Estimating Article Influence Scores for Open Access Journals

Bree Norlander, Peter Li, Jevin West
norlab@uw.edu, pzli97@uw.edu, jevinw@uw.edu

Problem

FlourishOA is a one-stop shop for authors, publishers, funders, librarians, and policy makers to find high-quality, cost-effective Open Access (OA) journals. It is the only web application that compares influence data with pricing data in an effort to distinguish legitimate OA journals from "predatory" OA journals. Specifically, FlourishOA provides Article Processing Charge (APC)[1] and Article Influence (AI)[2] data via visualizations and API. AI scores are retrieved from InCites Journal Citations Reports (JCR). However, the FlourishOA database contains thousands of journals not indexed in JCR. We would like to include more journals in our comparison tool. Can we predict AI scores for journals not represented in JCR?

1. Article Processing Charge: the cost to publish in the given journal
2. Article Influence Score: a measure of influence per article "calculated as a journal's Eigenfactor Score divided by the number of articles in that journal, normalized so that the average article in the Journal Citation Reports has an Article Influence Score of 1" (West, Bergstrom, & Bergstrom, 2010, p.239)

Solution

1. Fuzzy match[3] FlourishOA journals to Microsoft Academic Graph (MAG) Journals to retrieve MAG identification number.

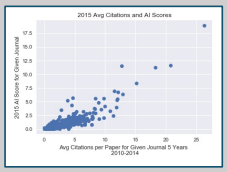
3.Strip out common mismatches, use Levenshtein threshold to output potential matches, & manually filter for correct titles.

2. Query MAG API to count all incoming citations to a specified journal in 2010, 2011, 2012, 2013, and 2014. Also count the number of articles in specified journal for same years.

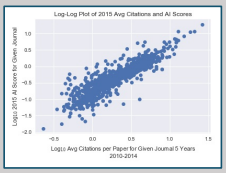
These first two steps formed the bulk of the work. Because the MAG Database via Microsoft Cognitive Services Academic Knowledge API[4] contains such a large amount of data in several entity attributes, the API calls are complicated and can run for several days. Our API call tests would often bump up against the rate limit and we had to adjust the script to account for these potential errors.

4. Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. *In Proceedings of the 24th International Conference on World Wide Web(WWW '15 Companion). ACM, New York, NY, USA, 243-246.* http://dx.doi.org/10.1145/2740908.2742839

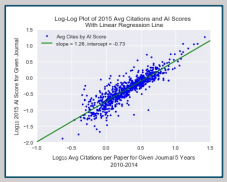
3. Calculate average citations per article per journal over the five year span. Compare with known AI scores.



4. Transform the data via a log-log plot to better account for the density of data in the lower left quadrant of the graph.



5. Regress the log-transformed known $\log_{10}$ AI Scores on $\log_{10}$ Avg Citation Counts to form a model from which to predict unknown AIs.



6. Transform the predicted scores back, calculate RMSE (0.498) of predictions versus known AI values, and plot.

Alfred P. Sloan FOUNDATION

Acknowledgement:
This work was funded by a grant from the Alfred P. Sloan Foundation.

More about FlourishOA:
http://www.flourishoa.com
http://twitter.com/Flourish_OA

Information School
UNIVERSITY of WASHINGTON