Bibliometric analysis of CiteSeer data for countries

Dalibor Fiala

University of West Bohemia, Univerzitní 8, 30614 Plzeň, Czech Republic
Phone: +420 377 63 24 29, fax: +420 377 63 24 01, email: dalfia@kiv.zcu.cz

Abstract: This article describes the results of our analysis of the data from the CiteSeer digital library. First, we examined the data from the point of view of source top-level Internet domains from which the data were collected. Second, we measured country shares in publications indexed by CiteSeer and compared them to those based on mainstream bibliographic data from the Web of Science and Scopus. And third, we concentrated on analyzing publications and their citations aggregated by countries. This way, we generated rankings of the most influential countries in computer science using several non-recursive as well as recursive methods such as citation counts or PageRank. We conclude that even if East Asian countries are underrepresented in CiteSeer, its data may well be used along with other conventional bibliographic databases for comparing the computer science research productivity and performance of countries.

Keywords: CiteSeer, CiteSeer^{X}, citations, shares, countries, Internet domains.

1. Introduction

CiteSeer (CiteSeer) is a vast free Web digital library and search engine of mainly computer science papers that have been automatically acquired from various Web sites, stored, and analyzed to allow for searching and exploring its bibliographic data. Despite its free on-line as well as off-line availability and well structured data, it has been relatively rarely used in bibliometric studies particularly due to fears of incomplete and erroneous machine-generated data. We refer to the work by Fiala (2011) where a detailed overview of CiteSeer's features in the context of other established bibliographic databases is given.

The purpose of this study is to show: a) where CiteSeer has got its data (i.e. which Web domains it has visited to obtain them), b) which countries have contributed most to its digital library (in terms of the number of papers published by authors from these countries), and c) which countries have the most influence (in terms of citedness of "their" publications). We have thoroughly analyzed the CiteSeer data file from December 13, 2005 and have made

a quick look at the newer data provided by CiteSeer^X (CiteSeer^X) which replaced CiteSeer in April 2010 but is still a beta version at the time of writing this article (May 2011).

2. Related work

There have been a number of studies of research productivity (publications) and impact (citations) at the level of countries in recent years. There is a growing need for such scientometric indicators because they often reflect the quality of science policy in a specific country and may have influence on changes in science funding. From the many research papers discussing this topic, let us mention just one of the most recent by Albarrán et al. (2010), which compares the United States to the European Union in a detailed way in various fields of science.

While quite a lot of research efforts have been devoted to bibliometrics of chemistry, biology, or humanities, relatively few scientometric studies have been concerned with the field of computer science. Bakri & Willett (2011) measure the performance of computer science research in Malaysia and Gupta et al. (2011) analyze the research output of Indian computer science. Wainer et al. (2009) compared the Brazilian computer science production to twelve other countries. Ma et el. (2008) did not limit their analysis to a particular country but evaluated the computer science research performance of universities around the globe and Guan & Ma (2004) evaluated China and five other countries. Different sources of bibliographic data for the scientometric evaluation of computer science publications were examined by Bar-Ilan (2010) and by Franceschet (2010). The latter author also presents an overview of literature comparing citation data from various data sources for a specific scientific field. Furthermore, Franceschet (2010b) investigated the influence of computer science journal and conference papers on the scientific community.

Unlike our paper, most of the articles above have mainly exploited the well-known and manually-maintained bibliographic database Web of Science (Web of Science) or its variants. As far as CiteSeer as a data source is concerned, some researchers have already used it for bibliometric purposes: Zhou et al. (2007) explored CiteSeer documents to discover temporal communities of collaborating authors in the domains of databases and machine learning. On the other hand, Hopcroft et al. (2004) tracked evolving communities in the whole CiteSeer paper citation graph. An et al. (2004) conducted a component analysis of the CiteSeer paper citation graph in several research domains and CiteSeer^X data were used by Wu et al. (2010) in order to enhance collaborative networks with topic information. Zhao & Strottman (2007) and Zhao & Logan (2002) analyzed co-citations in CiteSeer documents in

the XML research field and a similar study for computer graphics was reported by Chen (2000). Bar-Ilan (2006) used CiteSeer data for a citation analysis of the works of a famous mathematician. A kind of citation analysis for acknowledgements was also performed by Giles & Councill (2004). Feitelson & Yovel (2004) examined citation ranking lists obtained from CiteSeer and predicted future rankings of authors.

Our study is the first of its kind that attempts to measure the productivity and impact of computer science research conducted by countries by analyzing CiteSeer data.

3. Data

The last CiteSeer data originate from December 2005 and they contain roughly 717 thousand publications with 1.8 million references within CiteSeer. On the other hand, CiteSeer^X (data from March 2011) provides more than 1.3 million publications with almost 15 million references within CiteSeer^X. This means that the citation graph with publications as nodes and references as edges has become much denser over the past six years – the mean number of references in a publication increased from 2.5 in 2005 to 11.2 in 2011.

Let us have a look at a few obvious differences between CiteSeer (CS) / CiteSeer^X (CS^X) and Web of Science / Scopus (Scopus) – two well-known databases of scientific literature. Both CiteSeer and CiteSeer^X collect (or collected) its data in the same way: they crawl the Web starting from some seed pages submitted by their engineers or by individual users (authors) and pick up freely accessible documents (mostly PDF or PostScript files) that have the potential to be research papers in computer science, mathematics, or related fields. Web crawling as well as information extraction (titles, author names, references, etc.) occurs automatically, without human intervention. The contents of CiteSeer and CiteSeer^X depend generally on the content and structure of the Web. On the other hand, both Web of Science and Scopus use a great deal of human labour to receive publications (mainly journal issues and conference proceedings) and to index them. Unlike CiteSeer and CiteSeer^X, WoS and Scopus cover all scientific fields. Which publication sources are indexed and which are not is decided by the editorial boards of both "human-made" databases. Another big difference between CiteSeer and CiteSeer^X on one side and WoS and Scopus on the other is that the first two are free whereas the latter two are subscription-based.

4. Methods

4.1 Data collection

Data collection methods were different for CiteSeer and for CiteSeer^X. For CiteSeer, there was a single archive data file created in December 2005 (the most recent CiteSeer data) that we merely downloaded from the CiteSeer Web site and unpacked into 2 GB of 72 XML-like files. As for CiteSeer^X, we were forced to use one of the harvesting tools referenced on its Web site to gain off-line access to its current repository. The harvest itself took a few days in March 2011 and resulted in a regular 3.7 GB XML file which we further split up into 73 files to process them more smoothly in main memory. We developed software[1] that parsed the data files and stored information about publications, authors, and citations in a relational database. We were then able to query the database and obtain the information presented in the following sections. The software also had capabilities to compute more complex values such as HITS and PageRank.

4.2 Internet domains and countries

Gathering statistics about Internet top-level domains (TLD) is quite smooth and accurate given that the "source" property for each document is almost always present and error free. The situation gets considerably worse when we try to assemble similar statistical data for the distribution of countries whose authors produced the publications collected by CiteSeer. As far as CiteSeer^X is concerned, unfortunately, it does not provide any information on the addresses or affiliations of the authors of its publications – not only for "new" publications, but also for "old" publications for which this information is present in CiteSeer. Therefore, we could not use CiteSeer^X data for our experiments with countries. Let us hope that future versions of CiteSeer^X (the current one is still a beta) will have such information included.

4.3 Missing data and name unification

In CiteSeer, there is a problem with missing data. For almost each document, there are authors assigned to it but only for some of the authors there is also an address affiliated with him/her. Strictly said, from the total of 1.66 million authors (without any name unification or disambiguation), we had no address information at our disposal for about 690 thousand or 42% of them, let alone the accuracy of such information.

Thus, to obtain the data shown later in Figure 2, we proceeded in the following way: We discarded publications without any address information for any of its authors. This resulted in only 439 thousand being kept. (For these publications, one author at least had some address information included.) Then, we tried to unify country names used in the addresses. This task consisted in obtaining a list of countries and territories owning a top-level Internet

[1] http://textmining.zcu.cz/downloads/sciento.php

domain. After some cleansing, 243 countries or territories were left. Next, we attempted to unify country names by replacing common synonymic variants of each of those 243 countries with one standard name.

For instance, in the case of the United States of America, we had to count in names like “United States”, “U.S.A.”, “U.S.A”, “U.S.”, “USA”, or “US”. Since U.S. postal addresses often do not contain any mention of “USA” or its variants and only display the name or abbreviation of a federal state such as “California” or “CA”, we also needed to take this into account and counted such occurrences as “USA”. Other types of unification included considering often independently appearing entities such as England, Scotland, Wales and Northern Ireland as one country (United Kingdom) or, in contrast, keeping territories of one country separate such as Hong Kong, Taiwan, and Macau from China or Reunion and Martinique from France. Finally, we processed international postal country codes in the addresses as well, thus yielding Czech Republic for an address “CZ-30416” with respect to the prefix “CZ-” as an example.

4.4 Comparison with the Web of Science and Scopus

Since the CiteSeer data we examined were from December 2005, we restricted our analysis to a 10-year period from 1996 to 2005. This decade is the most probable one, in which CiteSeer was collecting its documents. Moreover, Scopus itself does not generally capture citations to documents published before 1996, which is also a good reason for 1996 as a decade’s start with regard to possible future comparisons of citations. In September 2010, we were querying on-line Web services of both WoS and Scopus and generated the rankings in Tables 3 and 4. As for WoS, we opted to limit our search to the “Science Citation Index Expanded” database, to the “article” document type, and to the publications from the journals included in the seven computer science subject categories of the Journal Citation Reports® Science Edition 2009. In this way, we arrived at the total of 148 838 publications, which is 100% for the relative shares in Table 3. As far as Scopus is concerned, querying was easier in that the subject area (computer science) could be specified directly in the query and the exact results number was always disclosed. The final 325 614 “article” documents form 100% for the relative shares in Table 4. Due to the search limits of both WoS and Scopus, it was sometimes necessary to split up “big” queries into subqueries and to combine their results.

Alternatively, WoS as well as Scopus provide programming interfaces that enable submitting queries and obtaining results without needing to interact with their Web front-ends. However, the basic APIs included in the subscription do have queries and results restrictions that are similar to those on their Web sites.

4.5 Citations and recursive indicators

In addition to measuring shares of individual countries in the publications indexed by CiteSeer, we wished to determine the influence of countries by examining citations they receive. Thus, we derived a citation graph of countries from the citation graph of publications. In the directed publication citation graph, there were 717 thousand nodes (publications) and 1.76 million edges (citations between publications). This accounts for roughly 2.45 citations per paper so, obviously, many citations (or references) are missing in CiteSeer. Let us recall that addresses of publications' authors were normalized by the approach described earlier. We aggregated citations by the country of the source and target publication. If there were more countries associated with a publication, a couple of citations came into being. We removed self-citations of countries as well.

Besides first-order methods such as in-degree and citations, there are recursive techniques as well that not only count citations but take also into account whether the citing node itself is frequently cited. Some of these methods are HITS introduced by Kleinberg (1999), PageRank defined by Brin and Page (1998), or weighted PageRank (e.g., Fiala et al., 2008). We applied these methods to the normalized country citation graph from CiteSeer and present the country rankings obtained in Table 6.

5. Results and discussion

5.1 Internet domains

One of the properties of each document item indexed by CiteSeer is its source. This is the URL (a Web page) from which the document has originally been downloaded. We were interested in the distribution of Internet top-level domains (TLD) among the sources of CiteSeer documents. This would reveal what regions of the Web the CiteSeer Web crawler has visited and to what extent. It might also help explain a possible bias in publication and citations shares of individual countries discovered later.

Figure 1 shows the shares of top twenty top-level Internet domains as sources of CiteSeer and CiteSeer^{X} documents. The charts are quite similar - approximately one third of all publications originate from *.edu* servers, followed by *.de*, *.uk*, *.fr*, and *.com* with the most notable change for *.org*, which grew from 3.62% to 9.42% between 2005 and 2011. Although *.edu*, *.com*, and *.org* domains do not necessarily mean U.S. Web sites, we shall not be too far from the truth if we count them along with *.gov* as U.S. sites and claim that about a half of all CiteSeer documents have been gathered in the United States with a small increase by several percentage points from 2005 to 2011. In 2005, only 25 documents had no source URL

affiliated with them and they are included in those almost 10% of "other" domains. In 2011, this number is considerably higher – almost 17 thousand – and the share of "other" domains is as much as 12%. A complete list of the top 100 CiteSeer source domains is available in Table 1 with their respective ranks and shares in CiteSeer^{X}. After a quick look at the table, we may notice that a couple of non-country TLDs have significantly increased their shares such as *.org* (moving from rank 6 to rank 2), *.net* (from 29 to 18), or *.info* (from 62 to 41) while the main country-code TLDs remain relatively stable or even slightly decline. There is one remarkable exception, *.in*, which increases its rank from 37 to 25 and its share from 0.20% to 0.55% between the years 2005 and 2011. In this context, it is interesting to see that the postion of *.cn* (38) remains unchanged in both CiteSeer and CiteSeer^{X}.

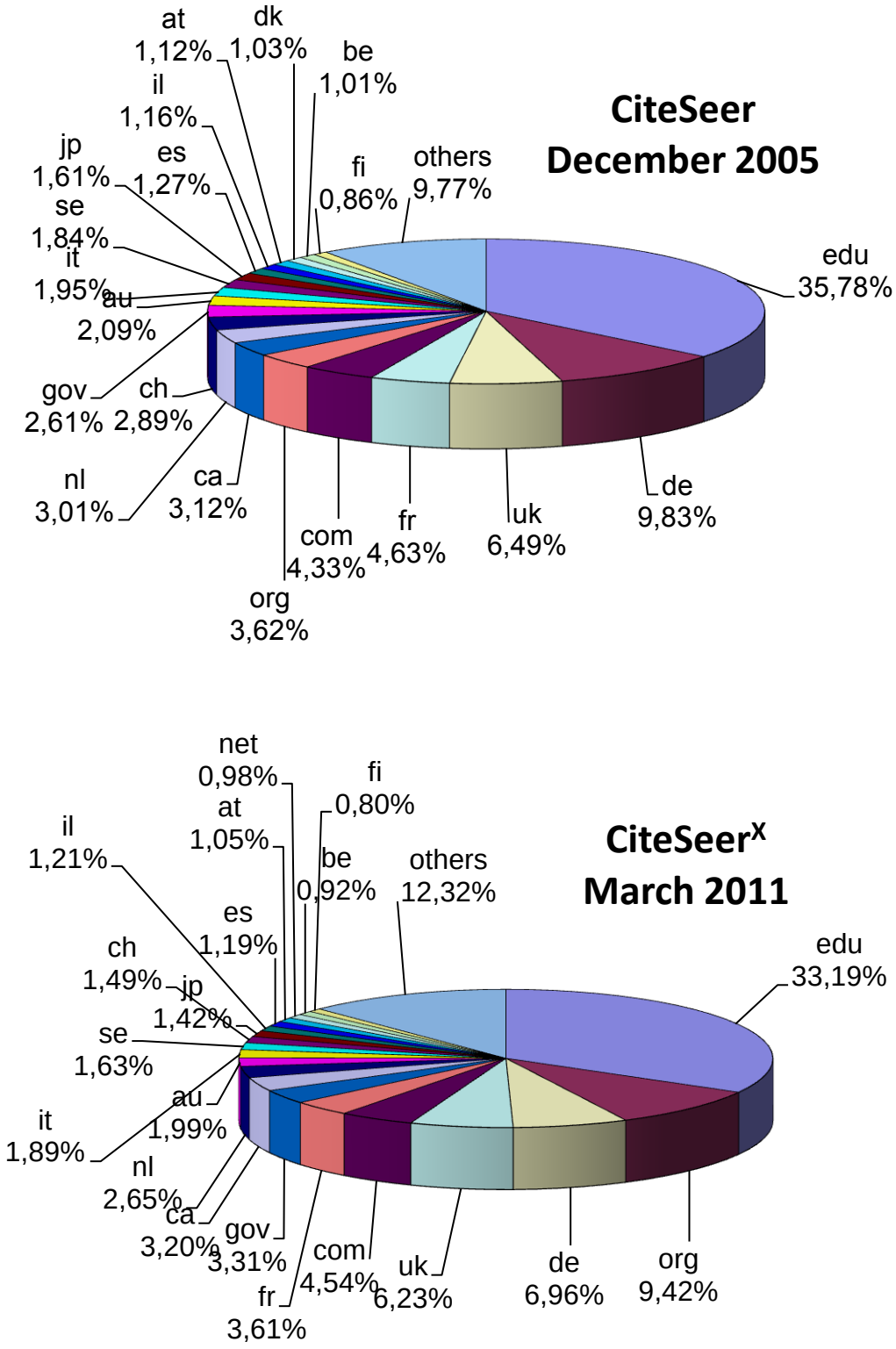

Fig. 1 Shares of Internet domains from which CiteSeer and CiteSeer^{X} documents have been collected

Table 1 Top 100 Internet top-level domains (TLD) by publications in CiteSeer compared to CiteSeer^{X}

Dec 2005				March 2011			Dec 2005				March 2011		
No.	TLD	# Pub.	%	No.	# Pub.	%	No.	TLD	# Pub.	%	No.	# Pub.	%
1	edu	256 433	35.78	1	442 756	33.19	51	ua	273	0.04	68	280	0.02
2	de	70 446	9.83	3	92 821	6.96	52	ar	226	0.03	52	739	0.06
3	uk	46 544	6.49	4	83 049	6.23	53	hr	197	0.03	55	546	0.04
4	fr	33 172	4.63	6	48 190	3.61	54	cy	175	0.02	63	338	0.03
5	com	31 051	4.33	5	60 570	4.54	55	yu	165	0.02	58	488	0.04
6	org	25 922	3.62	2	125 650	9.42	56	uy	146	0.02	69	175	0.01
7	ca	22 368	3.12	8	42 671	3.20	57	ee	137	0.02	56	536	0.04
8	nl	21 544	3.01	9	35 411	2.65	58	ir	135	0.02	54	548	0.04
9	ch	20 686	2.89	13	19 908	1.49	59	bg	116	0.02	60	369	0.03
10	gov	18 694	2.61	7	44 179	3.31	60	co	109	0.02	81	79	0.01
11	au	14 976	2.09	10	26 547	1.99	61	ve	105	0.01	72	123	0.01
12	it	13 976	1.95	11	25 188	1.89	62	info	91	0.01	41	2 507	0.19
13	se	13 178	1.84	12	21 721	1.63	63	lv	65	0.01	71	155	0.01
14	jp	11 522	1.61	14	18 911	1.42	64	my	65	0.01	59	462	0.03
15	es	9 092	1.27	16	15 851	1.19	65	py	54	0.01	169	0	0.00
16	il	8 287	1.16	15	16 162	1.21	66	to	54	0.01	66	285	0.02
17	at	8 056	1.12	17	14 013	1.05	67	is	52	0.01	67	284	0.02
18	dk	7 360	1.03	21	10 250	0.77	68	lt	52	0.01	53	581	0.04
19	be	7 270	1.01	19	12 261	0.92	69	ps	51	0.01	65	286	0.02
20	fi	6 145	0.86	20	10 705	0.80	70	lu	46	0.01	75	109	0.01
21	kr	4 791	0.67	29	6 404	0.48	71	mt	30	0.00	77	106	0.01
22	gr	4 336	0.60	22	9 077	0.68	72	mk	27	0.00	87	50	0.00
23	pt	4 229	0.59	24	7 604	0.57	73	lb	26	0.00	72	123	0.01
24	no	3 977	0.55	27	6 697	0.50	74	ma	26	0.00	79	95	0.01
25	br	3 973	0.55	31	6 109	0.46	75	ph	25	0.00	76	107	0.01
26	cz	3 844	0.54	30	6 305	0.47	76	gb	24	0.00	93	24	0.00
27	ie	3 522	0.49	26	6 708	0.50	77	nu	21	0.00	80	91	0.01
28	hk	3 470	0.48	23	7 759	0.58	78	et	18	0.00	106	14	0.00
29	net	2 847	0.40	18	13 091	0.98	79	aero	15	0.00	109	11	0.00
30	mil	2 527	0.35	35	4 054	0.30	80	fm	15	0.00	62	345	0.03
31	nz	2 427	0.34	28	6 448	0.48	81	id	15	0.00	78	96	0.01
32	pl	2 202	0.31	34	4 417	0.33	82	sa	15	0.00	57	514	0.04
33	tw	2 056	0.29	33	4 981	0.37	83	biz	10	0.00	88	49	0.00
34	mx	1 978	0.28	42	2 301	0.17	84	cu	10	0.00	98	20	0.00
35	hu	1 905	0.27	37	3 805	0.29	85	name	10	0.00	64	290	0.02
36	sg	1 725	0.24	32	5 572	0.42	86	rs	10	0.00	102	15	0.00
37	in	1 423	0.20	25	7 342	0.55	87	tc	10	0.00	99	19	0.00
38	cn	1 265	0.18	38	3 396	0.25	88	ws	9	0.00	84	65	0.00
39	tr	1 208	0.17	40	2 800	0.21	89	mu	6	0.00	113	9	0.00
40	ru	1 176	0.16	44	1 892	0.14	90	mo	5	0.00	85	61	0.00
41	cl	1 054	0.15	47	1 657	0.12	91	om	5	0.00	122	5	0.00
42	si	801	0.11	43	1 900	0.14	92	li	4	0.00	113	9	0.00
43	za	785	0.11	45	1 735	0.13	93	tv	4	0.00	96	21	0.00
44	int	621	0.09	39	3 256	0.24	94	ac	3	0.00	110	10	0.00
45	th	474	0.07	51	844	0.06	95	af	3	0.00	126	4	0.00
46	us	462	0.06	36	3 954	0.30	96	cx	3	0.00	100	18	0.00
47	sk	459	0.06	49	1 439	0.11	97	pg	3	0.00	169	0	0.00
48	su	447	0.06	61	353	0.03	98	ae	2	0.00	86	52	0.00
49	cc	333	0.05	48	1 630	0.12	99	am	2	0.00	119	7	0.00
50	ro	277	0.04	50	1 014	0.08	100	ge	2	0.00	102	15	0.00

Nowadays, most open access repositories are located within North America and Europe (Repository66) and, therefore, it is logical that even Asian researchers might prefer placing their manuscripts in the repositories of these regions, which further increases the prevalence of American and European top-level Internet domains crawled by CiteSeer.

5.2 Countries

After unifying country names in the available addresses as described in Section 4.3, we tried to assign all 439 thousand publications to one or more country depending on how many authors from which countries they had. About 25 thousand publications could not be assigned to any country, i.e. it was impossible to make use of the information in their address field to identify a standard country by the above approach. Thus, only 414 thousand documents (58% of 717 thousand) were finally assigned to one or more country. We counted the assignments to countries and found out country shares that are demonstrated relatively as well as absolutely in Figure 2 and in Table 2. Note, however, that the relative shares in Figure 2 differ from those presented in Table 2.

The relative shares in Figure 2 sum up to 100% constituted by a total of 449 thousand publication-country assignments, which is not equal to 414 thousand publications due to international co-authorships. (Albarrán et al. (2010) call the publication-country assignments "extended articles".) Even though the number of such assignments is only less than 10% greater than that of publications, it does not necessarily imply a relatively low number of international publications in CiteSeer. We may rather assume that addresses in international papers are more difficult to be processed by a machine (CiteSeer) and, therefore, they are often missing or erroneous and do not appear in our cleansed data.

In Figure 2, the top twenty most represented countries take almost 93% of "extended articles". The first country is the United States with a four-fold greater share (42.59%) than the second most "prolific" country – Germany (10.65%). At the third position, there is a tie between France and the United Kingdom (both 5.35%). As a remarkable point, two developing countries have entered the Top 20 – India and Brazil with shares of 0.67% and 0.64%, respectively. The number (or share) of publications not assigned to any country is not visible in Figure 2.

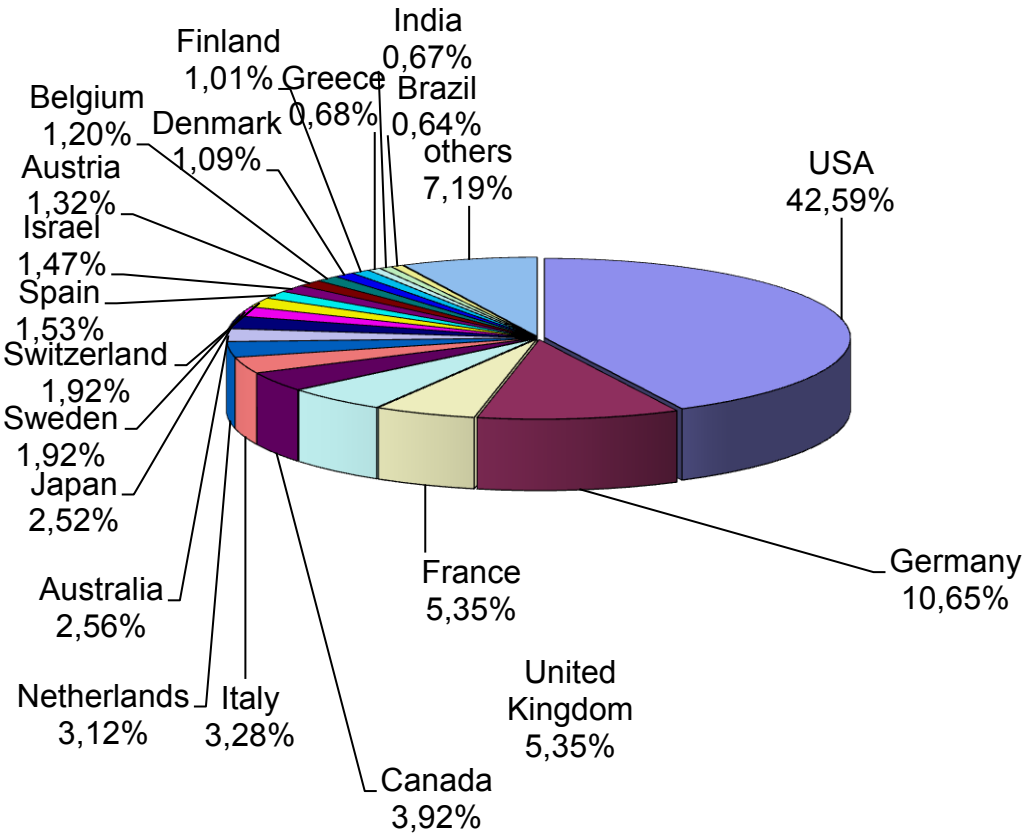

Fig. 2 Shares of countries to which publications are assigned in CiteSeer

The relative shares in Table 2 are smaller than those in Figure 2 because the base (100%) is much larger – 717 thousand, which is the original number of CiteSeer documents. These relative shares are important for they help us compare CiteSeer publication shares with those from the Web of Science and Scopus where the number of all documents can be determined, but the number of publication-country assignments is unknown. The absolute numbers in Table 2 are the numbers of publications assigned to a country and they were input in Figure 2. If, hypothetically, each CiteSeer article was assigned to exactly one country, the sum of counts in Table 2 would be approximately 717 thousand and the total share 100% (the rest after rank 100 is negligible). If each document was assigned to two or more countries (i.e. all papers are internationally co-authored), the sum of counts would be more than 717 thousand and the total share more than 100 %. A further discussion of the results in Table 2 will follow in the next section along with a comparison to the Web of Science and Scopus.

Table 2 Top 100 countries by publications in CiteSeer

Rank	Country	Public.	Share	Rank	Country	Public.	Share
1	USA	191 363	26.70%	51	Belarus	119	0.02%
2	Germany	47 866	6.68%	52	Venezuela	114	0.02%
3	France	24 052	3.36%	53	Egypt	107	0.01%
4	United Kingdom	24 042	3.35%	54	Latvia	102	0.01%
5	Canada	17 630	2.46%	55	Uruguay	96	0.01%
6	Italy	14 718	2.05%	56	Serbia and Mont.	94	0.01%
7	Netherlands	14 022	1.96%	57	Lithuania	93	0.01%
8	Australia	11 496	1.60%	58	Lebanon	66	0.01%
9	Japan	11 328	1.58%	59	Tunisia	66	0.01%
10	Sweden	8 639	1.21%	60	Colombia	60	0.01%
11	Switzerland	8 611	1.20%	61	Malta	60	0.01%
12	Spain	6 876	0.96%	62	Armenia	55	0.01%
13	Israel	6 616	0.92%	63	Iceland	55	0.01%
14	Austria	5 934	0.83%	64	Panama	53	0.01%
15	Belgium	5 411	0.75%	65	Vietnam	44	0.01%
16	Denmark	4 882	0.68%	66	Cuba	42	0.01%
17	Finland	4 533	0.63%	67	Morocco	39	0.01%
18	Greece	3 038	0.42%	68	Macau	37	0.01%
19	India	3 002	0.42%	69	Pakistan	36	0.01%
20	Brazil	2 889	0.40%	70	Indonesia	34	0.00%
21	Portugal	2 650	0.37%	71	Saudi Arabia	34	0.00%
22	Russia	2 351	0.33%	72	Puerto Rico	32	0.00%
23	Hong Kong	2 238	0.31%	73	Philippines	31	0.00%
24	Norway	2 215	0.31%	74	Kuwait	30	0.00%
25	Singapore	1 897	0.26%	75	Algeria	25	0.00%
26	Taiwan	1 808	0.25%	76	Bangladesh	24	0.00%
27	New Zealand	1 703	0.24%	77	Costa Rica	23	0.00%
28	China	1 600	0.22%	78	Jordan	21	0.00%
29	Poland	1 564	0.22%	79	Kenya	14	0.00%
30	Czech Republic	1 453	0.20%	80	Liechtenstein	14	0.00%
31	South Korea	1 450	0.20%	81	Macedonia	14	0.00%
32	Hungary	1 423	0.20%	82	Nigeria	14	0.00%
33	Ireland	1 366	0.19%	83	Moldova	13	0.00%
34	Mexico	1 071	0.15%	84	Oman	11	0.00%
35	Turkey	775	0.11%	85	Cameroon	9	0.00%
36	Slovenia	659	0.09%	86	Jamaica	9	0.00%
37	Chile	489	0.07%	87	Martinique	9	0.00%
38	South Africa	472	0.07%	88	Netherlands Antilles	9	0.00%
39	Romania	450	0.06%	89	Sri Lanka	9	0.00%
40	Argentina	445	0.06%	90	Reunion	8	0.00%
41	Thailand	335	0.05%	91	United Arab Emirates	8	0.00%
42	Ukraine	306	0.04%	92	Uzbekistan	8	0.00%
43	Bulgaria	299	0.04%	93	Ethiopia	7	0.00%
44	Cyprus	285	0.04%	94	Vatican	7	0.00%
45	Slovakia	250	0.03%	95	Bahrain	6	0.00%
46	Luxembourg	242	0.03%	96	Fiji	6	0.00%
47	Iran	215	0.03%	97	Guinea	6	0.00%
48	Croatia	149	0.02%	98	Mozambique	6	0.00%
49	Estonia	141	0.02%	99	Nicaragua	6	0.00%
50	Malaysia	131	0.02%	100	Uganda	6	0.00%

5.3 Comparison with the Web of Science and Scopus

To get a clue how reliable CiteSeer data are and to see how distant or close to other well-known bibliographic data sources they are, it was necessary to perform a couple of comparisons and measurements. Based on the amount of available information on publication shares of countries from the previous section, we decided to compare these country shares to those obtained from the Web of Science and Scopus – two established manually maintained bibliographic databases. The goal was to create rankings of countries by the number of "their" publications in the field of computer science and to compare them to the CiteSeer ranking in Table 2.

Table 3 Top 30 computer science countries by Web of Science in 1996 – 2005

Rank	Cite-Seer	Country	Publications	Share	Citations	Average citations	h-index
1	1	*USA*	52 579	35.33%	904 339	17.20	258
2	4	*United Kingdom*	11 515	7.74%	160 691	13.95	125
3	9	*Japan*	8 902	5.98%	72 379	8.13	82
4	2	*Germany*	8 554	5.75%	114 075	13.34	108
5		China	8 348	5.61%	92 050	11.03	86
6	5	*Canada*	7 630	5.13%	102 609	13.45	105
7	3	*France*	7 159	4.81%	97 801	13.66	102
8		Taiwan	6 690	4.49%	66 762	9.98	76
9	6	*Italy*	6 587	4.43%	76 837	11.66	87
10		South Korea	4 753	3.19%	42 720	8.99	65
11	12	*Spain*	4 421	2.97%	50 272	11.37	76
12	8	*Australia*	4 196	2.82%	54 625	13.02	82
13	7	*Netherlands*	3 503	2.35%	55 459	15.83	88
14	19	*India*	3 103	2.08%	27 613	8.90	55
15	13	*Israel*	3 014	2.03%	46 385	15.39	82
16		Singapore	2 695	1.81%	32 015	11.88	66
17		Russia	2 246	1.51%	7 879	3.51	33
18	18	*Greece*	2 153	1.45%	20 283	9.42	50
19	15	*Belgium*	1 849	1.24%	29 343	15.87	65
20	11	*Switzerland*	1 838	1.23%	37 542	20.43	78
21	10	*Sweden*	1 766	1.19%	23 825	13.49	57
22	20	*Brazil*	1 449	0.97%	14 601	10.08	46
23		Poland	1 440	0.97%	15 948	11.08	50
24	17	*Finland*	1 408	0.95%	23 137	16.43	59
25	14	*Austria*	1 357	0.91%	17 065	12.58	51
26		Turkey	1 284	0.86%	13 160	10.25	44
27	16	*Denmark*	1 045	0.70%	16 645	15.93	53
28		Hong Kong	858	0.58%	10 909	12.71	47
29		Ireland	806	0.54%	8 202	10.18	38
30		Hungary	791	0.53%	8 072	10.20	41

Table 4 Top 30 computer science countries by Scopus in 1996 - 2005

Rank	Cite-Seer	Country	Publications	Share	Citations	Average citations	h-index
1	1	*USA*	87 591	26.90%	1 731 096	19.76	360
2		China	26 004	7.99%	149 019	5.73	104
3	4	*United Kingdom*	21 545	6.62%	292 929	13.60	163
4	9	*Japan*	21 231	6.52%	141 346	6.66	106
5	2	*Germany*	18 125	5.57%	213 144	11.76	143
6	3	*France*	14 570	4.47%	187 746	12.89	136
7	5	*Canada*	13 001	3.99%	191 347	14.72	135
8	6	*Italy*	12 133	3.73%	147 608	12.17	117
9		South Korea	10 370	3.18%	84 225	8.12	91
10		Taiwan	10 238	3.14%	106 810	10.43	95
11	12	*Spain*	8 035	2.47%	87 291	10.86	94
12	8	*Australia*	7 105	2.18%	96 481	13.58	103
13	19	*India*	5 997	1.84%	58 432	9.74	80
14	7	*Netherlands*	5 966	1.83%	93 431	15.66	110
15		Hong Kong	5 382	1.65%	78 625	14.61	94
16		Russia	5 177	1.59%	16 783	3.24	45
17	13	*Israel*	4 767	1.46%	81 874	17.18	108
18		Singapore	4 230	1.30%	51 347	12.14	79
19	18	*Greece*	3 932	1.21%	38 669	9.83	66
20	10	*Sweden*	3 916	1.20%	69 242	17.68	85
21	11	*Switzerland*	3 618	1.11%	75 824	20.96	111
22	15	*Belgium*	3 479	1.07%	55 409	15.93	86
23		Poland	3 165	0.97%	25 992	8.21	57
24	17	*Finland*	2 867	0.88%	37 645	13.13	73
25	20	*Brazil*	2 860	0.88%	24 543	8.58	55
26		Turkey	2 496	0.77%	23 679	9.49	57
27	14	*Austria*	2 371	0.73%	27 242	11.49	66
28	16	*Denmark*	1 818	0.56%	26 444	14.55	64
29		Portugal	1 527	0.47%	15 513	10.16	50
30		Hungary	1 500	0.46%	16 459	10.97	50

In addition to article counts, we also found out numbers of citations to the articles, average citations per article, and h-indices as defined by Hirsch (2005) for individual countries. In both Table 3 and Table 4, countries are ordered descendingly by the number of publications and the countries from the top 20 CiteSeer countries (see Table 2) are marked with their CiteSeer rank in the second column. When looking at the rankings, we may immediately note that three East Asian countries (mainland China, South Korea, and Taiwan) are under-represented in CiteSeer. Both WoS and Scopus place them in the Top 10 whereas in CiteSeer they are at ranks around 30. The corresponding top-level Internet domains *.cn*, *.kr*, and *.tw* in Table 1 are also relatively lowly ranked, which might suggest that CiteSeer did not crawl these Web regions so extensively as it should have regarding their real scientific productivity

in computer science. Otherwise, we cannot see any striking discrepancies between CiteSeer on one side and WoS and Scopus on the other.

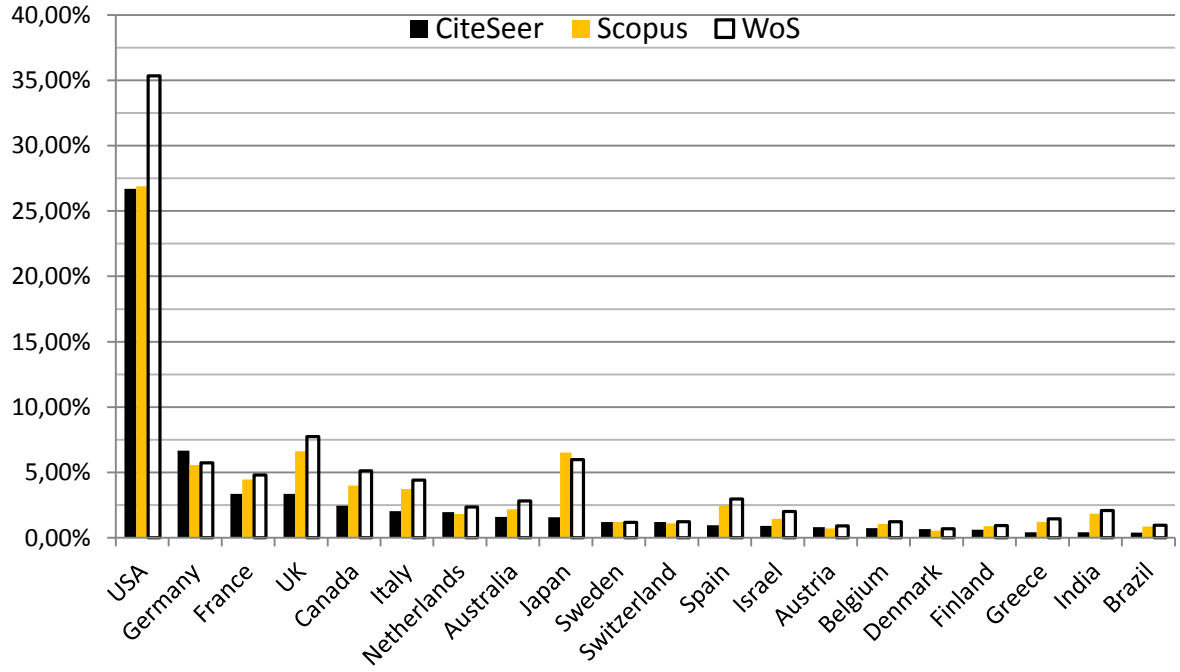

Fig. 3 Publication shares of top 20 CiteSeer countries in Scopus and WoS

Publication shares of the top 20 CiteSeer countries in CiteSeer, WoS, and Scopus are shown in Figure 3. There are no evident outliers or differences either, except perhaps for a greater USA share in WoS. In Figure 4, we show Spearman's rank correlation coefficients between the rankings of CiteSeer and Scopus, CiteSeer and WoS, and Scopus and WoS for the top 10, 20, 30, 40, and 50 CiteSeer countries. All the coefficients are significant at the 0.01 level (two-tailed) except those around 0.65 in the top ten, which are significant at the 0.05 level. Not surprisingly, the rankings from Scopus and WoS are always very highly positively correlated (0.96 – 0.99). But as for CiteSeer, it is also positively correlated with the highest correlation being about 0.86 in the top 50. We may conclude that the ranking by publications from CiteSeer (Table 2) is relevant and quite competitive compared to the rankings from both WoS and Scopus. As there is no simple way of obtaining the total count of citations to all computer science publications published from 1996 to 2005 from the Web sites of WoS and Scopus, which would be necessary to determine the relative citation shares in Tables 3 and 4, we do not present a comparison plot similar to Figure 3 for citations. But we do show, in analogy to Figure 5, how citation-based rankings correlate with each other in Figure 5. As we can see, the rankings of countries based on citations from CiteSeer correlate quite positively

(0.79 – 0.90) with those from Scopus and WoS. All the coefficients in Figure 5 are significant at the 0.01 level (two-tailed).

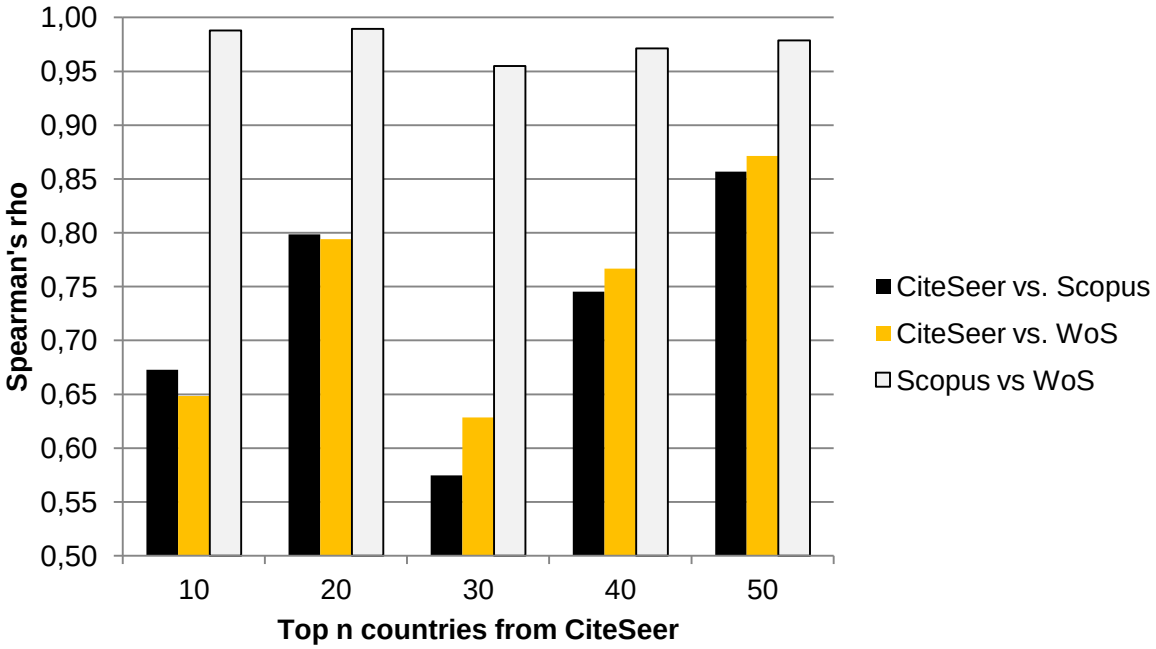

Fig. 4 Correlations of country publication rankings of CiteSeer, Scopus, and WoS

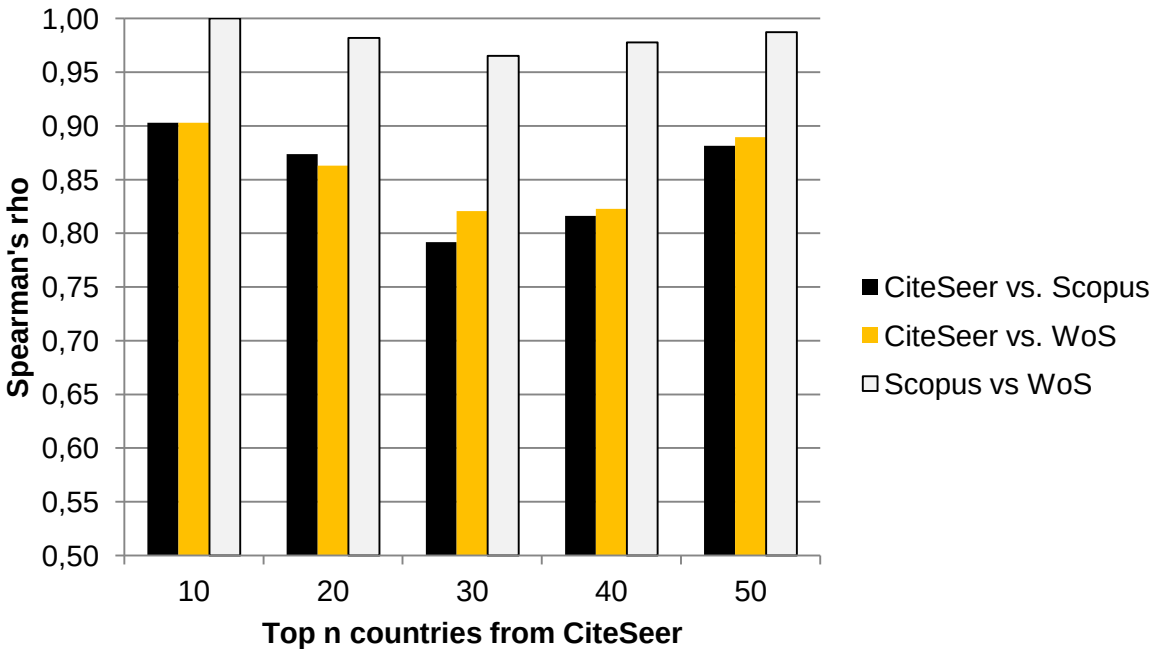

Fig. 5 Correlations of country citation rankings of CiteSeer, Scopus, and WoS

5.4 Citations and recursive indicators

Finally, the resulting directed graph of citations between countries had 243 nodes (countries) and 2 472 edges (citations between them). There were no parallel edges in the graph. Instead, a weight was assigned to each edge denoting from how many parallel edges the edge was created. The sum of weights in the whole graph was about 1.5 million.

In Table 5, we can see the top 80 countries ordered descendingly by their in-degree in the country citation graph. In the first case ("In-degree") the edge weights are all set to one, in the second case ("Citations") they are left as they are. Both rankings place USA, Germany, and the United Kingdom at the top with approximately 48%, 8%, and 6% of all citations, respectively. The rank four in In-degree is tied by Canada and France with the same number of citing countries (74) but, in total, France is cited more often by foreign countries and is positioned ahead of Canada in Citations. A similar behaviour may be observed with several other countries. The country rankings in Table 6 were obtained by applying recursive techniques, but despite their much higher computational costs they do not seem to provide any striking new information, though. We found the five rankings in Tables 5 and 6 to be very highly positively correlated with each other with Spearman's ρ between 0.97 and 1 (all significant at the 0.01 level two-tailed).

Table 5 Top 80 countries by in-degree and citations in CiteSeer

In-degree						Citations					
R.	Country	In	R.	Country	In	R.	Country	Cites	R.	Country	Cites
1	USA	98	41	Slovakia	26	1	USA	728 289	41	Romania	641
2	Germany	82	42	Chile	24	2	Germany	122 389	42	Chile	590
3	United Kingdom	75	43	Jordan	22	3	United Kingdom	89 933	43	Jordan	425
4	Canada	74	44	Argentina	21	4	France	82 632	44	Slovakia	416
5	France	74	45	Bahrain	21	5	Canada	76 148	45	Thailand	409
6	Australia	66	46	South Africa	21	6	Italy	52 570	46	South Africa	328
7	Netherlands	66	47	Bulgaria	20	7	Netherlands	42 252	47	Venezuela	321
8	Switzerland	66	48	Croatia	18	8	Israel	33 701	48	Bahrain	246
9	Italy	64	49	Estonia	18	9	Switzerland	33 185	49	Croatia	222
10	Israel	63	50	Venezuela	18	10	Japan	32 433	50	Estonia	190
11	Japan	62	51	Uruguay	15	11	Australia	27 484	51	Ukraine	183
12	Sweden	62	52	Egypt	14	12	Belgium	21 356	52	Bulgaria	179
13	Spain	58	53	Lebanon	14	13	Sweden	21 211	53	Uruguay	179
14	Austria	55	54	Serbia & Mt.	14	14	Austria	13 975	54	Panama	165
15	Denmark	55	55	Lithuania	13	15	Finland	13 953	55	Lebanon	147
16	Finland	54	56	Latvia	12	16	Spain	13 543	56	Iceland	141
17	Singapore	53	57	Malta	12	17	Denmark	12 744	57	Egypt	138
18	Belgium	52	58	Panama	11	18	India	10 882	58	Iran	119
19	Greece	50	59	Belarus	10	19	Greece	7 304	59	Lithuania	108
20	India	48	60	Fiji	10	20	Singapore	6 165	60	Latvia	103
21	Hong Kong	45	61	Iceland	10	21	Mexico	5 618	61	Fiji	97
22	Portugal	45	62	Bangladesh	9	22	Hong Kong	5 419	62	Serbia & Mt.	81
23	Russia	45	63	Iran	9	23	Portugal	5 398	63	Macau	62
24	Brazil	43	64	Pakistan	8	24	Brazil	5 056	64	Belarus	55
25	Taiwan	42	65	Ukraine	8	25	Taiwan	3 828	65	Pakistan	54
26	China	40	66	Saudi Arabia	7	26	South Korea	3 413	66	Saudi Arabia	50
27	New Zealand	40	67	Moldova	6	27	Russia	3 218	67	Liechtenstein	42
28	Poland	40	68	Macau	5	28	Norway	3 008	68	Kuwait	40
29	Ireland	39	69	Morocco	5	29	New Zealand	2 978	69	Moldova	35
30	Hungary	38	70	Costa Rica	4	30	Ireland	2 952	70	Bangladesh	23
31	Mexico	37	71	Kuwait	4	31	Hungary	2 816	71	Reunion	21
32	Norway	37	72	Vietnam	4	32	China	2 385	72	Vietnam	21
33	Czech Republic	36	73	Armenia	3	33	Poland	1 696	73	Costa Rica	18
34	Cyprus	35	74	Colombia	3	34	Slovenia	1 389	74	Armenia	16
35	South Korea	34	75	Indonesia	3	35	Cyprus	1 162	75	Indonesia	15
36	Turkey	34	76	Tunisia	3	36	Turkey	1 089	76	Monaco	14
37	Slovenia	33	77	Antarctica	2	37	Luxembourg	920	77	Morocco	13
38	Luxembourg	29	78	Congo	2	38	Czech Republic	837	78	Tunisia	12
39	Thailand	27	79	Ethiopia	2	39	Argentina	721	79	Antarctica	10
40	Romania	26	80	Jamaica	2	40	Malta	649	80	Colombia	9

Table 6 Top 80 countries by HITS, PageRank and weighted PageRank in CiteSeer

HITS		PageRank		Weighted PageRank	
R. Country	R. Country	R. Country	R. Country	R. Country	R. Country
1 USA	41 Slovakia	1 USA	41 Thailand	1 USA	41 Romania
2 Germany	42 Chile	2 Canada	42 Chile	2 Germany	42 Chile
3 UK	43 Jordan	3 Germany	43 Jordan	3 UK	43 Malta
4 Canada	44 Argentina	4 UK	44 South Africa	4 France	44 Thailand
5 France	45 Bahrain	5 France	45 Argentina	5 Canada	45 Jordan
6 Netherlands	46 South Africa	6 Israel	46 Bahrain	6 Italy	46 Venezuela
7 Italy	47 Bulgaria	7 Italy	47 Bulgaria	7 Netherlands	47 South Africa
8 Australia	48 Croatia	8 Switzerland	48 Venezuela	8 Israel	48 Bahrain
9 Switzerland	49 Venezuela	9 Netherlands	49 Croatia	9 Japan	49 Croatia
10 Japan	50 Estonia	10 Australia	50 Estonia	10 Switzerland	50 Estonia
11 Sweden	51 Uruguay	11 Japan	51 Uruguay	11 Australia	51 Bulgaria
12 Israel	52 Egypt	12 Sweden	52 Egypt	12 Sweden	52 Panama
13 Spain	53 Lebanon	13 Austria	53 Lebanon	13 Belgium	53 Iceland
14 Finland	54 Serbia & Mt.	14 Spain	54 Serbia & Mt.	14 Austria	54 Lebanon
15 Denmark	55 Lithuania	15 Denmark	55 Lithuania	15 Finland	55 Uruguay
16 Austria	56 Latvia	16 Belgium	56 Latvia	16 Spain	56 Egypt
17 Belgium	57 Malta	17 Finland	57 Malta	17 Denmark	57 Ukraine
18 Singapore	58 Panama	18 Greece	58 Panama	18 India	58 Lithuania
19 Greece	59 Iceland	19 India	59 Iceland	19 Greece	59 Iran
20 India	60 Belarus	20 Singapore	60 Belarus	20 Mexico	60 Fiji
21 Hong Kong	61 Fiji	21 Russia	61 Fiji	21 Singapore	61 Latvia
22 Russia	62 Iran	22 Portugal	62 Iran	22 Hong Kong	62 Serbia & Mt.
23 Portugal	63 Bangladesh	23 Hong Kong	63 Bangladesh	23 Brazil	63 Liechtenstein
24 Taiwan	64 Ukraine	24 Brazil	64 Pakistan	24 Portugal	64 Pakistan
25 Brazil	65 Pakistan	25 New Zealand	65 Ukraine	25 Taiwan	65 Saudi Arabia
26 Ireland	66 Saudi Arabia	26 Taiwan	66 Saudi Arabia	26 New Zealand	66 Belarus
27 Poland	67 Moldova	27 Poland	67 Moldova	27 South Korea	67 Macau
28 China	68 Morocco	28 Ireland	68 Morocco	28 Russia	68 Vietnam
29 New Zealand	69 Macau	29 China	69 Kuwait	29 Norway	69 Kuwait
30 Norway	70 Costa Rica	30 Norway	70 Macau	30 Hungary	70 Moldova
31 Hungary	71 Kuwait	31 Hungary	71 Costa Rica	31 Ireland	71 Monaco
32 Mexico	72 Vietnam	32 Mexico	72 Vietnam	32 China	72 Reunion
33 South Korea	73 Armenia	33 Czech Rep.	73 Armenia	33 Poland	73 Costa Rica
34 Czech Rep.	74 Indonesia	34 Cyprus	74 Indonesia	34 Slovenia	74 Indonesia
35 Turkey	75 Tunisia	35 South Korea	75 Tunisia	35 Cyprus	75 Tunisia
36 Cyprus	76 Colombia	36 Turkey	76 Colombia	36 Turkey	76 Morocco
37 Slovenia	77 Reunion	37 Slovenia	77 Ethiopia	37 Slovakia	77 Armenia
38 Luxembourg	78 Liechtenstein	38 Luxembourg	78 Liechtenstein	38 Czech Rep.	78 Vatican
39 Thailand	79 Neth. Antilles	39 Romania	79 Reunion	39 Luxembourg	79 Bangladesh
40 Romania	80 Ethiopia	40 Slovakia	80 Puerto Rico	40 Argentina	80 Colombia

6. Conclusions and future work

We have presented a thorough study of CiteSeer data with focus on countries and territories with which authors of publications indexed by CiteSeer are affiliated. The main contributions of the study are the following:

- We show from which parts of the Web CiteSeer and CiteSeer^X gathered its documents in terms of shares of top-level Internet domains in article sources.
- We analyze country shares in CiteSeer publications. (Unfortunately, CiteSeer^X does not have the information needed for this kind of analysis.)
- We compare the CiteSeer ranking to country shares of computer science publications from the Web of Science and Scopus to test the reliability of the productivity ranking.
- We submit CiteSeer data to a citation analysis and determine the most influential countries in terms of in-degree, citations, HITS, PageRank, and weighted PageRank.

Based on our analysis, we have obtained the following key results:

- Both CiteSeers collected computer science papers mainly from North American domains, followed by the domains of developed European and Asian countries. The top domains are *.com*, *.de*, *.edu*, *.fr*, *.org*, and *.uk*.
- United States is by far the greatest producer of computer science research papers although West European countries are, relatively at least, very competitive. Germany, France, and the United Kingdom can be named as a few examples.
- CiteSeer rankings of countries by publications and citations are very similar to those generated by the Web of Science or Scopus with a notable difference that CiteSeer apparently underestimates the potential of mainland China, South Korea, and Taiwan.
- Recursive techniques such as PageRank do not provide much new information on the influence of countries compared to simple citation counts. More or less, they confirm that popularity and prestige are close terms in the rankings of countries.

The study presented in this paper is the first of its kind that seeks to determine the most influential countries in computer science by analyzing the free CiteSeer digital library data. It complements the paper by Fiala (2011), which is concerned with individual authors in CiteSeer. From the papers listed in the literature review, the research conducted by Wainer et al. (2009) is closest to ours in that it evaluates the scientific output in computer science of several (thirteen) countries. However, it just examines publications from the Web of Science and Scopus from 2001 to 2005 and is not at all concerned with citations. Even less countries (six) are explored by Guan & Ma (2004) for the period of 1993 - 2002. Both studies, in

accordance with our results, document a clear superiority of the USA over the rest of the world in computer science research. Unfortunately, there seems to be no previous complex computer science study for countries with which we could compare our findings.

Although CiteSeer data are far from complete and precise (in our experience, some 10% of the existing information might be erroneous), we may conclude that CiteSeer is a free digital library of valuable data and may be successfully used in bibliometric studies, possibly along with other well-known bibliographic databases, as we have shown in this paper. Let us underline in this place that the results we present depend solely on the content and quality of CiteSeer data. If other regions of the Web had been crawled, if Asian paper repositories had been preferred by authors (see Section 5.1), or if the information extraction from papers done by CiteSeer had been more precise and complete, the outcomes of our analysis could have been different. Let us hope in this respect that CiteSeer^{X} will acquire data in a more standardized and transparent way and that it will enrich its metadata with the information on addresses and affiliations as well. Our future work on CiteSeer will concentrate on the citation analysis of institutions and on other reliability measures of CiteSeer data as well as on exploring further differences between the data in CiteSeer and CiteSeer^{X}.

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. Many thanks are due to the anonymous reviewers for their useful comments.

References

Albarrán, P., Crespo, J. A., Ortuño, I., & Ruiz-Castillo, J. (2010). A comparison of the scientific performance of the U.S. and the European Union at the turn of the 21st century. *Scientometrics*, 85(1), 329-344.

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.

Bakri, A., & Willett, P. (2011). Computer science research in Malaysia: A bibliometric analysis. *Aslib Proceedings: New Information Perspectives*, 63(2-3), 321-335.

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.

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics*, 83(3), 809-824.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World Wide Web Conference*, Brisbane, Australia, 107-117.

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

CiteSeer. http://citeseer.ist.psu.edu.

CiteSeer^{X}. http://citeseerx.ist.psu.edu.

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

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, 76(1), 135-158.

Fiala, D. (2011). Mining citation information from CiteSeer data. *Scientometrics*, 86(3), 553-562.

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.

Franceschet, M. (2010b). The role of conference publications in CS. *Communications of the ACM*, 53(12), 129-132.

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.

Guan, J., Ma, N. (2004). A comparative study of research performance in computer science. *Scientometrics*, 61(3), 339-359.

Gupta, B. M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999-2008. *Scientometrics*, 86(2), 261-283.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.

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.

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

Ma, R., Ni, C., & Qiu J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, 76(2), 245–260.
Repository66. http://maps.repository66.org.
Scopus. http://www.scopus.com.
Wainer, J., Xavier, E. C., & Bezerra, F. (2009). Scientific production in computer science: A comparative study of Brazil and other countries. *Scientometrics*, 81(2), 535-547.
Web of Science. http://apps.isiknowledge.com.
Wu, C.-L., & Koh, J.-L. (2010). Hierarchical topic-based communities construction for authors in a literature database. *Lecture Notes in Computer Science*, 6097, 514-524.
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. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM'07)*, Omaha, Nebraska, USA, 745-750.