Journal of Theoretical and Applied Information Technology
XX[st] Month 201x. Vol. x No.x

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

FROM CITESEER TO CITESEER^X: AUTHOR RANKINGS BASED ON COAUTHORSHIP NETWORKS

DALIBOR FIALA
University of West Bohemia, Department of Computer Science and Engineering, Czech Republic
E-mail: dalfia@kiv.zcu.cz

ABSTRACT

CiteSeer was a digital library and a search engine gathering its mainly computer science research papers from the World Wide Web. After a few years of stagnation, it was definitely replaced with a new version called CiteSeer^X in April 2010. As both CiteSeers provide(d) freely available metadata on the articles they index(ed), it is possible to analyze two different data sets to see the differences between CiteSeer and CiteSeer^X. More specifically, we examined the article metadata from CiteSeer (downloaded in December 2005) and from CiteSeer^X (harvested in March 2011) with a view of creating rankings of prestigious computer scientists. Since the free article metadata acquired from the Web site of CiteSeer^X differ from those in CiteSeer in that they do not systematically include cited references, the only possibility of creating such rankings is to base them on the coauthorship networks in both CiteSeers. In this study, we produce these rankings using 12 different ranking methods including PageRank and its variants, compare them with the lists of ACM A. M. Turing Award and ACM SIGMOD E. F. Codd Innovations Award winners and conclude that the rankings generated from CiteSeer^X data outperform those from CiteSeer.

Keywords: *CiteSeer, CiteSeer^X, Coauthorships, Citations, Researchers, PageRank*

1. INTRODUCTION AND RELATED WORK

CiteSeer [1] was a digital library and a search engine specialized mainly in computer science literature that gathered its content by autonomously crawling the World Wide Web and downloading and parsing potentially relevant documents [2]. After some time of running in parallel with a new version, finally, in April 2010, the "old" CiteSeer officially ceased to exist and was replaced by the new CiteSeer^X [3], which is, however, still in a beta version at the time of writing this paper (May 2013). In fact, the old URL redirects to the new one now. Anyway, in the last years of its existence, CiteSeer was no more updated. On the other hand, CiteSeer^X has been continuously updated since its creation until now. Although there have been enough studies based on CiteSeer data, some of which will be mentioned in the related work section, research dealing with CiteSeer^X has been somewhat rare so far, probably partly due to the relative novelty and presumed immaturity of CiteSeer^X. Also, even though the nature of CiteSeer data invites bibliometric analyses, there have been few of them, perhaps as a result of the presence of errors in the data that have been created using automated text processing tools. In spite of this, some papers have reported a successful usage of CiteSeer data for bibliometric purposes (see more on this in the following paragraphs).

This study tries to analyze the freely available article metadata of CiteSeer and CiteSeer^X (obtainable from their respective Web sites) and to answer the following main research questions: a) What is the structure of these article metadata of CiteSeer and CiteSeer^X and what are the basic characteristics of the coauthorship networks generated from them? b) Can the coauthorship networks of CiteSeer and CiteSeer^X be used to rank computer scientists? c) And, if yes, which CiteSeer generates better rankings if they are compared to the lists of prestigious computer science award winners (ACM A. M. Turing Award and ACM SIGMOD E. F. Codd Innovations Award)?

Numerous studies have explored CiteSeer or CiteSeer^X data for non-bibliometric purposes, mainly to test various graph-theoretic approaches. An et al. [4] analyzed the citation graph of CiteSeer (then called ResearchIndex) in terms of connectivity. Chakrabarti and Agarwal [5] made use of CiteSeer citation data to test their unified ranking model on real-world graphs. Chakrabarti et al. [6] utilized the CiteSeer corpus and query logs to test new techniques of personalized PageRank

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

computation on entity-relation graphs. Hopcroft et al. [7] tracked evolving communities of computer science research papers by exploring the CiteSeer citation graph from 1998 and 2001. Joorabchi and Mahdi [8] used CiteSeer documents to evaluate the performance of their automatic classification of research papers according to a standard library classification scheme. Popescul et al. [9] employed CiteSeer data to train and test their new classifier that categorized research papers into publication venues. Šingliar and Hauskrecht [10] performed a component analysis of a partial CiteSeer citation graph. Zhou et al. [11] used thousands of CiteSeer documents in the construction of a real-world network to test their graph partitioning algorithm for the discovery of temporal communities of computer science researchers. Chen et al. [12] proposed a system based on the coauthorship network of CiteSeer^X to recommend potential collaborators. He et al. [13] designed a recommender system suggesting cited references for a given article based on the many citation contexts available in CiteSeer^X. Abstracts from CiteSeer^X documents were employed in the construction of hierarchical topic-based communities of authors by Wu and Koh [14].

Fewer studies have been bibliometric. CiteSeer was used as one of the data sources providing citation data for the citation analysis of the works of a famous mathematician by Bar-Ilan [15]. Feitelson and Yovel [16] took advantage of CiteSeer's citation counts of highly cited researchers in their predictive model of future citation-based ranks of researchers. Giles and Councill [17] investigated acknowledgements in the papers of the CiteSeer archive including its citation graph and determined the most acknowledged entities as well as their citation counts. Goodrum et al. [18] analyzed the most cited documents in the CiteSeer database and found out their publication type and age, among others. Zhao [19] explored the CiteSeer citation graph in the XML research field and identified highly productive and influential scientists. Zhao and Logan [20] carried out a similar study and concluded that citation analysis based on CiteSeer (at least in the XML domain) is as valid as that based on established data sources. And, finally, Zhao and Strotmann [21], again in the XML research field, conducted an author co-citation analysis of CiteSeer documents and compared the results with an analysis based on ISI Science Citation Index. Krumov et al. [22] constructed a coauthorship network from CiteSeer^X data and examined the relation of coauthorship patterns to the impact of scientific publications.

Unlike our research, most of the above studies have not dealt with the CiteSeer citation or coauthorship graph as a whole – they have been mostly concerned with a part of it only. Furthermore, none of them has analyzed CiteSeer as well as CiteSeer^X at the same time. In this context, this study is unique in that it examines the whole coauthorship graphs of both CiteSeers. It is an extension to our previous work, in which a citation analysis of the whole CiteSeer citation graph with a view of identifying prominent computer scientists was carried out [23] and a bibliometric analysis of all CiteSeer metadata aimed at finding the most productive and influential countries in computer science was conducted [24]. The usefulness of coauthorships in the assessment of researchers was shown by Yan and Ding [25] who determined the impact of authors in the informetrics research community by applying the PageRank algorithm to a coauthorship network. For the evaluation of the author rankings resulting from our analyses, we use the same technique (comparing the rankings with the lists of computer science award winners) as in other studies [23, 26-28].

2. DATA AND METHODS

In the present study, we examined two data sets – CiteSeer and CiteSeer^X. Because CiteSeer was no more updated in the last years of its existence, the most recent data file that we could obtain was from December 2005. On the other hand, CiteSeer^X has been continuously updated since its creation until now and we took a snapshot if its metadata in March 2011. Thus, there is a roughly six-year age difference in the two data files, the analysis of which we present in this study. We downloaded CiteSeer metadata straight from its Web site as an archive file and we harvested CiteSeer^X metadata from its Open Archives Initiative collection [29]. The freely available metadata for each article in CiteSeer generally include its title, abstract, authors, authors' addresses and affiliations, source URL, document format and language, cited references, and publication year and download date. However, addresses and affiliations, references, and publication years are often missing, incomplete, or erroneous. On the other hand, the article metadata harvested from CiteSeer^X include information on the document publisher, but addresses and affiliations are entirely absent and references (or citations) do not appear systematically.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In total, there were 716768 "core" (i.e., with article full texts) publication records in CiteSeer and 1334000 "core" publication records in CiteSeer^X. Thus, the number of records almost doubled between 2005 and 2011. As complete citations between publications are not available in the CiteSeer^X metadata we had (unlike CiteSeer), the only possibility of constructing comparable author citation graphs from both CiteSeers is to base them on the coauthorship networks (similarly to Yan and Ding, 2011) that can be easily built from both metadata sets. From a coauthorship (or collaboration) network with publications and their respective authors, we can obtain a graph of authors, in which every two coauthors of a publication are connected with an undirected edge. To avoid parallel edges in the case of many publications being written by the same coauthors, the edge will be assigned a weight denoting the number of joint publications. Next, each undirected edge is replaced with two oppositely directed edges both retaining the original weight. As a result, a citation graph of authors based on the collaboration network has been created. The basic statistics of such author citation graphs generated from the article metadata of CiteSeer and CiteSeer^X can be seen in figure 1.

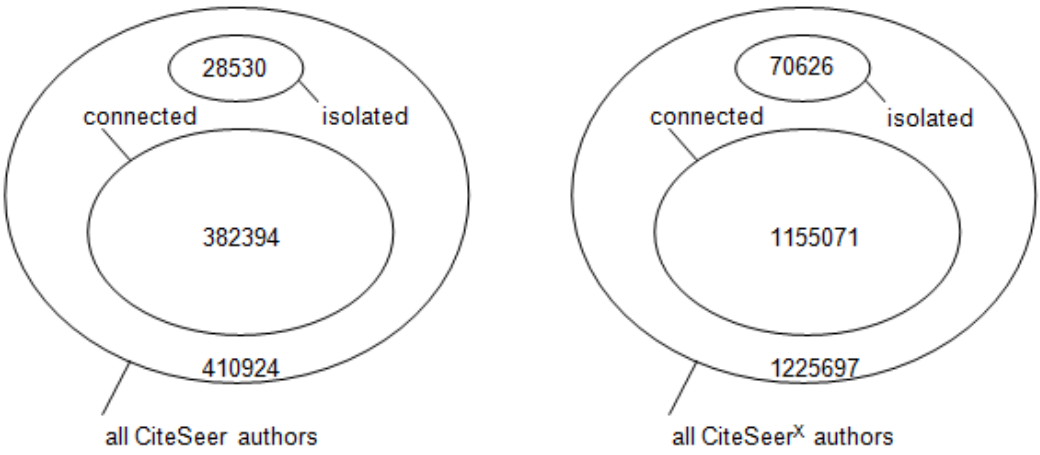


Figure 1: Basic Statistics of the Coauthorship Graphs in Both CiteSeers

Without disambiguation or duplicates removal, we found a total of 1 663 044 author records in CiteSeer and 3837226 in CiteSeer^X (not visible in figure 1). After transforming author names into upper case, we identified 410924 "distinct" authors in CiteSeer and 1225697 "distinct" authors in CiteSeer^X. These are the actual numbers of nodes in the author citation graphs. We must underline that name unification and disambiguation is a very tedious and time-consuming task and is not the concern of this research. We examine the data from CiteSeer "as is", without any pre- or postprocessing and this may have influence on the rather high per-author citation counts below. Prior to the elimination of parallel edges in the author citation graphs, there were 4764960 citations (formerly collaborations) between authors in CiteSeer (11.6 per author) and 16023138 in CiteSeer^X (13.1 per author) excluding self-citations of all authors. After eliminating the parallel edges, there were 2466446 and 9607486 edges left, which were assigned weights as described above. As for the authors, their number tripled between 2005 and 2011, but the percentage of isolated authors remained almost the same (7% and 6%, respectively) compared to the total number of authors. "Connected authors" are those who cite or are cited, which is equivalent here, because the citation graph is based on symmetric collaborations. Finally, we can conclude that the linkage density of the CiteSeer coauthorship graphs did not change between 2005 and 2011.

To analyze the citation graphs, we decided to apply the same 12 ranking methods used also by Fiala [23], which were described in detail in another paper [27]. In this section, we will briefly summarize the rationale of these methods. In the citation analysis, we can basically choose from simple (first-order, non-recursive) methods such as citation counts (in fact, a "weighted" in-degree) or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in-degree ("unweighted") or from more complicated (higher-order, recursive) methods such as HITS [30] or the notoriously known PageRank [31], which were originally conceived for the World Wide Web but later also applied to other network types such as author citation networks to identify influential actors. The "standard" PageRank (*PR*, by Brin and Page) can be modified so as to better reflect the features of bibliographic networks. For instance, the formerly unweighted edges can be assigned weights that denote the number of citations between two authors and thus give rise to a "weighted PageRank" (*PR-W*). The weighted PageRank formula can be further extended with some additional information such as the number of collaborations (*PR-C*), publications (PR-P), all coauthors (*PR-AC*), all distinct coauthors ((*PR-ADC*), all collaborations (*PR-AColl*), coauthors (*PR-CA*), or distinct coauthors (*PR-DCA*) that can all have influence on the weight of the directed edge between two authors. Thus, we get 12 ranking methods in total (*Cites*, *InDeg*, *HITS*, *PR*, *PR-W*, *PR-C*, *PR-P*, *PR-AC*, *PR-ADC*, *PR-AColl*, *PR-CA*, and *PR-DCA*), all of which will be used in our analysis. (For all the PageRank-like methods, we used a damping factor *d* of 0.9, a Spearman correlation-based convergence criterion and a maximum of 50 iterations.)

3. RESULTS AND DISCUSSION

We were interested in the changes that occurred in the CiteSeer data from 2005 to 2011. First, we had a look at the distribution of publications based on the number of their authors. Figure 2 shows such a histogram. There we can observe some similarities and discrepancies between the two CiteSeers. For instance, both digital libraries have a significant amount of publications with no authors and this amount remains relatively the same. The cause of this may be the inability of the underlying algorithms to correctly identify author names. From this point of view, the parsing quality does not seem to improve over the years. The most frequent number of authors per paper is two in both cases, but there is a difference in the second most frequent number – this is one author in CiteSeer but three authors in CiteSeer^X. There may be several reasons for this phenomenon including the general increase in the average number of authors per paper in computer science between 2005 and 2011 or the concentration of CiteSeer^X on a specific subfield of computer science with a higher number of authors. However, finding a precise explanation was not the aim of this study.

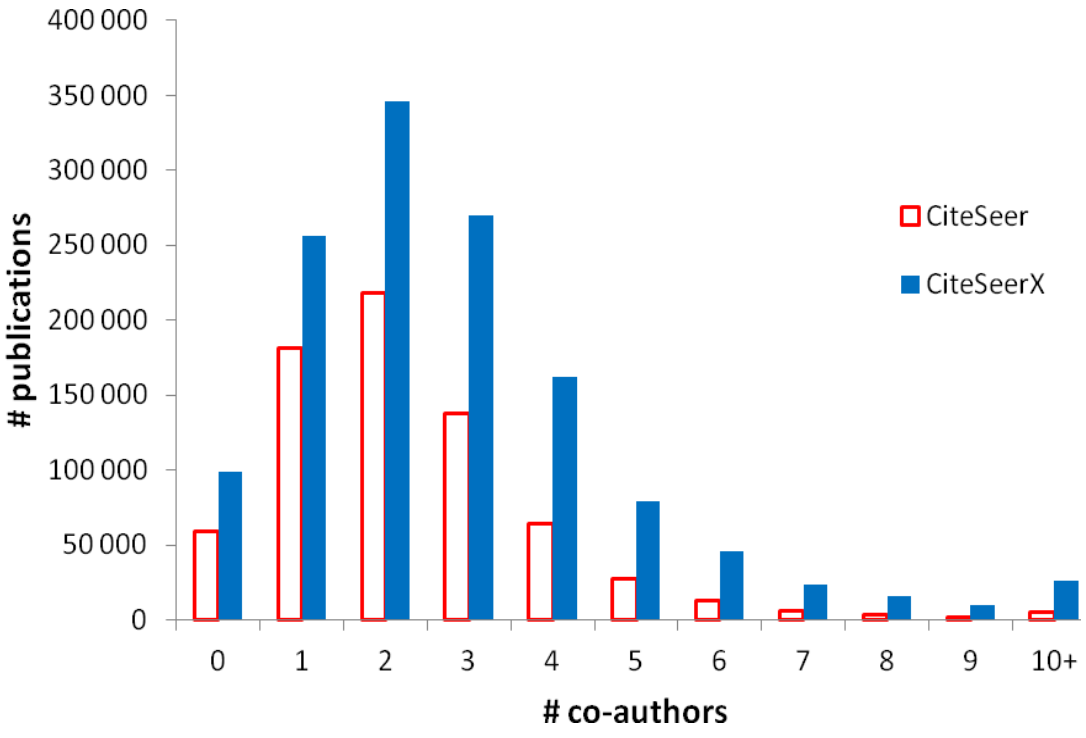
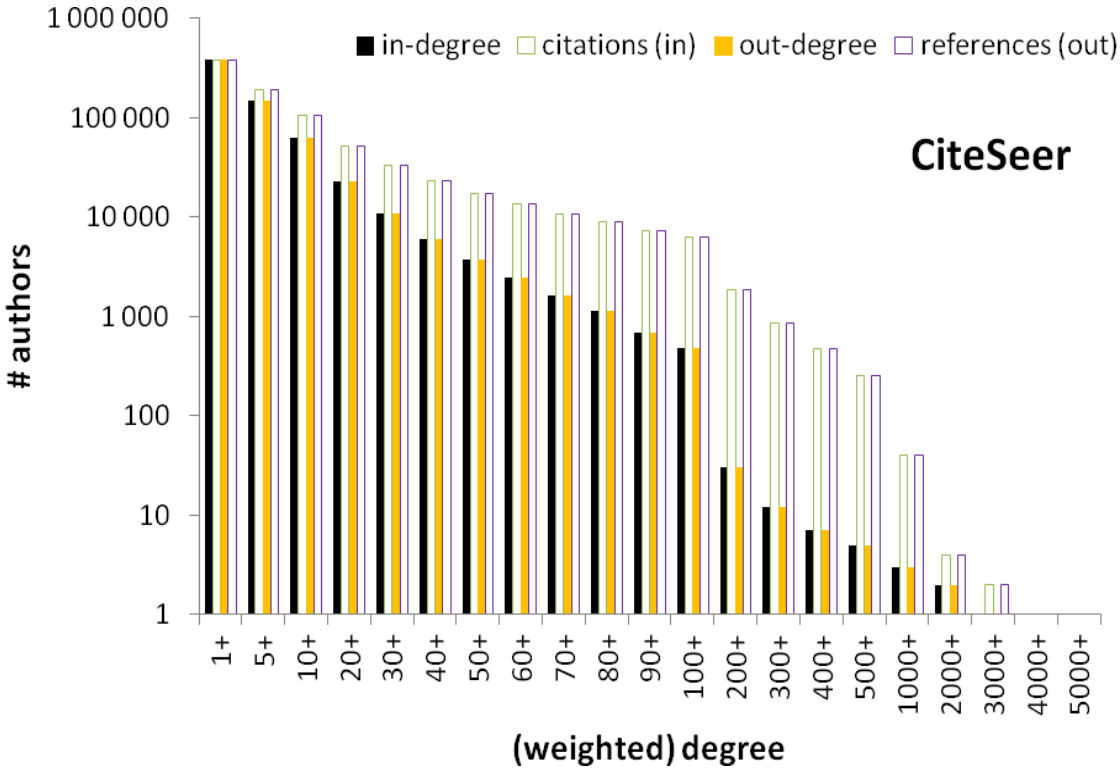


Figure 2: Coauthor Distribution of Publications in Both CiteSeers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



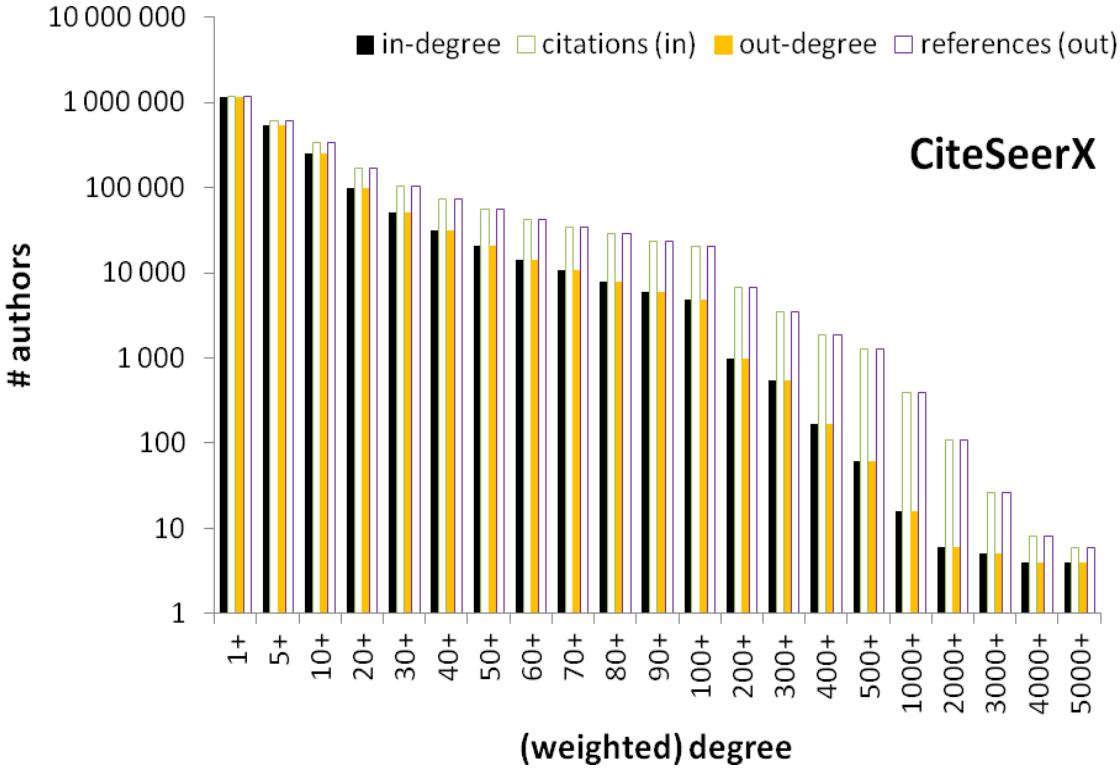


Figure 3: Distribution of Authors by (Weighted) In- and Out-degree in Both CiteSeers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

As far as the "density" of the graph of citations between publications is concerned, a great deal is revealed from the cumulative histogram charts in figure 3. The bars represent authors (i.e., graph nodes) with a specific magnitude of (weighted) in-degree or (weighted) out-degree. All the indicators are always larger in CiteSeer^X due to the overall greater number of nodes and edges in the graph. We call the weighted in-degree "citations" and the weighted out-degree "references". Evidently, for a weighted degree, the weights of in-coming (or out-going) edges are summed up. Since the directed graphs under study are based on symmetric collaborations, in-degrees and weighted in-degrees are equal and so are out-degrees and weighted out-degrees. The charts use a logarithmic Y-axis scale to better display bars in their tales. Thus, for instance, some 0.13% of authors have an in-degree of 100 or more in CiteSeer, whereas it is 0.41% in CiteSeer^X. Also, CiteSeer^X includes some authors that have more than 5000 citations, but CiteSeer does not. What authors are the most cited in both CiteSeers is shown in table 1.

Table 1 presents the top 40 authors by citations and in-degree in CiteSeer and CiteSeer^X. (Names in italics cannot be printed in full due to space limitations.) As we can see, there is a lot of noise in the results due to errors in the metadata. As a consequence, the most cited "researchers" turn out to be "Senior Member", "Student Member", or "Ph. D" in both CiteSeers, which are the words frequently occurring close to proper names on papers' title pages that were incorrectly parsed and classified as such. Nevertheless, some well known computer science researchers' names (such as "Jack Dongarra" or "Ian Foster") appear in the top 40 results from CiteSeer. In CiteSeer^X, less known scientists are in the top results, e.g. "R. R. Barton". An interesting extension to table 1 is table 2, in which the top 40 authors determined by three other methods (HITS, PageRank, and weighted PageRank) are presented. The HITS ranking differs the most from the others – it contains no noise and its researchers are mostly unknown. On the other hand, the PageRank and weighted PageRank rankings are noisy and include well known as well as little known computer science authors such as "Jack Dongarra", "Ian Foster", "Takeo Kanade", "R. R. Barton", or "Vladik Kreinovich".

As it is impossible to show all the 12 rankings in full, we focused our attention to two sets of researchers whose ranks generated by all the methods are visualized in the charts in figure 4 and in figure 5. In the first set, there are ACM A. M. Turing Award ("Nobel Prize" in computer science) winners from the years 1991 - 2010. In the second, there are ACM SIGMOD E. F. Codd Innovations Award winners ("Nobel Prize" in databases) from 1992 to 2011. The time spans for both prizes were selected as the last 20 available years at the time of our experiments. All the charts are displayed on the logarithmic scale and lower ranks mean better ranks (e.g. a rank of 10 is better than a rank of 100). By looking at the charts, we can immediately see a striking feature in all of them – the award winners generally receive bad ranks by HITS. This is supported by the fact we observed in table 2 – no well known researchers were placed at the top by HITS. Another clearly visible property of all the charts is the very good performance of simple citation counts (*Cites*). In principle, the award winners achieve good ranks by citation counts and, therefore, citations can be considered a "good" ranking in contrast to the much more computationally expensive HITS. And finally, PageRank (*PR*), itself also a computationally expensive method, performs comparably to citations but better than HITS and some of its variants are of the same quality or even slightly better than the standard PageRank (most notably *PR-W* for Codd Award winners in CiteSeer^X, see the lower chart in figure 5). All the three findings are in accordance with those reported by Fiala [23] on the normal author citation graph of CiteSeer. As for the individual scientists, the best ranked Turing Award winners (according to their median rank) are "Pnueli" and "Rivest" in CiteSeer and "Gray" and "Rivest" in CiteSeer^X and the best ranked Codd Award winners (according to their median rank) are "Garcia-Molina" and "Stonebraker" in CiteSeer and "Garcia-Molina" and "Widom" in CiteSeer^X. (Awardees whose names were absent in the data are missing in the charts. These are "Selinger" for the Codd Award in CiteSeer, "Feigenbaum", "Yao", "Nygaard", "Naur", and "Allen" for the Turing Award in CiteSeer and "Allen" for the Turing Award in CiteSeer^X.)

Journal of Theoretical and Applied Information Technology
XX[st] Month 201x. Vol. x No.x

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Top 40 Authors by Citations and In-degree in CiteSeer (CS) and CiteSeer^X (CS^X)

CS	Citations	CS^X		CS	In-degree	CS^X	
Senior Member	4390	Ph. D	28641	Senior Member	2570	Ph. D	10811
Student Member	3676	Senior Member	23136	Student Member	2185	Senior Member	10305
Fachbereich Informatik	2515	Prof Dr	21032	Ph. D	1795	Student Member	7771
Ph. D	2513	Student Member	17173	Fachbereich Informatik	823	Prof Dr	7114
Michael H. Bohlen	1898	Email Alerting	6105	Prof Dr	780	Email Alerting	3843
Kristian Torp	1895	J Neurophysiol	5128	Mathematisch Centrum	481	Jr.	2876
Christian S. Jensen (Codirector	1883	The Erwin	4960	Copyright Stichting	480	Et Al	1797
Richard T. Snodgrass (Codirector	1883	Jr.	4845	G. W. Evans	393	United States	1686
Heidi Gregersen	1880	H. Wahl	3467	H. B. Nembhard	393	J Neurophysiol	1639
Alex Waibel	1877	R. R. Barton	3397	P. A. Farrington	393	The Erwin	1488
Jack Dongarra	1795	V. Kekelidze	3258	D. T. Sturrock	392	Key Words	1149
Christian S. Jensen	1446	M. Martini	3255	Associate Member	311	*Technische Universität*	1146
Sudha Ram	1410	A. Gonidec	3204	Computer Science	287	*Schrödinger International*	1112
Deborah Estrin	1380	A. Ceccucci	3190	Forest Service	282	Forest Service	1110
Curtis E. Dyreson	1360	L. Gatignon	3180	Key Indicators	282	Computer Science	1054
Dieter Pfoser	1344	*Schrödinger International*	3179	E. Dvorkin (Eds	273	R. R. Barton	1009
Giedrius Slivinskas	1288	A. Gianoli	3079	Ian Foster	267	IEEE Computer Society	959
Renato Busatto	1272	A. Norton	3079	S. Idelsohn	265	Fachbereich Informatik	941
Janne Skyt	1244	W. Bartel	3076	Thme Rseaux Et Systmes	256	Prof Dr. -ing	818
Douglas C. Schmidt	1235	V. Falaleev	3054	Rwth Aachen	253	M. Sc	752
Mathematisch Centrum	1228	W. Kubischta	3051	Ecole Normale	248	Supervisor Prof	731
Copyright Stichting	1227	D. Cundy	3050	Jack Dongarra	248	Editorial Board	728
Hector Garcia-Molina	1166	A. Belousov	3039	Sophia Antipolis	244	Associate Member	698
Sebastian Thrun	1159	G. Bocquet	3039	Arthur C. Smith	239	Ipan Mohanty	673
Michael Stonebraker	1154	P. Hristov	3032	Member IEEE	220	Wildlife Service	664
Bongki Moon	1153	N. Molokanova	3018	P. L. Frabetti	216	Lt Col	663
H. Niemann	1104	F. Petrucci	2997	Alle Rechte Vorbehalten	214	Assoc Prof	662
J. Engler	1075	A. Zinchenko	2996	Vladik Kreinovich	211	Member IEEE	659
Prof Dr	1066	P. Dalpiaz	2996	Sun Microsystems	209	III	657
P. Doll	1052	E. Barrelet	2976	IEEE Computer Society	206	Ulrich H. E. Hansmann	638
D. Heck	1049	V. Boudry	2964	M. Martini	197	Gutachter Prof	631
Ian Foster	1033	P. L. Frabetti	2943	Christian S. Jensen	196	Olav Zimmermann (Editors	626
K. Daumiller	1028	V. Brisson	2940	*Technische Hochschule*	196	Sophia Antipolis	609
G. W. Evans	1024	Et Al	2927	Andrei Shleifer	194	B. Biller	608
H. B. Nembhard	1024	M. Savrié	2909	INRIA Rocquencourt	193	J. A. Joines	604
P. A. Farrington	1024	P. Baranov	2848	A. Ceccucci	192	J. D. Tew	603
D. T. Sturrock	1020	M. Velasco	2824	Mario Gerla	189	J. Shortle	603
K. Bekk	1020	K. Bekk	2820	Politecnico Di Milano	189	M. -h. Hsieh	603
H. Bozdog	1013	H. Bozdog	2790	D. Cundy	188	Principal Investigator	603
Don Towsley	1005	D. Bruncko	2763	Ron Kikinis	188	S. G. Henderson	603

Journal of Theoretical and Applied Information Technology
XX[st] Month 201x. Vol. x No.x

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Top 40 Authors by HITS, PageRank, and Weighted PR in CiteSeer (CS) and CiteSeer^X (CS^X)

CS	HITS	CS^X	CS	PageRank	CS^X	CS	PageRank (weighted)	CS^X
D. Cundy	H Collaboration		Senior Member	Ph. D		Senior Member	Ph. D	
H. Wahl	A. Belousov		Student Member	Senior Member		Student Member	Senior Member	
A. Ceccucci	V. Boudry		Ph. D	Student Member		Ph. D	Prof Dr	
V. Kekelidze	V. Brisson		Fachbereich Informatik	Prof Dr		Fachbereich Informatik	Student Member	
G. Bocquet	D. Bruncko		Prof Dr	Email Alerting		Prof Dr	Email Alerting	
A. Gianoli	A. Babaev		Mathematisch Centrum	Jr.		Mathematisch Centrum	Jr.	
P. L. Frabetti	G. Buschhorn		Copyright Stichting	The Erwin		Copyright Stichting	The Erwin	
L. Gatignon	W. Bartel		Key Indicators	United States		Jack Dongarra	J Neurophysiol	
N. Doble	E. Barrelet		G. W. Evans	Et Al		G. W. Evans	United States	
A. Gonidec	P. Baranov		H. B. Nembhard	Key Words		H. B. Nembhard	*Schrödinger International*	
B. Gorini	B. Delcourt		P. A. Farrington	*Schrödinger International*		P. A. Farrington	Et Al	
G. Barr	S. Egli		D. T. Sturrock	Computer Science		D. T. Sturrock	Forest Service	
J. Duclos	A. De Roeck		Forest Service	*Technische Universität*		Computer Science	*Technische Universität*	
A. Lacourt	G. Eckerlin		Associate Member	Forest Service		Alex Waibel	R. R. Barton	
D. Schinzel	V. Efremenko		Computer Science	J Neurophysiol		Turku Centre	Fachbereich Informatik	
M. Martini	E. Elsen		Arthur C. Smith	IEEE Computer Society		Vladik Kreinovich	Key Words	
A. Norton	Ch. Berger		Vladik Kreinovich	Fachbereich Informatik		Douglas C. Schmidt	Prof Dr. -ing	
B. Panzer-Steindel	F. Eisele		E. Dvorkin (Eds	R. R. Barton		Forest Service	Computer Science	
Yu. Potrebenikov	G. Cozzika		S. Idelsohn	Supervisor Prof		Key Indicators	Vladik Kreinovich	
A. Lai	J. Cvach		Member IEEE	M. Sc		Don Towsley	Assoc Prof	
W. Kubischta	M. Fleischer		Ecole Normale	Prof Dr. -ing		*Technische Hochschule*	M. Sc	
P. Grafstrom	A. Fedotov		Thme Rseaux Et Systmes	Associate Member		Deborah Estrin	Wildlife Service	
P. Hristov	L. Favart		Rwth Aachen	Member IEEE		E. Dvorkin (Eds	IEEE Computer Society	
A. Zinchenko	J. Ferencei		Key Words	*Schrodinger International*		Ian Foster	J. A. Joines	
H. Taureg	W. Braunschweig		Sophia Antipolis	III		Sebastian Thrun	B. Biller	
G. Tatishvili	G. Franke		Jack Dongarra	Assoc Prof		S. Idelsohn	J. D. Tew	
D. Madigojine	D. Clarke		Anthony M. Santomero	Vladik Kreinovich		Hector Garcia-Molina	J. Shortle	
F. Petrucci	L. Goerlich		M. Asce	Open Access Principal		Mario Gerla	M. -h. Hsieh	
S. Palestini	E. Gabathuler		Ian Foster	Investigator		Takeo Kanade	S. G. Henderson	
P. Dalpiaz	B. Andrieu		Turku Centre	Lt Col		Kang G. Shin	*Schrodinger International*	
M. Lenti	M. Erdmann		Takeo Kanade	Sophia Antipolis		Andrew B. Kahng	Lt Col	
I. Mikulec	G. Flügge		IEEE Computer Society	Gutachter Prof		Rwth Aachen	Jack Dongarra	
M. Savrie	J. Formánek		*Technische Hochschule*	John Wiley		Jason Cong	Supérieure Lyon	
D. Marras	R. Gerhards		Mario Gerla	David		Calton Pu	École Normale	
N. Molokanova	J. Gayler		Marie Curie	Civil Justice		Michael H. Bohlen	Supervisor Prof	
W. Funk	J. Feltesse		INRIA Rocquencourt	E. Onate		Manuela Veloso	Terrence J. Sejnowski	
C. Cheshkov	G. Bernardi		Scientiarum Fennicae	John		David E. Goldberg	Member IEEE	
O. Vossnack	J. Bürger		Politecnico Di Milano	J. A. Joines		Daniel Thalmann	Takeo Kanade	
R. Sacco	S. Burke		Sun Microsystems	Wildlife Service		Kristian Torp	Civil Justice	
V. Falaleev	U. Bassler		Alle Rechte Vorbehalten	Editorial Board		Heidi Gregersen	Ian Foster	

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

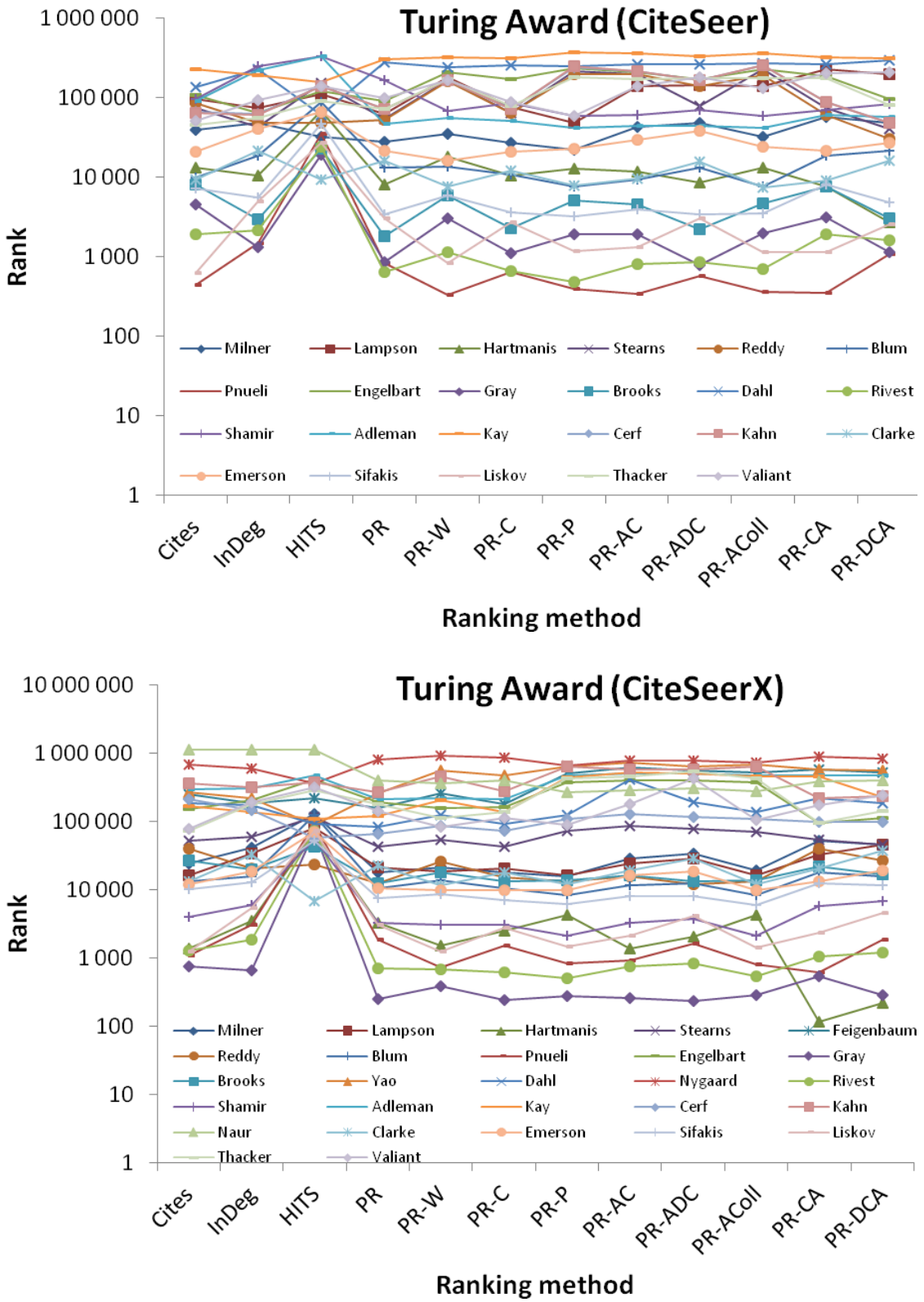


Figure 4: Ranks of Turing Award Winners by Various Methods in Both CiteSeers

Journal of Theoretical and Applied Information Technology
XX[st] Month 201x. Vol. x No.x

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

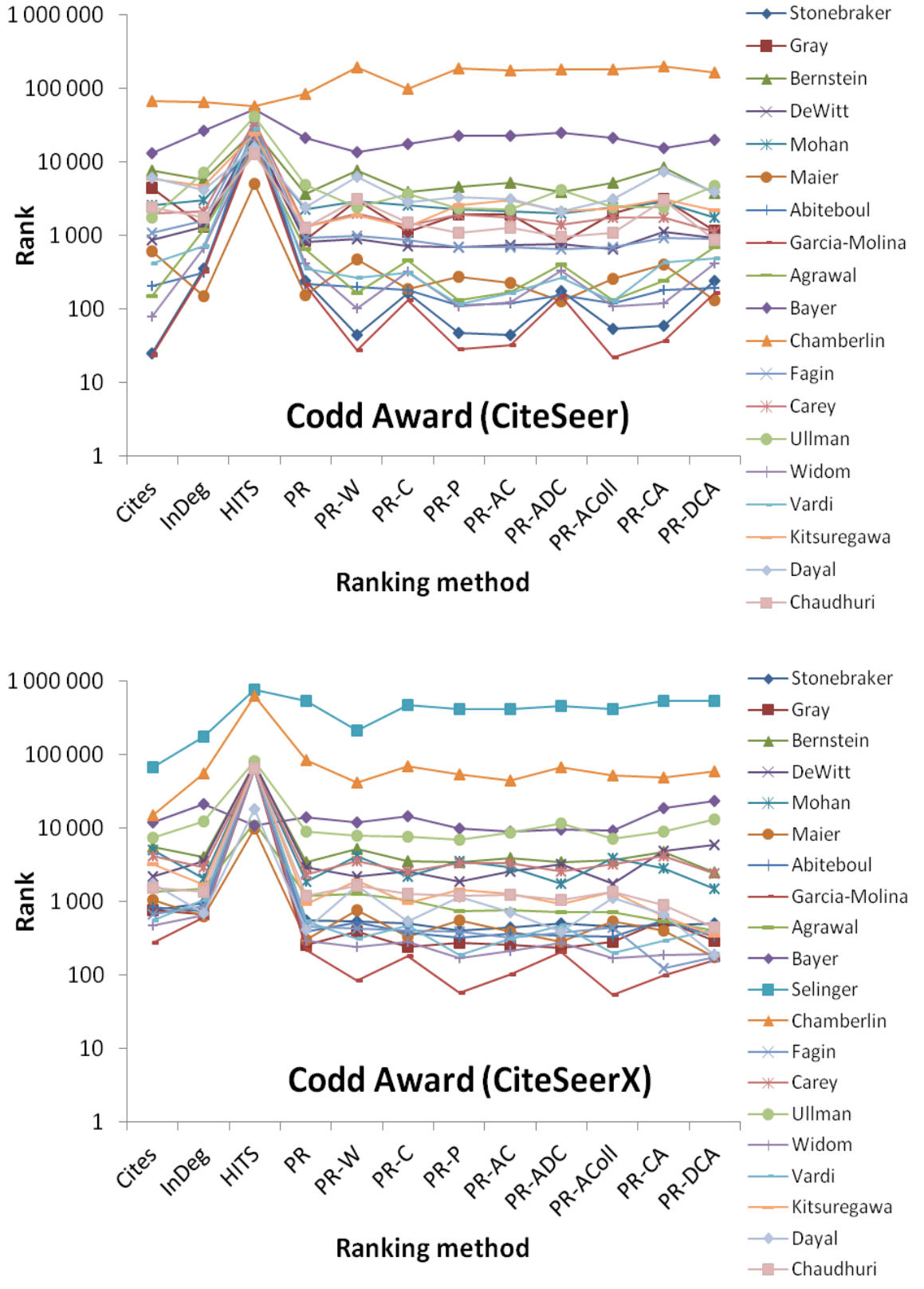


Figure 5: Ranks of Codd Award Winners byVarious Methods in Both CiteSeers

Journal of Theoretical and Applied Information Technology
XX[st] Month 201x. Vol. x No.x

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

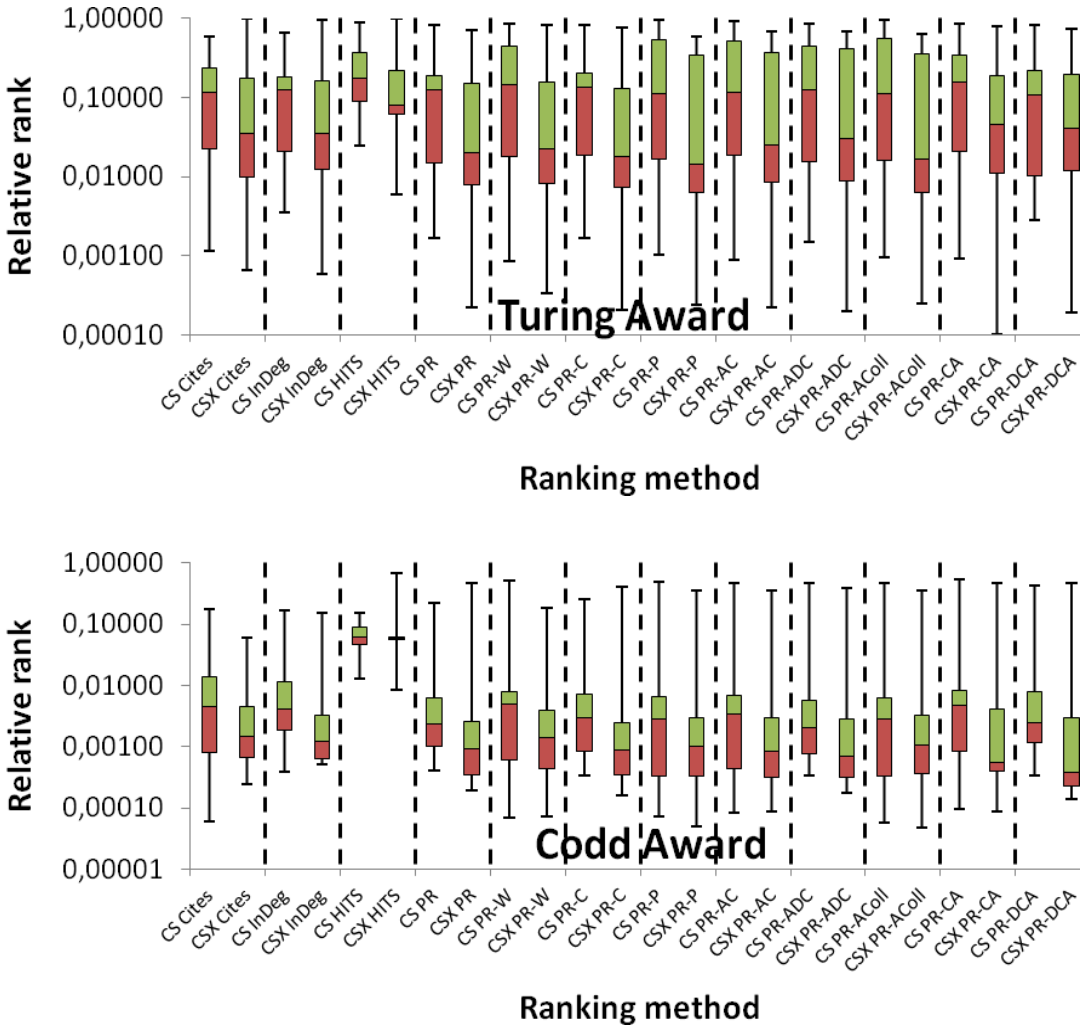


Figure 6: Boxplots of Relative Ranks Generated by Various Methods for Award Winners in Both CiteSeers

To answer the research question which of the two CiteSeers is better suited to evaluate computer science researchers, let us have a look at figure 6 and figure 7, in which charts comparing the ranks of Turing Award and Codd Award winners based on both CiteSeers are presented. figure 6 shows two boxplot charts (with the Y-axis on the logarithmic scale) depicting the relative ranks generated for the award winners by 12 methods in each CiteSeer. Thus, there are 24 different rankings for each of the awards. Relative ranks instead of absolute ranks are needed because the total number of researchers in CiteSeer and CiteSeer^X differs as explained earlier. In general, the ranks based on CiteSeer^X tend to be better (i.e., closer to 0) than those based on CiteSeer as we can see from the boxplots. We can also observe that the relative median rank of Turing Award winners in both CiteSeers roughly falls within top 10% and the relative median rank of Codd Award winners in both CiteSeers roughly falls within top 1% (except HITS). This might suggest that the coverage of general computer science literature (including theoretical computer science relevant to the Turing Award) in both CiteSeers is weaker than the coverage of database literature (relevant to the Codd Award). Another explanation may be that the Turing Award is a more life-time achievement prize than the Codd Award and that the main body of work of Turing Award winners was published in the years out of the scope of both CiteSeers. Similarly, the relative average and median ranks produced by 12 methods from two CiteSeer data sets for the winners of two awards are displayed in the charts in figure 7. Here the ranks of Turing Award winners based on CiteSeer^X are always clearly better than CiteSeer-based ranks and the ranks of Codd Award winners based on CiteSeer^X

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

are generally better than those in CiteSeer with the most notable exception being the relative average rank by HITS. As the basic characteristics of the coauthorship networks of both CiteSeers are similar (except for their size), the cause of the better ranks in CiteSeer^X seems to be its broader coverage of the relevant computer science literature.

4. CONCLUSIONS AND FUTURE WORK

CiteSeer and its current (yet still beta) version CiteSeer^X is a digital library and a search engine for computer science literature**,** whose article metadata have been successfully used for various purposes in the past. Some of the studies based on its data have been of bibliometric nature investigating its citation or coauthorship graphs. This paper belongs to such studies. Whereas CiteSeer has been discontinued and its most recent data come from December 2005, CiteSeer^X has been continuously updated until now. This research is concerned with CiteSeer^X data harvested from its Open Archives Initiative collection in March 2011. The number of articles covered by CiteSeer^X almost doubled between 2005 and 2011 and, unfortunately, the structure of the metadata on these articles freely obtainable from the respective Web sites changed considerably. These modifications do not enable the 2011 data to be analyzed in the same way as the 2005 data. The greatest difference is the general lack of the information on cited references in the article metadata. This fact excludes the possibility of a direct analysis of the CiteSeer^X citation graph acquired in this way. As a result, only its coauthorship network can be examined. The main contributions of this research are the following:

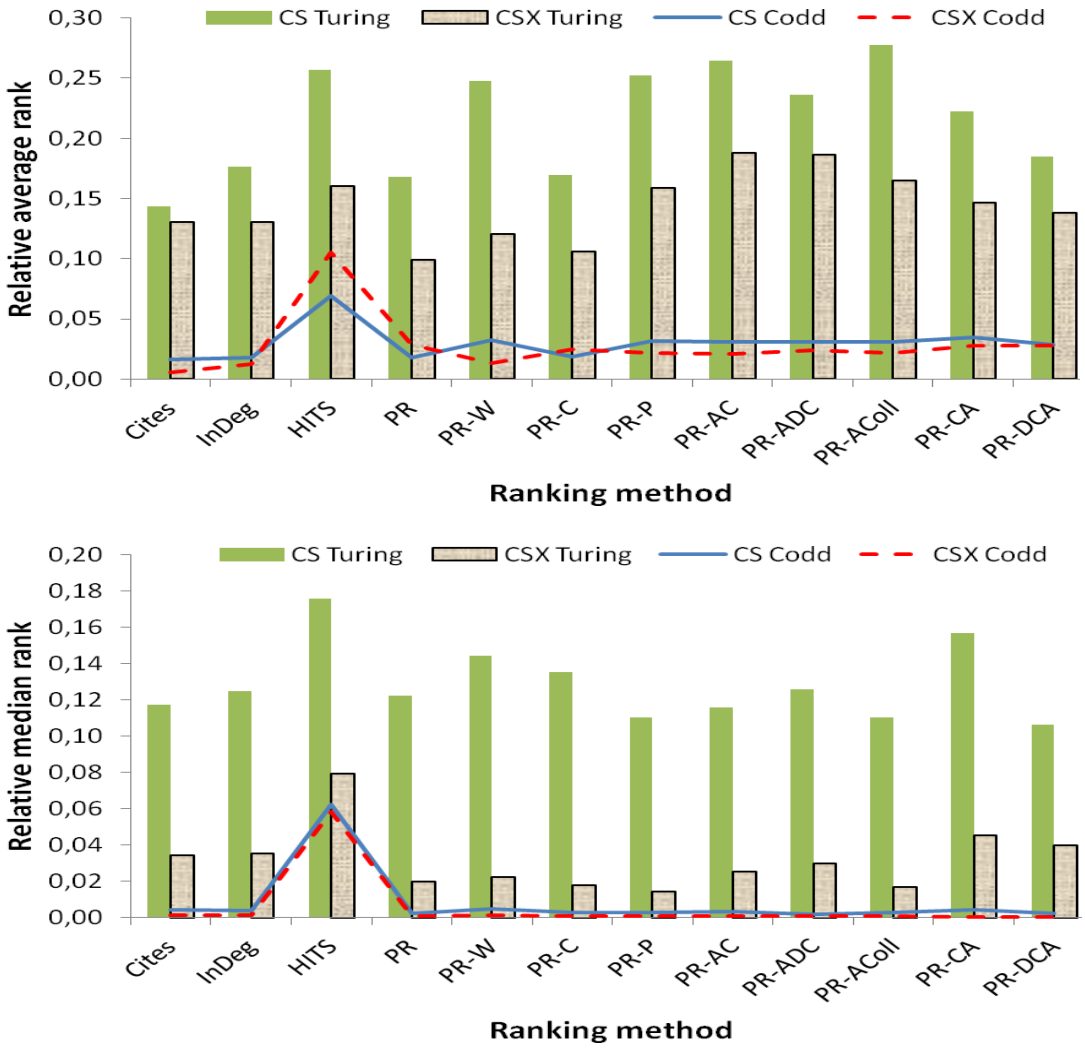


Figure 7: Relative Ranks by Various Methods for Award Winners in Both CiteSeers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- We compared the structure of the article metadata in CiteSeer and CiteSeer^X freely available via their Web sites and constructed coauthorship (or author collaboration) networks from both data sets.
- We treated the coauthorship networks as citation graphs (according to the model of Yan and Ding [25]) and created rankings of researchers using 12 different ranking methods such as citation counts, HITS, PageRank, or its variations.
- We concentrated on the ranks achieved by the winners of the ACM A. M. Turing Award from the years 1991 – 2010 and by the winners of the ACM SIGMOD E. F. Codd Innovations Award from the years 1992 – 2011 and compared the rankings in both CiteSeers.

We thereby obtained the following main results:

- The coauthorship graphs of both CiteSeers have similar characteristics, apart from their sizes (see figure 1, figure 2, and figure 3).
- The basic properties of the individual rankings based on coauthorship networks are the same as of those previously reported that were based on citation networks, which may indicate the usefulness of coauthorship networks for the ranking of researchers (see figure 4 and figure 5).
- The relative ranks of both Turing Award and Codd Award winners based on CiteSeer^X are generally better than CiteSeer-based ranks presumably resulting from the broader coverage of the relevant computer science literature in CiteSeer^X (see figure 6 and figure 7).

In the future, a natural continuation of this research would be the acquisition of the complete CiteSeer^X citation graph and its thorough analysis. It would be interesting to see how different the researcher rankings are between CiteSeer and CiteSeer^X (based on their citation graphs) and between CiteSeer^X (based on the citation graph) and CiteSeer^X (based on the coauthorship graph).

Acknowledgements. 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] CiteSeer, 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, 1999, pp. 67-71.

[3] CiteSeer^X, http://citeseerx.ist.psu.edu.

[4] 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, 2004, pp. 664–678.

[5] S. Chakrabarti and A. Agarwal, "Learning parameters in entity relationship graphs from ranking preferences", *Lecture Notes in Computer Science*, Vol. 4213, 2006, pp. 91–102.

[6] S. Chakrabarti, A. Pathak, and M. Gupta, "Index design and query processing for graph conductance search", *VLDB Journal*, Vol. 20, No. 3, 2011, pp. 445-470.

[7] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, 2004, pp. 5249–5253.

[8] A. Joorabchi and A.E. Mahdi, "An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata", *Journal of Information Science*, Vol. 37, No. 5, 2011, pp. 499-514.

[9] A. Popescul, L.H. Ungar, S. Lawrence, and D.M. Pennock, "Statistical relational learning for document mining", *Proceedings of the Third IEEE International Conference on Data Mining*, Melbourne (USA), 2003, pp. 275–282.

[10] T. Šingliar and M. Hauskrecht, "Noisy-OR component analysis and its application to link analysis", *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 2189–2213.

[11] D. Zhou, I. Councill, H. Zha, and C.L. Giles, "Discovering temporal communities from social network documents", *Proceedings of the Seventh IEEE International Conference on Data Mining*, Omaha (USA), 2007, pp. 745–750.

[12] H.-H. Chen, L. Gou, X. Zhang, and C.L. Giles, "CollabSeer: A search engine for collaboration discovery", *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, Ottawa (Canada), 2011, pp. 231-240.

[13] Q. He, D. Kifer, J. Pei, P. Mitra, and C.L. Giles, "Citation recommendation without author supervision", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, Hong Kong (China), 2011, pp. 755-764.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[14] C.-L. Wu and J.-L. Koh, "Hierarchical topic-based communities construction for authors in a literature database", *Lecture Notes in Computer Science*, Vol. 6097, 2010, pp. 514–524.

[15] J. Bar-Ilan, "An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes", *Information Processing and Management*, Vol. 42, No. 6, 2006, pp. 1553–1566.

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

[17] 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, 2004, pp. 17599–17604.

[18] 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, 2001, pp. 661–675.

[19] D. Zhao, "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 and Management*, Vol. 41, No. 6, 2005, pp. 1403–1418.

[20] 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, 2002, pp. 449–472.

[21] D. Zhao and A. Strotmann, "Can citation analysis of web publications better detect research fronts?", *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 9, 2007, pp. 1285–1302.

[22] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M.-T. Hütt, "Motifs in co-authorship networks and their relation to the impact of scientific publications", *European Physical Journal B*, Vol. 84, No. 4, 2011, pp. 535-540.

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

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

[25] E. Yan and Y. Ding, "Discovering author impact: A PageRank perspective", *Information Processing and Management*, Vol. 47, No. 1, 2011, pp. 125-134.

[26] A. Sidiropoulos and Y. Manolopoulos, "A citation-based system to assist prize awarding", *SIGMOD Record*, Vol. 34, No. 4, 2005, pp. 54–60.

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

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

[29] CiteSeer[X] OAI, http://citeseerx.ist.psu.edu/oai2.

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

[31] 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, 1998, pp. 107-117.