Modelling the Author Bias Between Two Online Computer Science Citation Databases

Vaclav Petricek Ingemar J. Cox
University College London
{v.petricek|i.cox}@cs.ucl.ac.uk

Hui Han
Yahoo! Inc.
huihan@yahoo-inc.com

Isaac G. Councill C. Lee Giles
Pennsylvania State University
{giles@ist.|igc2@}psu.edu

The 14th International World Wide Web Conference
May 10 - 14, 2005

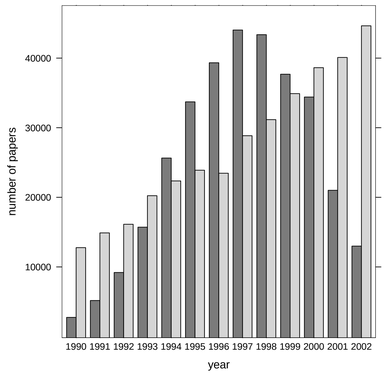
Objective

Compare properties of DBLP and CiteSeer – two online citation databases with different acquisition methods. The database entries in DBLP are inserted manually while the CiteSeer entries are obtained autonomously. There are advantages and disadvantages to both methods. In our analysis we focus on the biases that this difference introduces.

Datasets

CiteSeer was created by Steve Lawrence and C. Lee Giles in 1997 [2]. It currently contains over 716,797 documents. Each entry in CiteSeer is automatically entered from an analysis of documents found on the Web. Citation data is obtained by automatic parsing of documents.

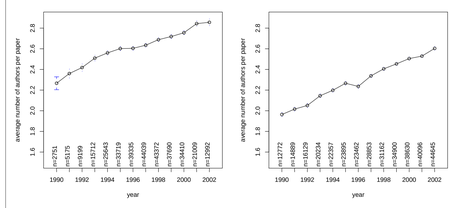
DBLP was operated by Micheal Ley since 1994 [4]. It currently contains over 550,000 computer science references by around 368,000 authors. In DBLP, each entry is manually inserted by a group of volunteers and occasionally hired students. The entries are obtained from conference proceeding and journals. Citation data has also been entered manually by a group of students.



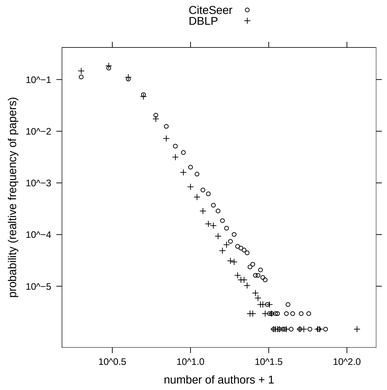
Publication year distribution for papers in CiteSeer (dark) and DBLP (light) databases

Author Bias

Average number of authors per paper in CiteSeer(left) and DBLP(right)



In both databases, the average number of authors is seen to be rising. Possible explanations include (i) funding agencies preference to fund collaborative research (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.



Probability histogram of number of authors. (double logarithmic scale.)

The relative frequency of n-authored papers in the two datasets. Note that the data is on a log-log scale. In fact, CiteSeer has relatively fewer papers published by one to three authors. 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 $n > 3$.
The distribution of number of authors follows a power law with $alpha = -0.23$ for DBLP and $\alpha = -0.24$ for CiteSeer.

Acquisition models

Definitions

all(i) is the number of papers with i authors published in all Computer Science.

dblp(i) is the number of papers in DBLP with i authors

citeseer(i) is the number of papers in CiteSeer with i authors

$\delta \in (0, 1)$ is the probability that an author puts a paper on a web site (homepage for example).

$\beta \in (0, 1)$ is the probability that an author submits a paper to CiteSeer.

$\gamma \in (0, 1)$ is the probability that CiteSeer crawler finds a paper that has one copy somewhere online.

DBLP model 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.

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

CiteSeer submission model 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, The first CiteSeer model is based on authors independently submitting their papers directly to the database with probability β.

$$\text{citeseer}(i) = (1 - (1 - \beta)^i) \cdot \text{all}(i) \tag{2}$$

Where $(1 - \beta)^i$ is the probability that none of the authors submits the particular paper to CiteSeer.

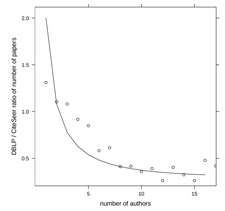
CiteSeer crawler model The second CiteSeer model assumes that the papers are obtained by a crawl of the Web.

$$\text{citeseer}(i) = (1 - (1 - \gamma\delta)^i) \cdot \text{all}(i) \tag{3}$$

We find that in fact, both models result in equivalent bias.

Results

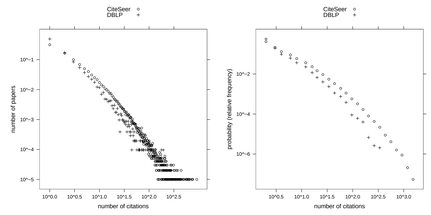
$$\frac{\text{dblp}(i)}{\text{citeseer}(i)} = \frac{\alpha}{(1 - (1 - \gamma\delta)^i)} \tag{4}$$



Fit of model (4) for $\alpha = 0.3$ and $\delta\gamma = 0.15$

For high numbers of authors the data points are converging to $\alpha = 0.3$. If our model is correct, this would suggest that the DBLP database covers approximately 30% of the entire CS literature.

Citation distributions for Computer Science



Citation distributions in CiteSeer and DBLP (left atomic and right exponentially binned).

In agreement with [1] we see that DBLP contains more low cited papers than CiteSeer.

number of citations	slope (α parameter of the power-law) Lehmann [3]	CiteSeer	DBLP
< 50	-1.29	-1.504	-1.876
> 50	-2.32	-3.074	-3.509

Results of linear interpolation of exponentially binned probabilities.

Our parameters are bigger in absolute value than [3] obtained for Physics, suggesting that **highly cited papers acquire a larger share of citations in Computer Science than in Physics**. There is also a significant difference between CiteSeer and DBLP.

Summary

We 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 **DBLP covers approximately 30% of the entire Computer Science literature**. Apart from that we show that citation distribution is more uneven in CS than in Physics.

Acknowledgments

The authors would like to thank Michael Ley for helping with understanding the DBLP. The work of the first author has been partially supported by Vize'97 Foundation, Mobility Fund of Charles University, and Bernard Bolzano Foundation.

References

[1] S. Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001.

[2] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[3] S. Lehmann, B. Lautrup, and A. D. Jackson. Citation networks in high energy physics. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 68(2):026113, 2003.

[4] M. Ley. Dblp: A www bibliography on databases and logic programming, 1997.

[5] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.