Mining citation information from CiteSeer data

Dalibor Fiala

University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic

Phone: 00420 377 63 24 29, fax: 00420 377 63 24 01, email: dalfia@kiv.zcu.cz

Abstract: The CiteSeer digital library is a useful source of bibliographic information. It allows for retrieving citations, co-authorships, addresses, and affiliations of authors and publications. In spite of this, it has been relatively rarely used for automated citation analyses. This article describes our findings after extensively mining from the CiteSeer data. We explored citations between authors and determined rankings of influential scientists using various evaluation methods including citation and in-degree counts, HITS, PageRank, and its variations based on both the citation and collaboration graphs. We compare the resulting rankings with lists of computer science award winners and find out that award recipients are almost always ranked high. We conclude that CiteSeer is a valuable, yet not fully appreciated, repository of citation data and is appropriate for testing novel bibliometric methods.

***Keywords**: CiteSeer, citation analysis, rankings, evaluation.*

Introduction

Data from CiteSeer have been surprisingly little explored in the scientometric literature. One of the reasons for this may have been fears that the data gathered in an automated way from the Web are inaccurate – incomplete, erroneous, ambiguous, redundant, or simply wrong. Also, the uncontrolled and decentralized nature of the Web is said to simplify manipulating and biasing Web-based publication and citation metrics. However, there have been a few attempts at processing the CiteSeer data which we will briefly mention.

Zhou et al. (2007) have investigated documents from CiteSeer to discover temporal social network communities in the domains of databases and machine learning. On the other hand, Hopcroft et al. (2004) track evolving communities in the whole CiteSeer paper citation graph. An et al. (2004) have constructed article citation graphs in several research domains by querying CiteSeer and have explored them in terms of components. Popescul et al. (2003) have classified CiteSeer articles into categories by venues. Šingliar and Hausknecht (2006) cluster CiteSeer papers by topics based on their references to authors. Author co-citation analysis of CiteSeer documents in the XML research field has been conducted by

Zhao and Strotmann (2007) and Zhao and Logan (2002) and in computer graphics by Chen (2000). Bar-Ilan (2006) has used CiteSeer for a citation analysis of the works of a famous mathematician. A kind of citation analysis, but this time for acknowledgements, has also been performed by Giles and Councill (2004). Chakrabarti and Agarwal (2006) use CiteSeer data in their experiments with learning ranking functions for real-world entity-relation graphs. Feitelson and Yovel (2004) have examined citation ranking lists obtained from CiteSeer and predicted future rankings of authors.

Most of the research activities mentioned above have been concerned with just a small part of the CiteSeer database, limited to a specific scientific field or even venue (conference or journal). Very few have dealt with the CiteSeer citation graph as a whole as we do in this study whose research questions are the following: What is the nature of CiteSeer data? Can sufficiently large citation and co-authorship graphs for publications and authors be constructed out of them? If yes, can we, based on those graphs, generate realistic rankings of salient researchers? In the rest of this paper, we will first describe the methods we work with, present the basic features of CiteSeer and its data and then show that we can answer yes to the last two questions.

Methods

In our previous work (Fiala et al. 2008 and Ježek et al. 2008), we have built on top of the well-known PageRank concept by Brin and Page (1998) and have modified this ranking function originally devised for the Web graph so as to evaluate author significance based on the citation as well as collaboration networks. The key concept is that a citation from a colleague is less valuable than that from a foreign researcher. Thus, cited authors should be penalized for the frequency of collaboration (co-authorship) with authors citing them. To add more information to the citation graph, we defined several parameters to weight its edges more discriminatively than purely by citation counts. These parameters, calculated from the collaboration graph, are the following:

a) $c_{u,v}$ is the number of common publications by authors u and v (i.e. the number of their collaborations, code-named COLLABORATION),

b) $f_{u,v}$ 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, code-named ALL_PUBLICATIONS),
c) $h_{u,v}$ 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*, code-named ALL_COAUTHORS,
d) $hd_{u,v}$ 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*, code-named ALL_DIST_COAUTHORS,
e) $g_{u,v}$ 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, code-named ALL_COLLABORATIONS),
f) $t_{u,v}$ is the number of co-authors (including duplicates) in common publications by authors *u* and *v*, code-named COAUTHORS,
g) $td_{u,v}$ is the number of distinct co-authors in common publications by authors *u* and *v*, code-named DIST_COAUTHORS.

Note that we make no distinction between authoring and co-authoring a publication. In either case, an author has published the publication. Also, for the sake of simplicity of parameters *h*, *hd*, *t*, and *td*, authors are considered as co-authors of themselves. For a much more detailed theoretical background as well as a practical example, we refer the reader to the article by Fiala et al. (2008).

Data

CiteSeer[1] gathers information mainly about computer science publications by crawling the World Wide Web, downloading, and automatically analyzing potential scientific publications (mostly PDF or PS files) and provides access to it via a Web interface and downloadable XML-like files that can be further processed by machines. The information in these XML files typically includes publication title, authors, their affiliations and addresses, abstract, and references. For our experiments, we chose the CiteSeer data files from December 13, 2005. These are the

[1] http://citeseer.ist.psu.edu

most recent data files prior to transforming CiteSeer into CiteSeer^X, which is dubbed "the next generation CiteSeer" and which is still in a beta version.

Possible data sources

CiteSeer is just one of many of bibliographic databases the most widely used of which are presented in Table 1. We may divide the databases into two groups according to their free availability or the way they are created and maintained. ACM Portal[2] consisting of the ACM Digital Library and of the ACM Guide along with Scopus[3] and Web of Science[4] are commercial subscription-based services (although some limited free access is provided by ACM) whereas CiteSeer, DBLP[5], and Google Scholar[6] are free for everyone with an Internet connectivity. On the other hand, CiteSeer and Google Scholar are automated systems while the databases of ACM Portal, DBLP, Scopus, and Web of Science are created and maintained mostly manually needing much human labour.

Table 1 Feature matrix of the main bibliometric systems as of October 4, 2010

	ACM Portal	CiteSeer^x	DBLP	Google Scholar	Scopus	Web of Science
Free	partly	yes	yes	yes	no	no
Automated	no	yes	no	yes	no	no
# records	1.59 mil.	32.23 mil.	1.46 mil.	NA	42.74 mil.	45.68 mil.
All bibl. data downloadable	no	yes	yes	no	no	no
Reference linking	yes	yes	partly	no	yes	yes
Citation linking	yes	yes	partly	yes	yes	yes
# citations for a publication	yes	yes	partly	yes	yes	yes
# citations for an author	yes	indirectly	partly indirectly	indirectly	yes	yes
domain coverage	computer science	computer science	computer science	general	general	general

As for the scope of the individual databases, the number of records in Table 1 means actually the number of all bibliographic records in the database, i.e. the number of research papers indexed plus the number of articles cited by the papers

[2] http://portal.acm.org

[3] http://www.scopus.com

[4] http://apps.isiknowledge.com

[5] http://dblp.uni-trier.de

indexed that are not in the database. For instance, the ACM Digital Library contains 290 thousand documents; 1.59 million records are available in the ACM Guide. CiteSeer^{X} actually owns 1.67 million documents only. DBLP is somewhat different – it is not a document repository, it merely stores bibliographic records so there is no need to make a distinction between documents and records. Google Scholar does not reveal any details about its database so, with certainty, we can only say that, in October 2010, it provides about 8.94 million results as a response to the query "the". (We are looking for documents containing the most frequent English word.) . Some of the results are documents but some of them are cited references only. Thus, Google Scholar currently provides access to no less than 9 million bibliographic records. Until now, solely estimates of the relative size of Google Scholar have been made by comparing its overlap with other bibliographic databases. Most of the papers on this topic are listed by Franceschet (2010).

While the absolute size of Google Scholar is unknown, a little bit more can be said about the documents it indexes – Meho and Yang (2007) report over 30 different document types in a sample of Google Scholar records such as journal articles, conference papers, dissertations, theses, technical reports, etc. (A similar earlier study by Goodrum et al. (2001) identified the following main document types in CiteSeer – journal articles, conference proceedings, technical reports, and books.) Indeed, regarding the same approach to obtaining documents by crawling the World Wide Web and looking for anything that looks like a research paper (a computer science research paper in the case of CiteSeer), one might expect that the document types covered by both Google Scholar and CiteSeer are almost the same.

Finally, the two huge human-made repositories of scientific literature, Scopus and Web of Science, make both available over 40 million records. Those 42.74 million records in Scopus can be really retrieved, for instance by searching for articles with an arbitrary title ("%"). If we restrict the search to articles published since 1996, we get the actual number of full-text documents in the database – 22.37 million. This number (of full-text documents) cannot be found out from the Web of Science.

Of the six databases, only CiteSeer and DBLP provide a full access to their bibliographic data in the form of one or more XML-like files. Unlike DBLP (see

[6] http://scholar.google.com

Fiala et. al. 2008), CiteSeer data records are substantially more linked by citations. The free availability of downloadable XML data and the high density of the citation graph are the key features that make CiteSeer the best tool for automated bibliometric and citation analyses despite its errors.

The other features in Table 1 describe more or less the user interface friendliness of the databases. In some of them, the user can go directly to the cited articles by clicking on the references in a paper (*reference linking*) or to the citing articles of the current paper (*citation linking*). We can get citation counts for an author directly or indirectly by counting citations to its publications (*citations for a publication* and *citations for an author*). These features are very limited in DBLP as it contains very few links between publications. The last aspect is the domain coverage of the databases – ACM Portal, CiteSeer, and DBLP cover mainly computer science whereas Google Scholar, Scopus, and Web of Science are general services. Let us recall that this paper deals with CiteSeer (and not CiteSeer^{X}) and that the relevant information in Table 1 is true for both of them except for the number of records.

Citation graphs

CiteSeer data are much larger than DBLP data analyzed by Fiala et al. (2008). There are more than 1.8 million citations between 717 thousand publications. We took the publication citation graph as it was and constructed an author citation graph out of it. The only data pre-processing we performed was transforming author names into upper case, removing duplicate authors, parallel edges, and self-citations. The resulting directed graph G of citations between authors has then some 411 thousand vertices (authors) and 4.8 million weighted edges (citations).

We made no attempt at disambiguating authors and publications, which is a complicated and time-consuming task. Thus, one author name may represent many real people and a single researcher may be referred to with several names, e.g. “Jack Dongarra” and “Jack J. Dongarra” at positions 9 and 13 of the first ranking in Table 6 (Online Resource 1). Also, automatic name recognition in CiteSeer produces errors and may identify absurd words as author names, e.g. “Senior Member” or “Student Member” at positions 2 and 4 of the first ranking in the same table. As for publications, there may also be duplicates and other inaccuracies. It is unclear whether CiteSeer groups all similarly looking publications found

on the Web into one and if so, with what precision this happens. Nevertheless, if this was not the case, one might easily bias CiteSeer citation counts by placing many copies of particular articles all over the Web. We can expect as well that small typos in paper titles may wrongly result in new or missing publications, etc.

All in all, computer-generated Web-based bibliographic data like in CiteSeer are always less reliable than those created by humans like in DBLP. This is one of the reasons why they have been so little used in bibliometric studies so far. On the other hand, they are much larger and much more up-to-date and we believe that the democratic, decentralized, and self-controlled nature of the Web itself makes it very difficult to manipulate Web-based bibliographic citations significantly and systematically. Zhao (2005) indicates that citation analyses based on CiteSeer may be as valid as those based on conventional data sources. Therefore, analyzing CiteSeer data makes sense and can bring new bibliometric insights into recent computer science publications.

Results

In the following tables and figures, we present the results of applying twelve different ranking methods to the amended citation graph of authors described earlier. The first five rankings are by pure citation counts (*Cites*), in-degree of author citation graph nodes (*InDeg*), HITS authorities (*HITS* - see Kleinberg 1999), PageRank (*PR*), and weighted PageRank (*w*). Next, we computed the previously defined parameters *c*, *f*, *g*, *h*, *hd*, *t*, and *td* from the collaboration graph, incorporated them into the PageRank formula (for details, see Fiala et al. 2008) and obtained rankings *a*) – *g*) corresponding to the numbering in section Methods.

Rankings

In addition to computing the ranks of all authors in the citation graph by each ranking method, we also compared each ranking with the list of ACM SIGMOD E. F. Codd Innovations Award winners (http://www.sigmod.org/awards) like Sidiropoulos and Manolopoulos (2005) to see how well they correlate with human-made charts of influential computer scientists. In Tables 2 and 3, we can see the ranks by all methods of 18 researchers awarded from 1992 to 2009. One of the researchers, Patricia Selinger, does not appear in any ranking. She is not present in the CiteSeer data we analyzed. For all rankings, we calculated three simple me-

trics characterizing the aggregate rank achieved by the awardees – worst rank, average rank, and median rank. The assumption is that the smaller are these values, the better is the ranking. In fact, an optimal ranking (including Patricia Selinger) equivalent to the human-made list in terms of these metrics, would have a worst rank of 18, an average rank of 9.5, and a median rank of 9.5.

Table 2 ACM Innovations Award winners and their ranks (part 1)

Year	Author	Cites	InDeg	HITS	PR	w
1992	Michael Stonebraker	137	87	170	35	36
1993	Jim Gray	194	132	132	287	367
1994	Philip Bernstein	1 477	1 767	2 055	4 884	4 749
1995	David DeWitt	27	38	84	75	43
1996	C. Mohan	2 634	2 419	3 996	4 945	4 958
1997	David Maier	458	284	604	375	521
1998	Serge Abiteboul	22	54	322	123	69
1999	Hector Garcia-Molina	14	14	58	89	63
2000	Rakesh Agrawal	3	9	112	41	15
2001	Rudolf Bayer	29 834	26 272	19 969	43 206	48 897
2002	Patricia Selinger					
2003	Don Chamberlin	5 497	4 577	4 474	7 162	9 125
2004	Ronald Fagin	512	587	1 160	701	774
2005	Michael Carey	161	163	220	308	306
2006	Jeffrey D. Ullman	228	205	476	609	575
2007	Jennifer Widom	7	15	103	81	29
2008	Moshe Vardi	217	326	1 622	447	441
2009	Masaru Kitsuregawa	16 497	12 603	7 972	27 477	42 133
	Worst rank	**29 834**	**26 272**	**19 969**	**43 206**	**48 897**
	Average rank	**3 407**	**2 915**	**2 561**	**5 344**	**6 653**
	Median rank	**217**	**205**	**476**	**375**	**441**

Table 3 ACM Innovations Award winners and their ranks (part 2)

Year	Author	a	b	c	d	e	f	g
1992	Michael Stonebraker	33	81	103	85	75	40	40
1993	Jim Gray	335	917	1 238	698	879	479	396
1994	Philip Bernstein	4 871	2 858	2 280	2 907	2 914	4 462	4 642
1995	David DeWitt	55	46	49	30	45	22	42
1996	C. Mohan	4 877	5 357	5 502	5 269	5 340	5 327	5 095
1997	David Maier	537	169	128	117	161	446	473
1998	Serge Abiteboul	76	43	47	36	42	44	66
1999	Hector Garcia-Molina	78	23	22	45	18	34	76
2000	Rakesh Agrawal	17	15	19	15	14	17	17
2001	Rudolf Bayer	48 600	52 676	54 482	51 648	52 522	49 505	49 098
2002	Patricia Selinger							
2003	Don Chamberlin	8 880	13 497	18 963	12 341	13 129	9 879	9 236
2004	Ronald Fagin	838	419	457	476	416	658	795
2005	Michael Carey	310	620	689	430	580	314	312
2006	Jeffrey D. Ullman	560	427	349	547	388	415	588
2007	Jennifer Widom	43	24	23	28	21	20	30
2008	Moshe Vardi	507	100	114	144	106	349	443
2009	Masaru Kitsuregawa	42 179	44 500	44 869	44 072	44 531	42 659	42 558
	Worst rank	**48 600**	**52 676**	**54 482**	**51 648**	**52 522**	**49 505**	**49 098**
	Average rank	**6 635**	**7 163**	**7 608**	**6 993**	**7 128**	**6 745**	**6 700**
	Median rank	**507**	**419**	**349**	**430**	**388**	**415**	**443**

The baseline ranking PR appears in a coloured column. It has a median rank of 375 which is outperformed only by ranking *c*) – ALL_COAUTHORS and by the both first-order methods *Cites* and *InDeg*. Its average rank 5 344 is forth best after *HITS*, *InDeg*, and *Cites*. The same holds for its worst rank 43 206. *HITS* is, somewhat surprisingly, the best ranking method as for the worst and the average rank. However, this is particularly thanks to the relatively high ranks (small numbers) for Rudolf Bayer and Masaru Kitsuregawa in comparison to the other rankings. On the other hand, it is the second worst ranking in terms of the median rank. Only method *a*) – COLLABORATION is worse in this respect.

A graphical presentation of the results in Tables 2 and 3 is given in Figure 1. Rakesh Agrawal, Jennifer Widom, and Hector Garcia-Molina are always top-ranked. While Rakesh Agrawal obtains the highest median rank of 16 and Hector Garcia-Molina never falls off the Top 100, Jennifer Widom's result is remarkable in that she received the award only in 2007 and thus could not attract citations after her nomination (CiteSeer data are from 2005). The rank series are quite stable – there are no evident outliers except a slight deterioration by *HITS* for the better ranked authors.

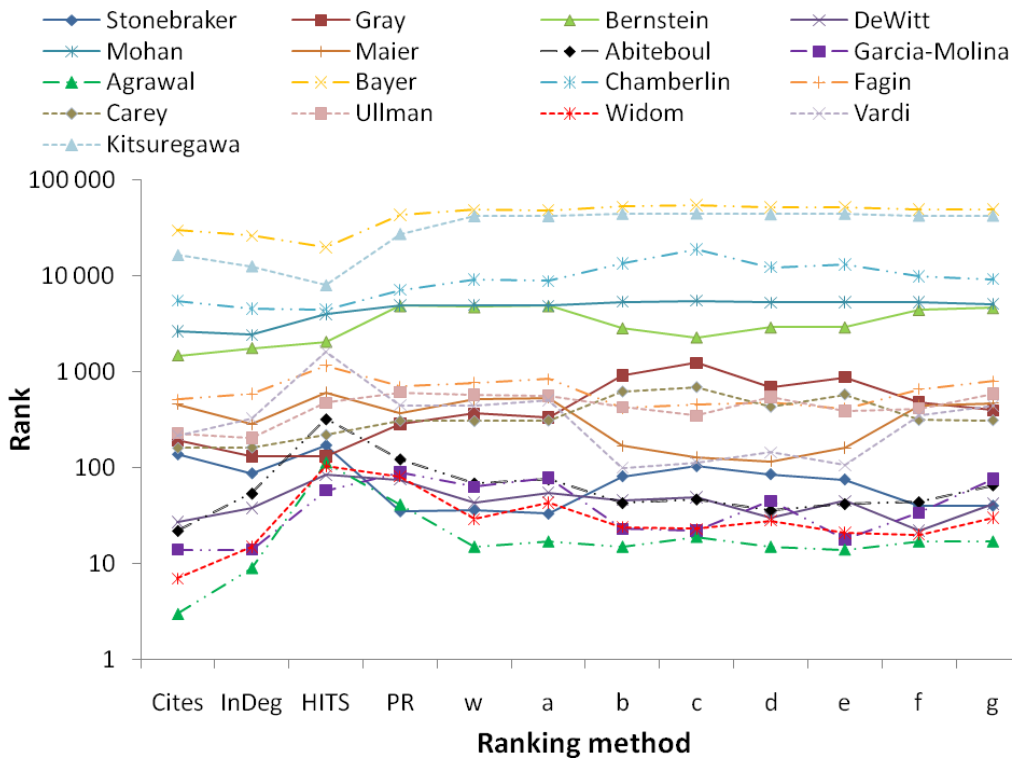

Fig. 1 ACM Innovations Award winners and their ranks

A complete overview of top 40 scientists in all rankings may be found in Tables 4 through 7 (Online Resource 1) with award recipients printed in bold. A simple look at the tables reveals that the number of award winners varies between 5 in *Cites* and *f* (COAUTHORS) or 4 in *InDeg* and *d* (ALL_DIST_CO-AUTHORS) and 1 in *PR* or even 0 in *HITS*. This suggests that as far as the top of each ranking is concerned, any improved PageRank (with some additional information from the collaboration graph) is closer to the real-world perception of a researcher's significance than the standard PageRank but is still at best as good as common (and far less computationally expensive) first-order methods based on simple citation counts.

The above tables may also be used for a prediction of future ACM SIGMOD E. F. Codd Innovations Award winners if we choose scientists active in the database field. Regarding the fact that citation and in-degree rankings have the largest overlap with the true list of awardees (see Table 2) and after consulting Scopus about the fields of interest of the top-ranked authors in Table 4 (Online Resource 1), Ramakrishnan Srikant and Christos Faloutsos seem to be the hot candidates. Scott Shenker, Sally Floyd, and Van Jacobson appear almost always among the top researchers in each ranking but as their interests do not focus on databases, they should be considered as candidates for other awards.

Conclusions

Current tools for analyzing social networks in the scientific community concentrate mainly on established citation indices such as ISI Web of Science or Scopus. These databases were originally not conceived to allow for a direct machine processing and, therefore, information scientists treat them manually or semi-manually. This approach results in very time-consuming analyses of relatively little data. On the other hand, the data from open access Web services such as CiteSeer are still rather underestimated as they are computer-generated and hence error-prone. However, their potential is great as their accuracy and completeness get higher and the general need for large and up-to-date bibliographic and citation databases grows.

In this paper, we present the results of our experiments with CiteSeer data. We show that sufficiently large citation and collaboration graphs for publications and authors can be created from these data. We analyze the citation graph of pub-

lication authors and present twelve rankings of the most influential researchers. In addition to common ranking methods such as counting citations or in-degree, we apply variations of the standard PageRank formula that combine information from both the citation and collaboration graphs. With respect to CiteSeer's drawbacks such as missing or wrong data, we argue that author rankings based on CiteSeer are realistic enough (by comparing them with true award recipients) so that they might be carefully used along with other data sources for the prediction of future computer science award winners. We conclude that CiteSeer, due to its free availability and well-structured large-scale data, is very well suited for citation analyses and testing of bibliometric methods despite its inherent errors. This work is the most comprehensive analysis of author citations based on CiteSeer data that we are aware of.

The remaining research issues are particularly the reliability of CiteSeer data, a more in-depth analysis of the CiteSeer collaboration graph, and differences between CiteSeer and CiteSeer^{X}. These and other topics including retrieving addresses, affiliations, and countries from CiteSeer shall be discussed in future studies.

Acknowledgements

This work[7] was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009. Many thanks go to the anonymous reviewers for their useful hints and comments and to Karel Ježek for his support of this project.

References

AN, Y., JANSSEN, J., MILIOS, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, 6(6): 664—678.

BAR-ILAN, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management*, 42(6): 1553—1566.

BRIN, S., PAGE, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the 7th World Wide Web Conference, Brisbane, Australia, pp. 107-117.

CHAKRABARTI, S., AGARWAL, A. (2006). Learning parameters in entity relationship graphs from ranking preferences. *Lecture Notes in Computer Science*, 4213: 91-102.

CHEN, C. (2000). Domain visualization for digital libraries. Proceedings of the International Conference on Information Visualization (IV2000), London, UK, pp. 261-267.

[7] The related software may found at http://textmining.zcu.cz/downloads/sciento.php.

FEITELSON, D. G., YOVEL, U. (2004). Predictive ranking of computer scientists using CiteSeer data. *Journal of documentation*, 60(1): 44-61.

FRANCESCHET, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, 83(1): 243-258.

FIALA, D., ROUSSELOT, F., JEŽEK, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1): 135-158.

GILES, C. L., COUNCILL, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51): 17599-17604.

GOODRUM, A. A., MCCAIN, K. W., LAWRENCE, S., GILES, C. L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature. *Information Processing and Management*, 37(5): 661-675.

HOPCROFT, J., KHAN, O., KULIS, B., SELMAN, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl. 1): 5249-5253.

JEŽEK, K., FIALA, D., STEINBERGER, J. (2008). Exploration and Evaluation of Citation Networks. Proceedings of the 12th International Conference on Electronic Publishing, Toronto, Canada, pp.351-362.

KLEINBERG, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5): 604-632.

MEHO, L. I., YANG, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13): 2105–2125.

POPESCUL, A., UNGAR, L. H., LAWRENCE, S., PENNOCK, D. M. (2003). Statistical relational learning for document mining. Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida, USA, pp. 275-282.

SIDIROPOULOS, A., MANOLOPOULOS, Y. (2005). A citation-based system to assist prize awarding. *SIGMOD Record*, 34 (4): 54–60.

ŠINGLIAR, T., HAUSKRECHT, M. (2006). Noisy-OR Component Analysis and its Application to Link Analysis. *Journal of Machine Learning Research*, 7: 2189–2213.

ZHAO, D. (2005). Challenges of scholarly publications on the Web to the evaluation of science - A comparison of author visibility on the Web and in print journals. *Information Processing & Management*, 41(6): 1403-1418.

ZHAO, D., LOGAN, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics*, 54(3): 449-472.

ZHAO, D., STROTMANN, A. (2007). Can citation analysis of web publications better detect research fronts? *Journal of the American Society for Information Science and Technology*, 58 (9): 1285-1302.

ZHOU, D., COUNCILL, I., ZHA, H., GILES, C. L. (2007). Discovering temporal communities from social network documents. Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM'07), Omaha, Nebraska, USA, pp. 745-750.