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**

EXTRACTING INFORMATION FROM CITESEER'S TEXTUAL DATA

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

ABSTRACT

This article deals with CiteSeer, a free online digital library and search engine of mainly computer science research papers. First, it discusses CiteSeer's features and structure and then it presents what useful information on publications and author collaborations can be extracted from its textual data. We show the basic properties of both the publication citation and author citation graph. Moreover, several parameters based on the structure of the collaboration graph of authors are discussed and their main statistical properties are shown.

Keywords: *CiteSeer, Publications, Citations, Researchers, Collaboration*

1. INTRODUCTION

CiteSeer [1] is a digital library and search engine that covers primarily computer science research literature. It acquires documents that are freely available on the Web and that it considers as computer science research papers. The papers are automatically gathered by crawling the Web; they are converted from mostly PDF and PostScript files into plain text, parsed and provided with metadata to finally produce a large corpus of tagged textual data that can be further analyzed. CiteSeer was first introduced in [2] and, in addition to searching computer science literature, its website provided services such as reference linking of articles and finding highly cited papers or authors. In 2010 CiteSeer was replaced with CiteSeer^X, which is, however, still in a beta version at the time of writing this article (March 2013). The present paper complements our study [3], in which we showed how citation information could be mined from CiteSeer data and used to detect influential computer scientists using several evaluation techniques. In this article we concentrate on the features and structure of CiteSeer data and analyze various parameters underlying the author rankings in [3] that can be calculated based on the investigation of the author collaboration graph extracted from CiteSeer's textual data.

Prior to [3], there was a preceding study [4], in which we profoundly defined the parameters discussed in Section 2, and later on we published another study [5], in which the parameters were further enhanced. Both analyses used the parameters to fine-tune a ranking algorithm for the evaluation of researchers based on PageRank [6] and HITS [7] applied to citation graphs of authors. Although [4] as well as [5] took advantage of data sources other than CiteSeer, the latter's data were used in various experiments elsewhere, e.g. in [8], [9], [10], [11], or [12].

2. DATA AND METHODS

In our analysis, we used a CiteSeer data file from December 2005, which represented the most recent freely available data before the transformation of CiteSeer into CiteSeer^X, in which the data format changed and the data were no longer available for download as a single file. We downloaded a 2 GB zipped archive which we unpacked and thus obtained 72 text files with almost 717 000 records similar to that in figure 1. Each record contains metadata on an article's title, authors, usually also cited references and other information. These records can then be processed to create citation graphs of publications and authors and collaboration graphs of authors. The results of the analysis of these networks in terms of various parameters are shown in Section 3.

CiteSeer data are much larger than DBLP data analyzed in [4]. There are more than 1.8 million citations between 717 thousand publications. Some publications (about 333 thousand) are entirely isolated – neither do they cite, nor are they cited by other publications. On the other hand, roughly 149 thousand publications cite and are cited at the same time. Of course, there are

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

```
<record>
<header>
<identifier>oai:CiteSeerPSU:2</identifier>
<datestamp>1997-11-01</datestamp>
<setSpec>CiteSeerPSUset</setSpec>
</header>
<metadata>
<oai_citeseer:oai_citeseer ...>
  <dc:title>The Graham Scan Triangulates Simple Polygons</dc:title>
  <oai_citeseer:author name="Xianshu Kong"></oai_citeseer:author>
  <oai_citeseer:author name="Hazel Everett"></oai_citeseer:author>
  <oai_citeseer:author name="Godfried Toussaint"> </oai_citeseer:author>
  <dc:subject>Xianshu Kong,Hazel Everett,Godfried Toussaint The Graham Scan…</dc:subject>
  <dc:description>The Graham scan is a fundamental backtracking technique...</dc:description>
  <dc:contributor>The Pennsylvania State University CiteSeer Archives</dc:contributor>
  <dc:publisher>unknown</dc:publisher>
  <dc:date>1997-11-01</dc:date>
  <oai_citeseer:pubyear>1991</oai_citeseer:pubyear>
  <dc:format>ps</dc:format>
  <dc:identifier>http://citeseer.ist.psu.edu/2.html</dc:identifier>
  <dc:source>http://www-cgrl.cs.mcgill.ca/~godfried/publications/tri.scan.ps.gz</dc:source>
  <dc:language>en</dc:language>
  <oai_citeseer:relation type="References">
    <oai_citeseer:uri>oai:CiteSeerPSU:97473</oai_citeseer:uri>
  </oai_citeseer:relation>
  <oai_citeseer:relation type="References">
    <oai_citeseer:uri>oai:CiteSeerPSU:154288</oai_citeseer:uri>
  </oai_citeseer:relation>
  <dc:rights>unrestricted</dc:rights>
</oai_citeseer:oai_citeseer>
</metadata>
</record>
```

Figure 1: A Sample CiteSeer Record

publications that cite but are not cited and vice versa. These and other relationships can be seen in figure 2.

CiteSeer data are not clean, the causes and implications of which will be discussed later on in more detail. There are several phases in which errors can be introduced in the whole automated process of the creation of this digital library. The first phase is the deposition of the source documents themselves that can be freely made by anyone and uploaded to a website already with errors. The second phase prone to errors is the plain text conversion and the third one is the parsing.

Figure 3, in turn, shows the cumulative distribution of in- and out-degrees in the directed graph of citations between publications on the logarithmic scale. Obviously, the bars in the first bin (1+) counting publications with in- or out-degree equal to one or more correspond to the small ovals in the middle of the diagram in figure 2. The out-degree is prevailing in the first three bins, then it begins declining with the value of twenty or more until there are no publications citing 300 or more articles. Regarding the in-degree, there are still some papers that receive 1000 or more citations. Another interesting aspect is the number of co-authors in a publication (figure not shown). The first three most common co-author numbers are two, one, and three. There are also over 50 thousand publications for which CiteSeer was unable to identify authors.

As for the methods we employed, for each author pair for which there is an edge in the author citation graph, we define the following coefficients that can be potentially used to enrich the citation graph with information from the collaboration graph by adding different weights to the edges and to help rank researchers more fairly by means of iterative PageRank-based algorithms (for details on these parameters and algorithms, see [4] and [3]):

- *c* as the number of common publications by authors 1 and 2
- *f* as the number of publications by author 1 plus the number of publications by author 2
- *h* as the number of all co-authors in all publications by author 1 plus the number of all co-authors in all publications by author 2

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

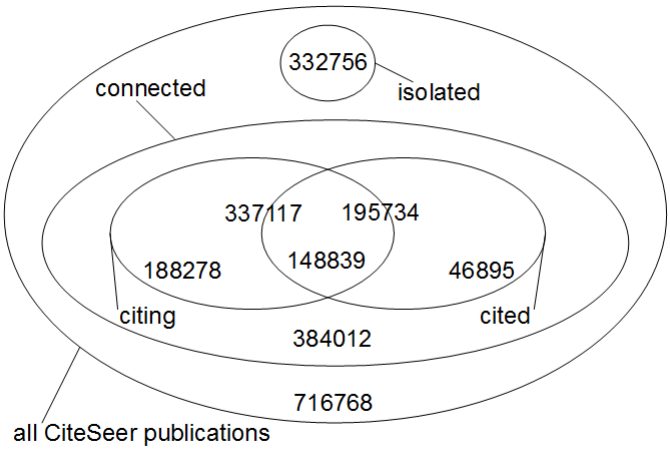

Figure 2: Numbers of Citing and Cited CiteSeer Publications

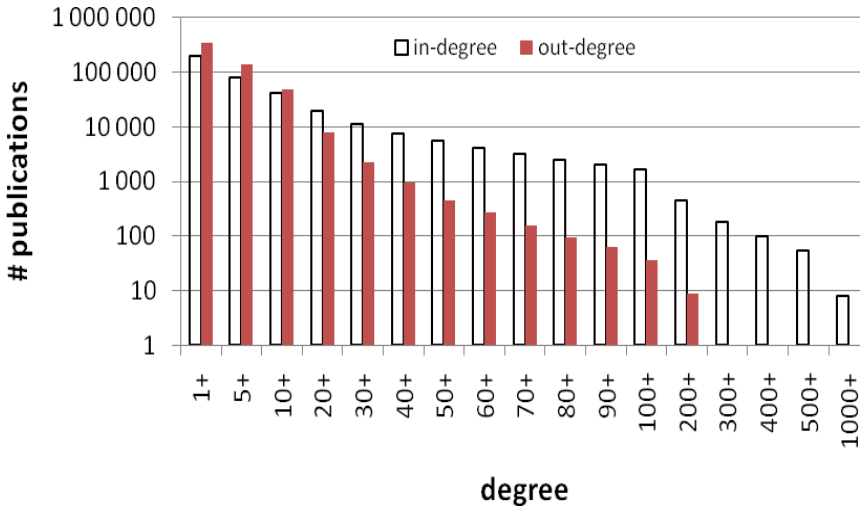

Figure 3: Cumulative Distribution of In- and Out-degrees in Publication Citation Graph

- *hd* as the number of all distinct co-authors in all publications by author 1 plus the number of all distinct co-authors in all publications by author 2
- *g* as the number of publications by author 1 where author 1 is not the only author plus the number of publications by author 2 where author 2 is not the only author
- *t* as the number of co-authors in common publications by authors 1 and 2
- *td* as the number of distinct co-authors in common publications by authors 1 and 2

3. RESULTS AND DISCUSSION

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). As with publications, some authors (171 thousand) are isolated from the rest while other authors cite or are cited by others (111 thousand in figure 4). The relation of those

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

who cite and are not cited to those who are cited but do not cite is approximately 3 : 1.

In the section on methods, we have defined a couple of parameters that can be retrieved from the data on publications and their authors (collaboration or co-authorship graph). Now, let us have a look at what values of these parameters and how often can be found. We can see frequency histograms of parameters *c*, *f*, *g*, *h*, *hd*, *t*, and *td* in figure 5, figure 6, figure 7, and figure 8. Not all collaborating authors also have an edge in the citation graph. We present parameter counts only for those author pairs that have an edge in the author citation graph (80 247 author couples in the 1+ bin). The histograms are cumulative and the chart bars decrease. For instance, the number of author pairs (ordered because each pair represents an edge in the directed graph of citations between authors) having three common publications or more is about 40 000 and the number of those collaborating exactly twice is approximately 15 000 (figure 5). Similarly, the number of author pairs that have the number of all publications (*f*) and that of all non-solo publications (*g*) 100 or more is a little more than 30 000 (figure 6).

While parameters *f* and *g* are relatively tightly bound, the histograms of *h* (all co-authors) and *hd* (all distinct co-authors), and *t* (co-authors in common publications) and *td* (distinct co-authors in common publications) differ more. For example, there is a big gap in the 100+ bin: the difference is almost 50 000 author pairs in favour of *h* (figure 7). Also, there are still well over 10 000 author pairs that have forty or more co-authors in common publications while there are hardly any that have a

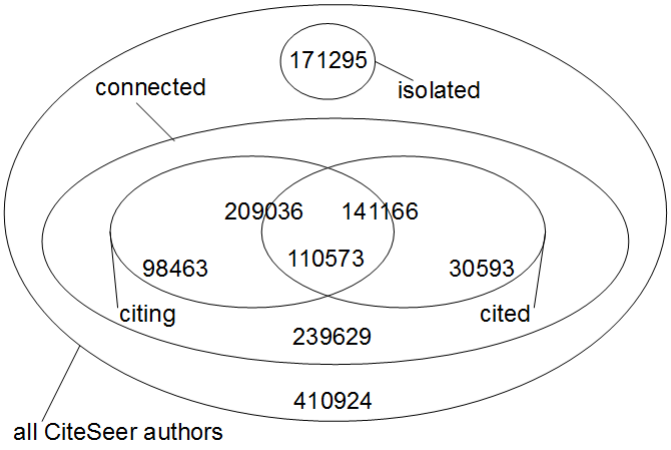

Figure 4: Numbers of Citing and Cited CiteSeer Authors

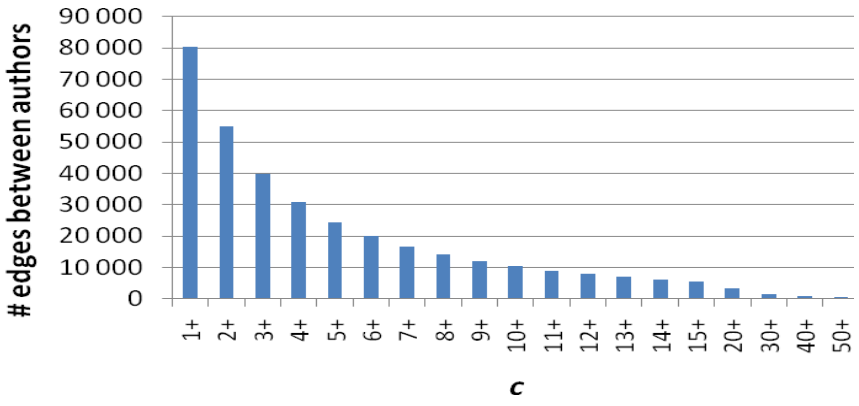

Figure 5: Cumulative Distribution of Values of Parameter c in Graph G

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

similar number of distinct co-authors in common publications (figure 8). In this context, let us note that there are inherent errors in the CiteSeer metadata resulting from a wrong parsing of the plain text generated from the original PDF or PostScript files. In this respect, any other author name disambiguation than a simple textual comparison would make little sense because it would require the data to be clean. In spite of this, we showed in [3] and [13] that CiteSeer data could be used in such analyses.

Table 1 complements figure 5, figure 6, figure 7, and figure 8 and presents a basic statistical overview of the distribution of the above mentioned parameters in the author citation graph for edges between authors having one common publication at least (more than 80 000 edges). Let us have a look at some of the values. One is the most frequent number of collaborations between cooperating authors (see mode of *c*) whereas the average number is just above five. The maximum number of common publications is 317. By examining the data, though, we find out that this most collaborating pair of authors is "Senior Member" and "Student Member". The first real author names are Dieter Fox and Wolfram Burgard with *c* equal to 191. The median of the number of all publications of two collaborating authors (*f*) is 86, the average count of non-solo publications (*g*) is equal to 126, The median of the number of all co-authors (*h*), all distinct co-authors (*hd*), co-authors in common publications (*t*), and distinct co-authors in common publications (*td*) is 287, 60, 12, and 5, respectively. The series is decreasing as the count criterion is getting stricter. A more thorough analysis of the collaboration patterns mined from CiteSeer is needed hence leaving enough space for some future research that will focus on the co-authorship network in a detailed way to detect, among others, the most intense collaborations in the field of computer science and related disciplines such as mathematics.

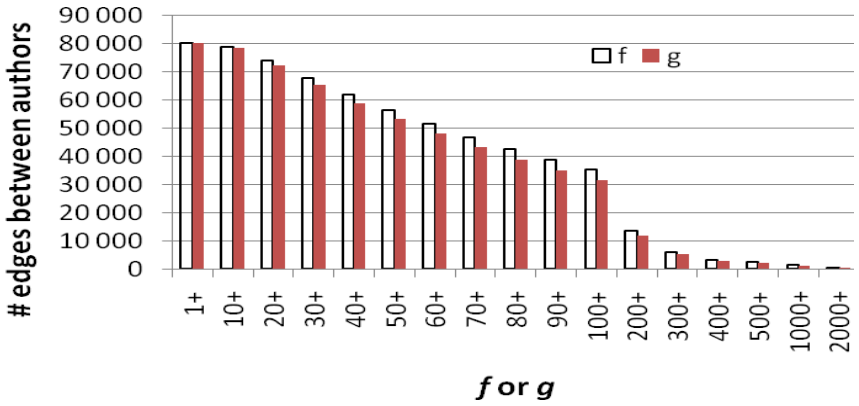

Figure 6: Cumulative Distribution of Values of Parameters f and g in Graph G

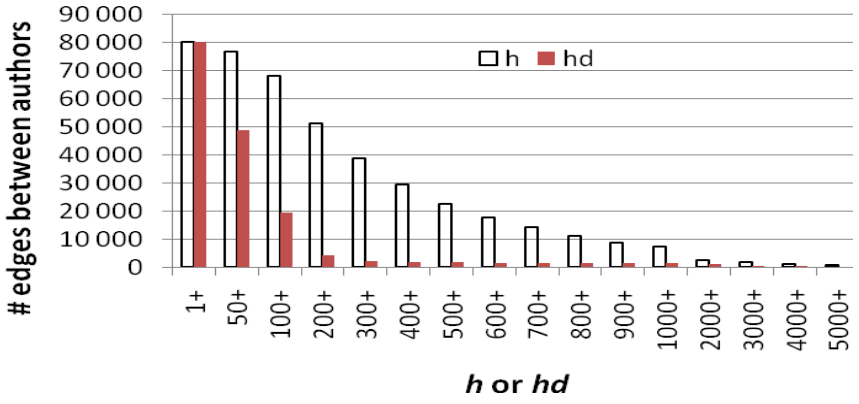

Figure 7: Cumulative Distribution of Values of Parameters h and hd in Graph G

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

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented CiteSeer, a free online digital library and search engine devoted to computer science-related research literature. In 2010, it was transformed into CiteSeer^X, which made it more difficult to automatically process its large corpus of textual data on hundreds of thousands of computer science research papers. Therefore, we analyzed the last freely available data file from the "old" CiteSeer containing almost 717 000 records with bibliographic metadata in tagged plain text. We described its properties and created citation networks of papers and authors and a collaboration network of authors. The main contributions of our work are the following:

- We showed the structure and the degree distribution of the paper citation graph.
- We showed the structure and indicated the degree distribution of the author citation network generated from the paper citation graph.
- We highlighted the basic statistical properties of the parameters defined in [3] and [4] that are based on the author collaboration graph and help assign weights to the edges in the author citation graph with the aim of rank influential researchers more fairly.

Our future work will include a deeper analysis of the collaboration patterns mined from the CiteSeer co-authorship network as well as exploring the correlation of author rankings based on the different parameters we discussed. We will also be concerned with the convergence rate of the new ranking methods that are iterative by definition. To conclude, we think that the text corpus of CiteSeer data on computer science research articles is still relatively little explored and that it is definitively worth studying in the future.

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.

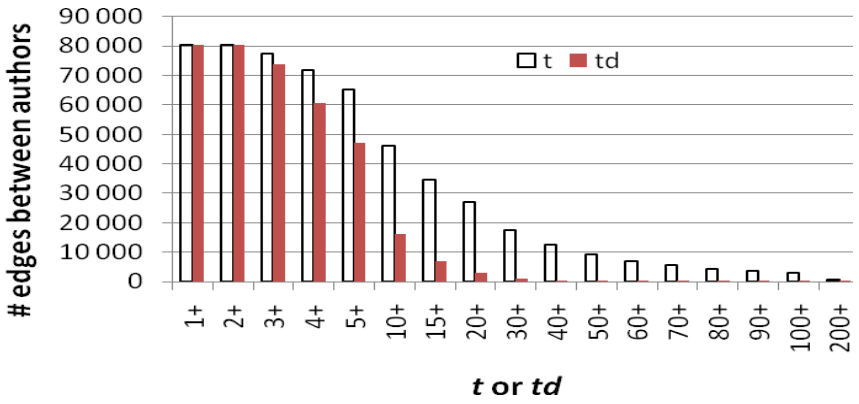

Figure 8: Cumulative Distribution of Values of Parameters t and td in Graph G

Table 1: Basic Statistics of Weight Parameters for Edges in G with Non-zero c

	c	*f*	*g*	*h*	*hd*	*t*	*td*
min	1	4	2	6	4	2	2
max	317	2 438	2 435	10 504	4 757	1 975	729
avg	5.06	136.22	126.45	498.07	117.72	24.64	7.12
std. deviation	8.13	188.67	185.49	782.45	316.71	52.05	7.52
median	2	86	77	287	60	12	5
mode	1	30	20	145	41	4	4

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

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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