these results, we can observe that, for the most part, the Conference PageRank seems to agree with both the human expert rankings, and the 2003 Impact Factor. There are some divergence between the ranking of human expert and the Conference PageRank, *e.g.*, the low Conference PageRank of IJCAI compared to ICML and AAAI. However, in these cases of disagreement between Conference PageRank and the human expert, the Impact Factor usually agrees with the Conference PageRank results. Furthermore, the Conference PageRank with discounted self-link ($w = 0.5$) does not seem to have an advantage over the Conference PageRank with full self-link ($w = 1.0$). Overall, a more through analysis by human experts is

required to determine if the Conference PageRank is a more valid ranking method than simply computing the Impact Factor for conferences.

Finally, to evaluate the Author PageRank algorithm (section 3.3.3), Table 4 contains the resulted rankings for a number of members of University of Alberta GAMES group. While the results seem reasonable to the author, more analysis is required to determine if Author PageRank is a suitable method to rank researchers.

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
SIGMOD	12.0×10^{-3}	11.9×10^{-3}	6.51×10^{-6}	1.74
VLDB	10.5×10^{-3}	10.5×10^{-3}	5.06×10^{-6}	1.52
ICDE	8.13×10^{-3}	8.22×10^{-3}	5.23×10^{-6}	1.25
DEXA	0.14×10^{-3}	0.14×10^{-3}	3.08×10^{-6}	0.27
IDEAS	0.26×10^{-3}	0.27×10^{-3}	3.12×10^{-6}	0.27

Table 1. Ranks for sample Database conferences

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
KDD	3.37×10^{-3}	3.26×10^{-3}	4.72×10^{-6}	1.68
ICDM	0.66×10^{-3}	0.69×10^{-3}	3.32×10^{-6}	0.35
SDM	1.21×10^{-3}	1.26×10^{-3}	4.04×10^{-6}	0.62
ADMA	0.09×10^{-3}	0.09×10^{-3}	3.00×10^{-6}	N/A

Table 2. Ranks for sample Data-Mining conferences

Conference Name	Conference PageRanak (w=1)	Cinference PageRank (w=.5)	Average Article PageRank	Impact Factor (CiteSeer 2003)
IJCAI	2.04×10^{-3}	2.10×10^{-3}	3.68×10^{-6}	1.10
AAAI	2.66×10^{-3}	2.63×10^{-3}	3.44×10^{-6}	1.49
ICML	2.89×10^{-3}	2.55×10^{-3}	4.30×10^{-6}	2.12
Canadian Conference on AI	0.11×10^{-3}	0.10×10^{-3}	3.03×10^{-6}	0.26

Table 3. Ranks for sample AI conferences

Author Name	Author PageRanak
Jonathan Schaeffer	50×10^{-6}
Russell Greiner	31×10^{-6}
Duane Szafron	30×10^{-6}
Michael Buro	11×10^{-6}
Vadim Bulitko	8×10^{-6}
Michael Bowling	8×10^{-6}
Robert Holte	4×10^{-6}

Table 4. Author PageRank for a number of members of the University of Alberta GAMES Group

5 Conclusion

In this paper, we demonstrated how the PageRank algorithm can be used to rank conferences, articles, and authors. While the resulting ranks seem intuitive and reasonable, more analysis is required to determine if these ranking methods are preferable alternatives to the current ranking methods, such as the Impact Factor.

Acknowledgments

The author would like to acknowledge Jiyang Chen for provoding access to his conference cluster in DBLP, and giving helpful hints on how to query Google Scholar. Our implimentation to query Google Scholar is based on an open source procedure, written by Nicolas Roussel, which masks the queries so they seem to be generated from Mozilla on a Windows machine.

References

[1] Sidney Bloch and Garry Walter. The impact factor: time for change. *Australian and New Zealand Journal of Psychiatry*, 35(5):563–568, 2001.

[2] Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. Journal status. *Scientometrics*, 69:669, 2006.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[4] Per O Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079):497–, 1997.

[5] Opthof T. Sense and nonsense about the impact factor. *Cardiovascular Research*, 33:1–7(7), January 1997.