Testing Ranking Algorithms on CiteSeer Data

Dalibor Fiala

Department of Computer Science and Engineering, University of West Bohemia
Univerzitní 8, 30614 Plzeň, Czech Republic
dalfia@kiv.zcu.cz

***Abstract*— This article describes how various ranking algorithms have been tested to evaluate researchers based on the data from a digital library called CiteSeer. We apply five well-known ranking methods such as citation counts, HITS, or PageRank and seven other methods derived from PageRank that take into account not only citation but also collaboration information to assess the importance of individual researchers. We compare the resulting rankings and show that some of them are highly correlated while others are not.**

***Keywords*— CiteSeer, researchers, rankings, PageRank, correlation, convergence**

I. Introduction

CiteSeer [1] was a digital library of mainly computer science papers autonomously collected from the Web [2] that was finally upgraded to CiteSeer^X in 2010. Due to the automated nature of how it gathers its data (crawling the Web, downloading potential research papers in PDF or PostScript format, converting to plain text, parsing, indexing, etc.), these data are prone to errors, which may be the reason why they have been used relatively little in the past informetric research. Nevertheless, there are a few studies which report results based on analyzing CiteSeer data. Apart from our previous [3], [4] and complementary research [5], [6], there have been studies on its citation graph [7]-[10] or other properties [11].

The aim of the analysis described in this article is to present various ranking techniques, some of which have been defined elsewhere by other researchers and are only shortly referred to and some of which have been introduced in our earlier studies and are briefly explained, and comment on the results of their application to CiteSeer citation and collaboration data in terms of correlation and convergence rather than individual ranks.

II. Data and Methods

We analyzed the last freely available CiteSeer data files released in December 2005. The files contained almost 717,000 publication records in a structure similar to XML. In total, there were about 1.8 million citations between these papers. We imported the data into a database and constructed a citation graph of authors and a coauthorship (or collaboration) graph. The author citation graph (or network), the mathematical definition of which will be given below, had some 411,000 nodes (authors) and 4.8 million edges (citations; parallel edges merged). All the experiments we conducted and whose results will be described later were carried out on this author citation graph using also some information from the collaboration network.

Our goal was to assess the importance of researchers who authored the papers indexed by CiteSeer and rank the authors using two first-order methods based on simple citation counts (*Cites* and *InDeg*), which only count citations or in-degree (or weighted and unweighted in-degree, in other words) and ten higher-order methods, which recursively reassign weights to nodes based on the weights of in-linking nodes. There are basically two groups of these algorithms. One is based on Kleinberg's *HITS* [12] and the other one on Google's PageRank (*PR*) by Brin and Page [13]. In our previous work [14], [15], we defined several PageRank variants that combined information from both the citation and collaboration networks with the key concept that citations do not have equal weights and that a citation from a colleague is less valuable than that from a foreign scientist. Thus, a high number of collaborations (common publications) of two authors reduces the weight of a citation between them, but this reduction depends on further factors and may be relaxed substantially, for instance when there are many other coauthors in their common publications. These factors are all based on the collaboration network and are denoted with the following parameters:

- $c_{u,v}$ is the number of common publications by authors u and v (i.e. the number of their collaborations; the variant relying purely on it called COLLABORATION),
- $f_{u,v}$ is the number of publications by author u plus the number of publications by author v (i.e. the total number of publications by those two authors; used in the variant called ALL_PUBLICATIONS),
- $h_{u,v}$ is the number of all co-authors (including duplicates) in all publications by author u plus the number of all co-authors (including duplicates) in all publications by author v (used in the variant ALL_COAUTHORS),
- $hd_{u,v}$ is the number of all distinct co-authors in all publications by author u plus the number of all distinct co-authors in all publications by author v (used in the variant called ALL_DIST_COAUTHORS),
- $g_{u,v}$ is the number of publications by author u where u is not the only author plus the number of publications by author v where v is not the only author (i.e. the total number of collaborations by those two authors; used in the variant called ALL_COLLABORATIONS),
- $t_{u,v}$ is the number of co-authors (including duplicates) in common publications by authors u and v (used in the variant called COAUTHORS),
- $td_{u,v}$ is the number of distinct co-authors in common publications by authors u and v (used in the variant called DIST_COAUTHORS).

If we wish to define the above concepts mathematically, then let $G = (A, E)$ be a directed, edge-weighted graph (author citation graph), A a set of vertices (authors), E a set of edges (citations between authors), and let $(u, v) \in E$ denote an edge from author u to author v, and let us associate a triple of weights $(w_{u,v}, c_{u,v}, b_{u,v})$ with each such edge. $c_{u,v}$ is defined above, $w_{u,v}$ is the number of times author u cites author v (thus, original parallel edges are united into one weighted edge), and $b_{u,v}$ is either equal to zero (resulting in ranking a in the results section) or to one of the earlier described values $f_{u,v}$ (b), $h_{u,v}$ (c), $hd_{u,v}$ (d), $g_{u,v}$ (e), $t_{u,v}$ (f), or $td_{u,v}$ (g) according to what additional information from the collaboration graph we would like to add to the citation graph. The rank score $R(u)$ for author u based on the scores of authors citing him or her is then computed recurrently as follows:

$$R(u) = \frac{1-d}{|A|} + d \sum_{(v,u)\in E} R(v) \frac{\sigma_{v,u}}{\sum_{(v,k)\in E} \sigma_{v,k}}$$

where

$$\sigma_{v,k} = \frac{w_{v,k}}{\frac{c_{v,k}+1}{b_{v,k}+1} \sum_{(v,j)\in E} w_{v,j}}$$

and d is the well-known damping factor we set to 0.9 in our calculations.

Note that if both parameters c and b are equal to zero, the first equation becomes the weighted PageRank formula (resulting in w ranking) and, in addition, if all weights w are set to one, the equation becomes the standard PageRank formula (resulting in *PR* ranking). With non-zero c and b coefficients we call the PageRank variants the bibliographic PageRank [14].

III. Results and Discussion

To compare the various rankings with each other, we show the number of common authors in the top 20 in Table 1 and Spearman’s rank correlation coefficients in Table 2. As we can see in the darker cells, there are three ranking pairs that have 18 or even 19 common authors in the top 20 – *Cites* and *InDeg*, *w* (weighted PageRank) and *a* (COLLABORATION), and *b* (ALL_PUBLICATIONS) and *e* (ALL_COLLABORATIONS). On the other hand, there are a couple of ranking pairs whose top 20 authors are almost entirely distinct and have just three elements in common – *Cites* and *PR*, *HITS* and *PR*, *PR* and *b* (ALL_PUBLICATIONS), and *PR* and *c* (ALL_COAUTHORS). The correlation metric used in Table 1 is very simple and does not reveal anything about the ordering of authors in the rankings. Nevertheless, we are usually interested in the top positions of rankings and, therefore, this kind of similarity measure may sometimes be useful.

The coefficients in Table 2 reflect the whole ordering in the rankings compared and are all significant at the 0.01 level two-tailed. (The number of identical authors in each ranking is 239,629 and authors with tied ranks are sorted alphabetically.) The best positive correlation (i.e. the highest similarity) have rankings *w* (weighted PR) and *a* (COLLABORATION), *w* (weighted PR) and *g* (DIST_COAUTHORS), *b* (ALL_PUBLICATIONS) and *e* (ALL_COLLABORATIONS), and *f* (COAUTHORS) and *g* (DIST_COAUTHORS). The least positively correlated is always *HITS* with *b* (ALL_PUBLICATIONS), *c* (ALL_COAUTHORS), and *e* (ALL_COLLABORATIONS), respectively.

After examining the correlation tables, we may draw three main conclusions:

- Citations and in-degree are quite close to each other as are *w* (weighted PageRank) and *a* (COLLABORATION), *w* and *g* (DIST_COAUTHORS), and *b* (ALL_PUBLICATIONS) and *e* (ALL_COLLABORATIONS).
- *HITS* is distinct from other rankings as far as the whole ranking is concerned whereas *PR* (standard PageRank) is distinct from others when comparing the top 20 authors.
- PR variations correlate very well with the standard PR as for the whole ranking and very badly as for the top 20 authors, which may indicate that adding information from the collaboration graph to the citation graph takes most effect at the top of rankings.

The preceding observations conform to the findings obtained from the analysis of data from another digital library DBLP [14] and suggest that the ranking methods behave roughly the same way when applied to small graphs (DBLP) and to much larger ones (CiteSeer). Convergence rates of all PageRank variations are shown in Fig. 1. As we can see, there are no evident differences between the methods. Each recursive algorithm computed in an iterative way converges in roughly 20 iterations. The convergence criterion is, again, the Spearman’s rank correlation coefficient (Y axis) between the current ranking and the ranking in the previous iteration. The actual PageRank values are not taken into account.

As far as name disambiguation and/or unification is concerned, we disambiguated neither publications nor authors. Strictly said, we did not unify publication titles or author names. Nor did we do any other data cleansing such as removing nodes wrongly labelled as publications or authors, etc. To do all of this manually in a graph with millions of nodes and edges is virtually impossible with limited human resources and within a reasonable time frame. Then, an intuitive approach would be to (randomly) choose a fraction of the original graph, create rankings using the above methods, perform a (semi) manual disambiguation (or cleansing) in that graph fraction, create rankings again, and compare the previous rankings with the current ones. The resulting correlation or precision/recall ratio of the rankings before and after cleansing should help us predict, how far the results yielded from the whole original graph are from the “true” results if the whole original graph was cleansed. Moreover, to improve the prediction, we should repeat the experiment with several graph fractions of various sizes to be able to extrapolate the precision/recall curve to the size of the original graph.

TABLE I

COMMON ELEMENTS IN TOP 20 AUTHORS OF DIFFERENT RANKINGS

	Cites	**InDeg**	**HITS**	**PR**	**w**	**a**	**b**	**c**	**d**	**e**	**f**	**g**
Cites	X	18	13	3	7	6	9	9	10	10	9	7
InDeg	18	X	13	4	8	7	9	9	10	10	10	8
HITS	13	13	X	3	6	5	6	5	6	6	7	6
PR	3	4	3	X	15	16	3	3	4	4	11	13
w	7	8	6	15	X	19	7	6	8	7	16	18
a	6	7	5	16	19	X	6	6	7	7	15	17
b	9	9	6	3	7	6	X	17	17	18	8	8
c	9	9	5	3	6	6	17	X	17	17	8	8
d	10	10	6	4	8	7	17	17	X	17	10	10
e	10	10	6	4	7	7	18	17	17	X	8	8
f	9	10	7	11	16	15	8	8	10	8	X	17
g	7	8	6	13	18	17	8	8	10	8	17	X

TABLE III

SPEARMAN'S RANK CORRELATION COEFFICIENTS OF ALL RANKINGS

	Cites	**InDeg**	**HITS**	**PR**	**w**	**a**	**b**	**c**	**d**	**e**	**f**	**g**
Cites	X	0.9959	0.9557	0.8969	0.8921	0.8918	0.8853	0.8760	0.8870	0.8859	0.8918	0.8919
InDeg	0.9959	X	0.9567	0.8996	0.8917	0.8914	0.8853	0.8766	0.8867	0.8859	0.8913	0.8914
HITS	0.9557	0.9567	X	0.8499	0.8350	0.8351	0.8213	0.8069	0.8250	0.8222	0.8337	0.8345
PR	0.8969	0.8996	0.8499	X	0.9943	0.9940	0.9841	0.9734	0.9864	0.9851	0.9930	0.9937
w	0.8921	0.8917	0.8350	0.9943	X	0.9996	0.9898	0.9800	0.9916	0.9907	0.9983	0.9994
a	0.8918	0.8914	0.8351	0.9940	0.9996	X	0.9881	0.9779	0.9902	0.9891	0.9974	0.9987
b	0.8853	0.8853	0.8213	0.9841	0.9898	0.9881	X	0.9954	0.9979	0.9996	0.9925	0.9910
c	0.8760	0.8766	0.8069	0.9734	0.9800	0.9779	0.9954	X	0.9937	0.9947	0.9843	0.9820
d	0.8870	0.8867	0.8250	0.9864	0.9916	0.9902	0.9979	0.9937	X	0.9983	0.9947	0.9932
e	0.8859	0.8859	0.8222	0.9851	0.9907	0.9891	0.9996	0.9947	0.9983	X	0.9933	0.9919
f	0.8918	0.8913	0.8337	0.9930	0.9983	0.9974	0.9925	0.9843	0.9947	0.9933	X	0.9990
g	0.8919	0.8914	0.8345	0.9937	0.9994	0.9987	0.9910	0.9820	0.9932	0.9919	0.9990	X

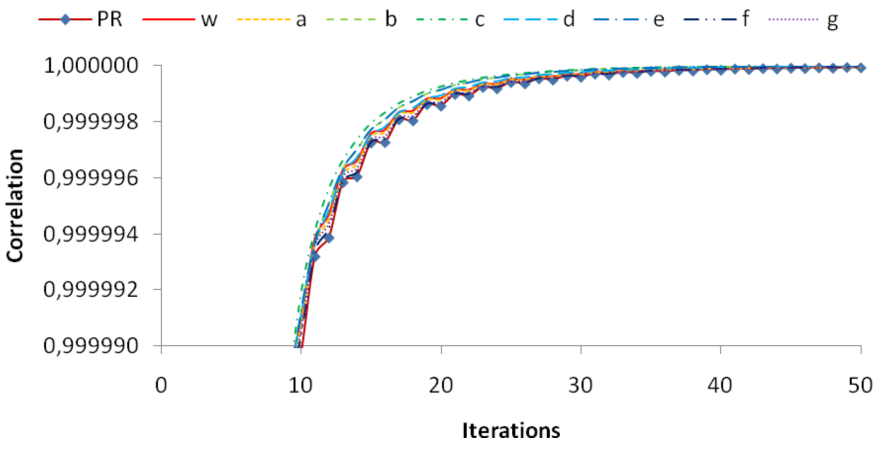

Fig. 1 Convergence of standard (PR), weighted (w) and bibliographic (a – g) PageRank

Of course, the fractional graph sizes must always stay within the limits of what can be done manually. Thus, from the graph of citations between publications, we randomly picked up subgraphs with 100, 500, 1,000, and eventually 5,000 and 10,000 nodes. Unfortunately, the numbers of edges (citations) within the subgraphs were 1, 3, 13, 155, and 474, respectively. In other words, the connectivity in the subgraphs is poor. They are too sparse for rank computations and their comparison with the original graph would be misleading. (Let us recall that the original graph has some 717,000 nodes and 1.8 million edges, i.e. roughly 2.5 edges for a node.) Author citation graphs generated from the publication citation subgraphs would also come up poorly as for their connectivity. Adding more nodes to the subgraphs using citations would be unfair as CiteSeer also has a closed set of nodes and it has no citations outside this set. We simply need subgraphs that are compatible with the original graph.

This leads us to the concept of synthetic graphs. The idea is to automatically generate a graph similar to the original publication citation graph in terms of the number of nodes, edges, degree distribution, etc., create rankings (via author citation graph), change the publication citation graph a little bit (i.e. inject errors to the extent CiteSeer is supposed to do), create rankings again, and compare the previous results with the current ones. Based upon this, it will be possible to make an estimate of the precision/recall loss in the rankings we present in this paper. And how many errors does CiteSeer inject in the real data? In our first experiments, we semi-manually found 0, 6, 24, 146, and 325 duplicate publication titles in the subgraphs with 100, 500, 1,000, 5,000, and 10,000 nodes. Also, we found 4, 47, and 108 false publications (e.g. author or venue names instead of publication titles or meaningless publication titles, etc.) in the graphs with 100, 500, and 1,000 nodes. These experiments must be extended and performed with the author citation graphs as well but can lead to a first rough estimate – 10% of CiteSeer data are false. However, their meaningfulness is shown in [3] where, with a few exceptions, we could not see anything in contradiction with the common sense.

IV. Conclusions and Future Work

In this paper, we described our experiments with various ranking algorithms that we used to evaluate the significance of individual researchers as may be determined on the basis of citation and collaboration networks acquired from a digital library of research papers called CiteSeer. Rather than in the ranks of researchers, however, we were interested in the properties of the different ranking methods and in the relationships of the resulting rankings.

The main contributions of this research are as follows:

- We created citation and collaboration networks of authors from the collection of papers indexed by CiteSeer.
- We applied 12 different ranking algorithms to the data to determine the importance of individual researchers.
- We compared the ranking methods and the corresponding author rankings in terms of convergence and correlation.

The key results we achieved were the following:

- Citations and in-degree-based rankings are quite close to each other in terms of both the shared top 20 authors and the correlation of the whole rankings.
- PageRank variations correlate very well with the standard PageRank as for the whole ranking and very badly as for the top 20 authors, which may indicate that adding information from the collaboration graph to the citation graph takes most effect at the top of rankings.
- As far as the convergence rate is concerned, there are no evident differences between the PageRank-based methods.

In our future work, we would like to concentrate on the successor of CiteSeer, CiteSeer^X, and analyze its citation data by means of ranking methods as it has been already done in [6] for its collaboration network.

Acknowledgment

This work was supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

References

[1] (2010) CiteSeer website. [Online]. Available: http://citeseer.ist.psu.edu

[2] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing", *IEEE Computer*, vol. 32, no. 6, pp. 67-71, 1999.

[3] D. Fiala, "Mining citation information from CiteSeer data", *Scientometrics*. vol. 86, no. 3, pp. 553-562, 2011.

[4] D. Fiala, "Bibliometric analysis of CiteSeer data for countries", *Information Processing and Management*, vol. 48, no. 2, pp. 242-253, 2012.

[5] D. Fiala, "Extracting information from CiteSeer's textual data", *Journal of Theoretical and Applied Information Technology*, (in press).

[6] D. Fiala, "From CiteSeer to CiteSeer^X: Author rankings based on coauthorship networks", *Journal of Theoretical and Applied Information Technology*, (in press).

[7] Y. An, J. Janssen, and E. E. Milios, "Characterizing and mining the citation graph of the computer science literature", *Knowledge and Information Systems*, vol. 6, no. 6, pp. 664–678, 2004.

[8] D. G. Feitelson and U. Yovel, "Predictive ranking of computer scientists using CiteSeer data", *Journal of Documentation*, vol. 60, no. 1, pp. 44-61, 2004.

[9] A. A. Goodrum, K. W. McCain, S. Lawrence, and C. L. Giles, "Scholarly publishing in the Internet age: A citation analysis of computer science literature", *Information Processing and Management*, vol. 37, no. 5, pp. 661–675, 2001.

[10] D. Zhao and E. Logan, "Citation analysis using scientific publications on the Web as data source: A case study in the XML research area", *Scientometrics*, vol. 54, no. 3, pp. 449–472, 2002.

[11] C.L. Giles and I.G. Councill, "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 51, pp. 17599-17604, 2004.

[12] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, vol. 46, no. 5, pp. 604-632, 1999.

[13] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117, 1998.

[14] D. Fiala, F. Rousselot, and K. Ježek, "PageRank for bibliographic networks", *Scientometrics*. vol. 76, no. 1, pp. 135-158, 2008.

[15] D. Fiala, "Time-aware PageRank for bibliographic networks", *Journal of Informetrics*, vol. 6, no. 3, pp. 370-388, 2012.