POSTER

Methods for Bibliometric Institutional Profiles for the Practitioner

Christina K. Pikas
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, m/s 5-17, Laurel, MD, USA, 20723-6099
Christina.Pikas@jhuapl.edu, ORCID: 0000-0002-4455-8036

INTRODUCTION

Recently, professional bibliometricians have suggested categories of stakeholders for bibliometric analyses: professional bibliometricians who develop the indicators and act as paid consultants; producers of data sets used for analyses; research managers who commission bibliometric analyses; and the researchers or authors whose work is measured (Leydesdorff, Wouters, & Bornmann, 2016). I argue that information professionals and librarians who work with researchers form an important and underappreciated stakeholder group that is neither "citizen" nor professional bibliometrician. A review of the library professional literature and conference proceedings such as those from the recent Association of College and Research Libraries meeting shows that not only do many librarians have a sophisticated and nuanced view of bibliometric indicators, but they serve as *local* (and salaried!) bibliometricians for their research institutions.

Accordingly, this poster furthers the discussion with practitioners who need to profile a research organization. It presents a case study of a research lab, The Johns Hopkins University Applied Physics Laboratory, commissioned on the occasion of the lab's 75th anniversary. The case study updated and revisited an earlier study by Berl (1986). The case study was developed by a practitioner with advice from professional bibliometricians on the SIGMETRICS listserv. Assuming access to either Scopus from Elsevier or Web of Science (WoS) from Clarivate, [1] the remaining steps can all be completed with tools and knowledge available to the sophisticated practitioner.

BACKGROUND

Profiling multidisciplinary organizations or even states or countries over time can be done for historical reasons as here or for competitive intelligence, strategic planning, or to assess potential collaboration partners (cf Kostoff et al., 2007). This requires normalization over disciplinary areas as well as over time because the various fields represented within an organization may have dramatically different citing behaviors. Professional bibliometricians often license entire data sets from producers and are thus able to calculate normalizations that compare articles to others in their domain (determined by assignment, clustering, or citation networks) and their publication year. For the librarian, a more straightforward citation source normalization method suggested by Leydesdorff and Opthof (2010) is used for discipline and binning is used for time.

METHODS

The data set consists of two parts. First, metadata for articles written by APL authors were retrieved from WoS. Specifically, *Johns Hopkins University Applied Physics Laboratory* was searched using the Organizations-Enhanced field. This field uses the vendor's matching algorithms to group articles that list APL in the address field but use different variations of the address. Although WoS covers journals published from 1900 to the present, searches using this field only retrieve articles back to 1966 due to incomplete author affiliation coverage. In any case, Berl (1986) reviewed APL's publications prior to 1980 so this was chosen as the starting date. The results were limited to journal articles (n=6,666 on April 3, 2015) and the metadata for the articles were retrieved in plain text, full record format as and for the 1,000 most cited articles, the metadata for the articles that cited them was retrieved. A spreadsheet was used to track filenames and associate citing article metadata with the articles.

The second part of the data includes the updated citations and citing articles for the original Berl (1986) articles.

Once the data were gathered, R (R Core Team, 2017) was used to loop through the data files to calculate the fractional times cited. Longitudinal k-means clustering was used to look for sleeping beauties and strong performers among the articles. Finally, latent Dirichlet allocation (LDA) was used to categorize major topics in the highly cited articles. The code required to recreate these analyses will be posted on GitHub (https://github.com/cpikas).

RESULTS

The top articles since 1980 are shown in Table 1. The fractional counting smoothed some of the disciplinary differences and changed the ordering from the strict time

[1] References to commercial offerings are provided for clarity and are not intended as an endorsement.

cited. The poster will show the results from year bins and statistical analyses.

Citation	TC	TC_F
Mirowski,M., Reid,P., Mower,M., et al. (1980) Termination of Malignant Ventricular Arrhythmias with an Implanted Automatic Defibrillator in Human-Beings. *New England Journal of Medicine* 303,322-324. doi:10.1056/NEJM198008073030607	850	42.43
Kanungo,T., Mount,D., Netanyahu,N., Piatko,C., Silverman,R. and Wu,A. (2002). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 881-892. doi:10.1109/TPAMI.2002.1017616	789	40.56
Spall,J.(1992). Multivariate Stochastic-Approximation using a Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control*, 37, 332-341. doi:10.1109/9.119632	568	32.98
Raney,R., Runge,H., Bamler,R., Cumming,I. and Wong,F. (1994) "Precision SAR Processing using Chirp Scaling," *IEEE Transactions on Geoscience and Remote Sensing*, 32, 786-799. doi:10.1109/36.298008	315	25.63
Ottman,G., Hofmann,H., Bhatt,A. and Lesieutre,G. (2002). Adaptive Piezoelectric Energy Harvesting Circuit for Wireless Remote Power Supply. *IEEE Transactions on Power Electronics*, 17, 669-676. doi:10.1109/TPEL.2002.802194	407	23.72
Brown,M., Burschka,D. and Hager,G. (2003). Advances in Computational Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 25(8), 993-1008, doi:10.1109/TPAMI.2003.1217603	414	22.14
Franson,J (1989) Bell Inequality for Position and Time. *Physical Review Letters*, 62, 2205-2208, doi:10.1103/PhysRevLett.62.2205	471	20.16
Sharpe,W., Yuan,B. and Edwards,R.(1997) A New Technique for Measuring the Mechanical Properties of Thin Films. *Journal of Microelectromechanical Systems* 6,193-199. doi:10.1109/84.623107	285	16.85
Murphy,J. and Aamodt,L.(1980) Photothermal Spectroscopy using Optical Beam Probing - Mirage Effect. *Journal of Applied Physics*, 51,4580-4588, doi:10.1063/1.328350	276	16.57
Ott,E. and Sommerer,J.(1994) Blowout Bifurcations - the Occurrence of Riddled Basins and on Off Intermittency. *Physics Letters A*, 18, 39-47, doi:10.1016/0375-9601(94)90114-7	378	15.34

Table 1 Most Highly Cited Articles

Sources	Articles
Journal of Geophysical Research-Space Physics	1243
Johns Hopkins APL Technical Digest	638
Geophysical Research Letters	492
Icarus	317
Journal of Geophysical Research-Planets	167
Astrophysical Journal	165
Planetary And Space Science	148
Science	124
Annales Geophysicae	88
Acta Astronautica	76

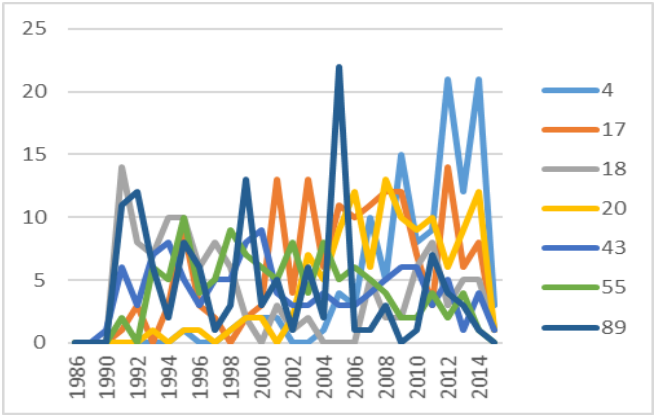
Table 2 Top Journals

Authors	Articles		Articles Fractionalized
Krimigis,SM	273	Lui,ATY	75.2
Meng,CI	233	Meng,CI	64.1
Lui,ATY	226	Krimigis,SM	53.3
Mitchell,DG	202	Newell,PT	49.7
Anderson,BJ	201	Cheng,AF	41.6
Newell,PT	191	Anderson,BJ	39.3
Roelof,EC	187	Roelof,EC	38.8
Solomon,SC	135	Sibeck,DG	36.5
Potemra,TA	132	Franson,JD	35.7
Ohtani,S	131	Greenwald,RA	33

Table 3 Top Authors

The most frequent journals are shown in Table 2. The top authors are shown in Table 3. Cross validation with log likelihood was used to determine the best number of topics for LDA (95). Topics were dominated by Space Weather, Magnetospheres, and related topics with 28 clusters. Geosciences for Earth and other bodies was second with 18 topics. Other topics were related to biomedicine, radar and communications, optics, and general engineering. The trends of topics with >110 articles are shown in Figure 1

Figure 1 Topic Trends Over Time



The longitudinal k-means clustering were used to identify citation trajectories (Genolini et al, 2016). The Berl papers easily showed a strong performer (on the Big Bang by Alpher), several typical performers with early spikes and then declining citations, and a few workhorse papers that continue to receive a low level of citation (see Figure 2). Papers 1980-2005 were mapped over PY+1 to PY+10.

While most showed a typical early peak, 10% gained in citations per year. These were typically review articles (see Figure 3).

Figure 2 Citation Trajectories for Early APL Articles

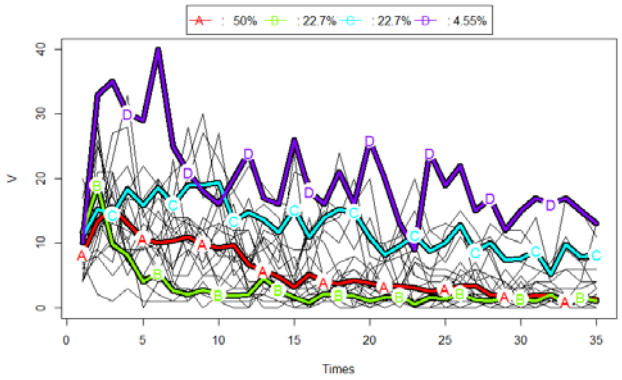
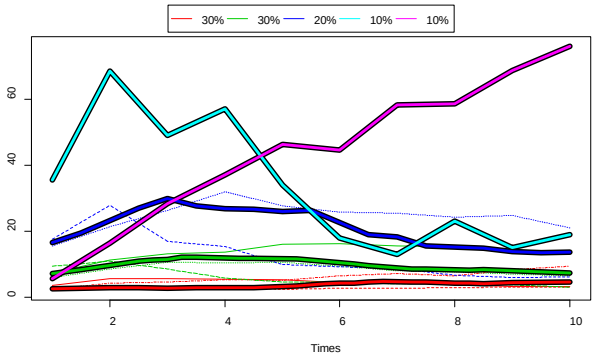


Figure 3 Citation Trajectories for 1980-2005 Articles



CONCLUSION

This work has shown that librarians with time and access to the databases can use source normalized techniques to profile an institution's work. Packages freely available for use in R and Python can make adoption and adaption of new techniques from the literature feasible.

ACKNOWLEDGMENTS

Loet Leydesdorff provided excellent advice that was very much appreciated.

REFERENCES

Berl, W. G. (1986). The 22 most frequently cited APLPublications. *Johns Hopkins APL Technical Digest, 7*, 221-232.

Genolini, C., Ecochard,R., Benghezal,M. Driss, R., Andrieu, S., & Subtil, F. (2016). kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS One, 11*, e0150738. doi:10.1371/journal.pone.0150738

Kostoff, R. N., del Rio, J. A., Cortes, H. D., Smith, C., Smith, A., Wagner, C., ... Tshiteya, R. (2007). Clustering methodologies for identifying country core competencies. *Journal of Information Science, 33*, 21. doi:10.1177/0165551506067124

Leydesdorff, L., & Opthof, T. (2010). Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. *Journal of the American Society for Information Science and Technology, 61*, 2365-2369. doi:10.1002/asi.21371

Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics, 109*, 2129-2150. doi:10.1007/s11192-016-2150-8

R Core Team (2017). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.