Modeling the Author Bias Between Two On-line Computer Science Citation Databases

Vaclav Petricek[1], Ingemar J. Cox[1], Hui Han[2], Isaac G. Councill[3], and C. Lee Giles[3]

[1]University College London, WC1E 6BT, Gower Street, London, United Kingdom, <{v.petricek| i.cox}@cs.ucl.ac.uk>

[2]Yahoo! Inc., 701 First Avenue, Sunnyvale, CA, 94089, <huihan@yahoo-inc.com>

[3]IST Pennsylvania State University, University Park, PA 16802, USA, <{giles@ist.| igc2@}psu.edu>

ABSTRACT

We examine the difference and similarities between two on-line computer science citation databases DBLP and CiteSeer. The database entries in DBLP are inserted manually while the CiteSeer entries are obtained autonomously. We show that the CiteSeer database contains considerably fewer single author papers. This bias can be modeled by an exponential process with intuitive explanation. The model permits us to predict that the DBLP database covers approximately 30% of the entire literature of Computer Science.

Categories and Subject Descriptors

H.1.m [**Information Systems**]: Models and Principles

General Terms

Theory, Measurement

Keywords

Acquisition bias, Bibliometrics, CiteSeer, DBLP

1. Introduction

Several public databases of research papers became available due to the advent of the Web [1, 7, 3, 2, 4, 5]. These databases collect papers in different scientific disciplines, index them and annotate them with additional metadata. The coverage and acquisition methods of these databases greatly vary. As author and document citation rates are increasingly being used to quantify the scientific impact of scientists, publications, journals and funding agencies, it is important to understand the limitations and biases introduced by different acquisition methods.

2. Datasets

Within the computer science community, there are two popular public citation databases. These are DBLP and CiteSeer. CiteSeer was created by Steve Lawrence and C. Lee Giles in 1997 [7]. It currently contains over 716,797 documents. DBLP was operated by Micheal Ley since 1994 [8]. It currently contains over 550,000 computer science references from around 368,000 authors. The two databases are constructed in very different ways. In DBLP, each entry is manually inserted by a group of volunteers and occasionally hired students. The entries are obtained from conference proceeding and journals. In contrast, each entry in CiteSeer is automatically entered from an analysis of documents found on the Web. There are advantages and disadvantages to both methods.

To appear at *WWW2005*, May 10–14, 2005, Chiba, Japan.

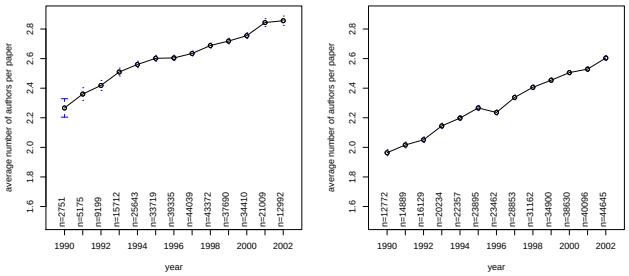

Figure 1: Average number of authors per paper for the years 1990 to 2002 in CiteSeer(left) and DBLP(right)

In our analysis we focus on the difference in data acquisition and the biases that this difference introduces.

3. Bias in number of authors

3.1 Average number of authors

We examined the average number of authors per paper for publications between 1990 and 2002, see Figure 1. In both datasets, the average is seen to be rising. It is uncertain what is causing this rise in multi-authorship. Possible explanations include (i) funding agencies preference to fund collaborative research and/or (ii) collaboration has become easier with the increasing use of email and the Web. However, we observe that the CiteSeer database contains a higher number of multi-author papers.

3.2 Bias in number of authors

Figure 2 examines the relative frequency of n-authored papers in the two datasets. Note that the data is on a log-log scale. It is clear that CiteSeer has far fewer single-authored papers. In fact, CiteSeer has relatively fewer papers published by one to three authors. This is emphasized in Figure 3 in which we plot the ratio of the frequency of n-authored papers for CiteSeer and DBLP. Here we see the frequency of single-authored papers in CiteSeer is only 77% of that ocurring in DBLP. As the number of authors increases, the ratio decreases since CiteSeer has a higher frequency of n-authored papers for $n > 3$. For high n, the ratio is somewhat random, reflecting the scarcity of data in this region. We therefore limit our analysis to numbers of authors where there are at least 100 papers in each dataset. This restricts the number of authors to less than 17.

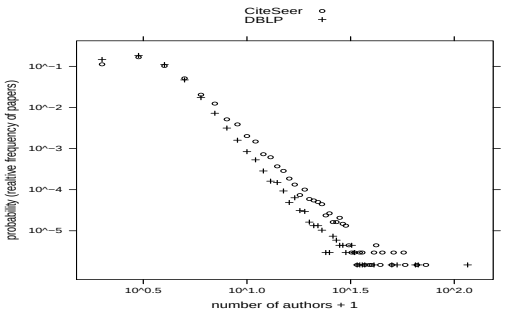

Figure 2: Probability histogram of number of authors. (double logarithmic scale.)

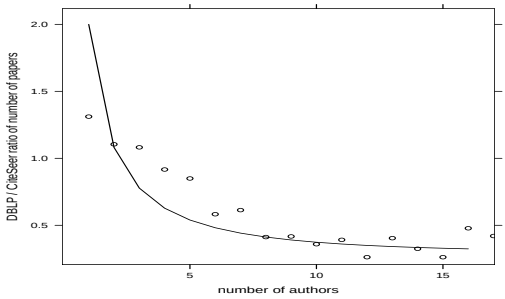

Figure 3: Fit of model (3) for values of $\alpha = 0.3$ and $\beta = \delta\gamma = 0.15$ for numbers of authors where there are at least 100 documents in both datasets in total.

As we see in Figure 2 the number of authors follows a power law corresponding to a line with slope approximately -0.23 for DBLP and -0.24 for CiteSeer. There is an obvious cut-off from the power law for papers with low number of authors. For CiteSeer, we hypothesize that (i) papers with more authors are more likely to be submitted to CiteSeer and (ii) papers with more authors appear on more homepages and are therefore more likely to be found by the crawler. These ideas are modeled in Section 3.3.

However none of these factors is relevant to DBLP, which also exhibits a similar drop off in single-authored papers. Other explanations may be that (i) single author papers are less likely to be finished and published, (ii) funding agencies encourage collaborative and therefore multi-authored research and (iii) it is an effect of limited number of scientists in the world [6].

3.3 Acquisition Models

To explain the apparent bias of CiteSeer towards papers with larger numbers of authors, we develop two possible models for the acquisition of papers within CiteSeer. We also provide a simple acquisition model for DBLP.

The first CiteSeer model is based on authors submitting their papers directly to the database. The second CiteSeer model assumes that the papers are obtained by a crawl of the Web. We find that in fact, both models are equivalent and therefore describe only the crawler model below.

To begin, let citeseer(i) be the number of papers in CiteSeer with i authors, dblp(i) the number of papers in DBLP with i authors and all(i) the number of papers with i authors published in all Computer Science. For DBLP, we assume a simple paper acquisition model such that there is a probability α that a paper is included in DBLP and that this probability is independent of the number of authors. For CiteSeer we assume that the acquisition method introduces a bias such that the probability, p(i) that a paper is included in CiteSeer is a function of number of authors, i, of that paper. That is,

$$\text{dblp}(i) = \alpha \cdot \text{all}(i) \tag{1}$$

$$\text{citeseer}(i) = \text{p}(i) \cdot \text{all}(i) \qquad = p(i) \cdot \frac{\text{dblp}(i)}{\alpha} \tag{2}$$

Let $\delta \in (0, 1)$ be the probability that an author puts a paper on a web site (homepage for example). Then the average number of copies of an i-authored paper on the Web is $i \cdot \delta$. Let us further assume that the crawler finds each available on-line copy with a probability γ. Then the probability that there will be exactly c copies of an an i-authored paper published on-line is $\binom{i}{c}\delta^c(1-\delta)^{i-c}$.

Given that the probability of finding a document with c copies online by a web crawl is $1-(1-\gamma)^c$ then the probability that CiteSeer will crawl an i-authored document, p(i) is p(i) $= 1-(1-\gamma\delta)^i$, where $(1-\gamma\delta)^i$ is the probability that no copy of an i-author paper is found by CiteSeer.

Substituting into Equation (2), we have

$$\text{r}(i) = \frac{\text{dblp}(i)}{\text{citeseer}(i)} = \frac{\alpha}{(1-(1-\gamma\delta)^i)} \tag{3}$$

We plot $r(i)$ for numbers of authors i where we have at least 100 papers available in Figure 3. We see the fit is not perfect suggesting that this is not the only mechanism involved.

The value of α is 0.30 - the value to which the data points are converging for high numbers of authors. If our model is correct, this suggests that the DBLP database covers approximately 30% of the entire Computer Science literature.

4. Summary

This paper compared two popular online science citation databases, DBLP and CiteSeer, which have very different methods of data acquisition. We showed that autonomous acquisition by web crawling, (CiteSeer), introduces a significant bias against papers with low number of authors (less than 4). We attempted to model this bias by constructing two probabilistic models for paper acquisition in CiteSeer. The model assumes that the probability of crawling a paper is proportional to the number of online copies of the paper and that the number of online copies is again proportional to the number of authors. This permits us to estimate that the coverage of DBLP is approximately 30% of the entire Computer Science literature.

5. REFERENCES

[1] Arxiv e-print archive, http://arxiv.org/.
[2] Compuscience database, http://www.zblmath.fiz-karlsruhe.de/comp/quick.htm.
[3] Corr, http://xxx.lanl.gov/archive/cs/.
[4] Cs bibtex database, http://liinwww.ira.uka.de/bibliography/.
[5] Sciencedirect digital library, http://www.sciencedirect.com, 2003.
[6] J. Laherrre and D. Sornette. Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales. *The European Physical Journal B - Condensed Matter*, 2(4):525–539, 1998.
[7] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.
[8] M. Ley. Dblp: A www bibliography on databases and logic programming, 1997.