Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science

Ryan Whalen
Northwestern University
2240 Campus Drive
Evanston, IL 60208 USA
+1 773 679 1344
r-whalen@northwestern.edu

Yun Huang
Northwestern University
2240 Campus Drive
Evanston, IL 60208 USA
+1 847 491 2104
yun@northwestern.edu

Anup Sawant
Northwestern University
2240 Campus Drive
Evanston, IL 60208 USA
+1 847 491 2104
anup.sawant@northewestern.edu

Brian Uzzi
Northwestern University
Room 355 Leverone Hall
Evanston, IL 60208
+1 847 491 8072
uzzi@kellogg.northwestern.edu

Noshir Contractor
Northwestern University
2240 Campus Drive
Evanston, IL 60208 USA
+1 847 491 2104
nosh@northwestern.edu

ABSTRACT
In this paper we advocate increased use of textual data to develop new bibliometric methods. To demonstrate text's potential we propose a new bibliometric method that combines natural language processing with traditional bibliometric techniques to improve high impact science predictions. Relying upon the vast amounts of scholarly data now available online, we assemble a universe of scientific topics and use article text to measure the topical distance between citing and cited papers. We show that accounting for topical distance improves our ability to predict scientific impact. Citations from both topically distant and proximate papers provide more insight into an article's impact potential than those from papers with middling similarity.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: Text analysis

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms
Management, Languages, Theory

Keywords
Bibliometrics, Citation Analysis, Topic Modeling, Network Analysis, Science of Science

1. INTRODUCTION
Advances in computational power, natural language processing techniques, and the availability of scholarly texts in online databases offer great potential to transform bibliometric methods. For decades, bibliometrics has relied extensively on citation analysis, using citations to measure diverse facets of scholarly activity including knowledge flows [1], scientific knowledge domains [9], and journal impact [3]. For the most part, these studies rely on binary citation counting: references amongst articles are either present or not, with no room for varied types of relationships between articles.

In recent years, researchers have attempted to provide more nuanced measures by weighting citations based on network structure [2], field-specific citation patterns [6], and by creating indices to more accurately reflect scholarly impact [5]. But these more nuanced measures ultimately still depend solely on the citations, ignoring the content of the articles in question. The advent of large scientific databases, including records of not only the citations between articles but also the text of the articles themselves, makes this an unnecessary concession. We now have sufficient access to data, methods, and computational power to consider not only the presence or absence of a citation, but also the content of both the citing and cited articles. This enables us to better understand the processes of knowledge flow and identify influential articles.

In this position paper we advocate for greater attention to article content in bibliometrics. In order to do so, we propose and demonstrate a proof of concept that shows how accounting for the content of citing and cited enables more accurate prediction of ultimate scientific impact.

1.1 Scientific Impact
Measuring scholarly impact is big business. By quantifying how researchers affect knowledge generation, it influences hiring, promotions, grant awarding, and tenure decisions. Most current measures rely on counting citations [5], sometimes also considering the journal in which the citing publication appeared [4], or online downloads and references to the article in social media [7]. These methods are all plagued by a common problem: the binary nature of citations. Even methods that adjust for the citation's source are necessarily imprecise as they rely on a journal's historical impact to weight the citation, rather than weighting based on the nature of the actual citing article.

We address this weakness in traditional bibliometric methods by using natural language processing techniques to take into account the content of both the citing and cited article. Doing so allows us to compute the topical similarity between the original article and the article that cites it. This in turn allows us to distinguish between articles that influence particularly diverse areas of science and those that are only cited by articles within a relatively narrow subject area.

We rely on the store of scholarly data now available online to determine which subject areas exist and build our universe of scientific topics. As articles are published in online databases, authors and editors provide keywords to assist individuals as they search for literature that is relevant to their research interests. We

use these keywords to create a universal topic set that we can compare article text to in order to generate individual topic sets for each article.

1.2 Topical Diversity & Scientific Impact

Uzzi et al. show that high impact science is most likely to arise when scientists draw upon conventional combinations of sources but also cite to sources that are not typically combined with the other resources they rely upon [10]. Our method builds upon this observation, but flips the focus from backward citations (the outgoing references that a paper makes to previous research) to forward citations (the incoming citations that an article receives from subsequent research). Focusing on backwards citations provides insight into how an article grounds itself in existing science and draws upon prior work as inspiration. On the other hand, focusing on forward citations as we do demonstrates endorsement by third party researchers.

In addition to the alterations to established methods that we make by flipping the focus from backward to forward citations, we also provide a more nuanced measure of topical distance. Previous work has often relied upon proxy measures—such as the journal that an article appears in—to assess article content. By focusing on the actual language used in each article we analyze, our method provides a more accurate way to measure topical distance.

2. METHOD

Our method begins with the full text of scientific articles. For this early proof-of-concept we use the full text of all articles published in the *Social Networks* journal. Our dataset includes the full text of 809 articles published in *Social Networks* since it began publishing in 1979 as well as the citations between them. As a measure of ultimate scientific impact we calculate the total number of citations each article receives in the *Thomson Reuters Web of Science*.

In order to generate topical distance scores between articles, we first define the topical universe. To do so, we extract all of the keywords used in the *Web of Science*. This includes both the keywords that authors identify as relevant, and those that *Web of Science* editors assign to each article. At this proof of concept stage we focus on bigram keyword phrases. After stemming and stop word removal, this leaves us with approximately 85,000 keyword phrases.

To determine the topics covered by each paper we parse the full text of each article detecting its use of any keyword phrases in our set and the number of times those keyword phrases appear in the article. This keyword extraction method generates a set of keywords for each paper. We then use term frequency minus inverse document frequency (TF-IDF) to weight keywords by their importance, creating a weighted keyword vector for each paper. These weighted keyword vectors allow us to measure the strength of the topical similarity between two papers by computing the cosine similarity between their keyword vectors. For each *Social Networks* article we locate the other papers within the dataset that cite it and measure their similarity. Papers with similar content will have scores approaching 1, while papers with dissimilar content will have scores approaching 0.

3. RESULTS

To test whether accounting for topical diversity of citations improves our impact predictions, we can compare simple citation counting measures with measures generated using our keyword extraction and content comparison method. Table 1 shows the results of a regression model that uses a simple count of citations from papers within our dataset as a predictor for the number as a predictor of citations that the article will receive from journals outside our dataset. This straightforward model shows that simply counting citations within our dataset is a significant predictor of an article's greater impact. The r^2 demonstrates that this simple binary citation counting model explains approximately 54% of the variation in total citations received.

Table 1

	Coefficient
Intercept	-12.84683 *** (2.56159)
Local Citations	7.516734 *** (.29701)
Adj R-squared = 0.544 *** p < 0.000	

Table 2 shows results from a similar model. However, instead of using a simple citation count variable we include four variables that count the number of citations in each quartile of our topical similarity measure.

Table 2

	Coefficient
Intercept	-4.997327 * (2.302369)
1st Quartile	16.6598 5*** (.8262323)
2nd Quartile	2.032704 (1.207965)
3rd Quartile	1.336103 (1.283674)
4th Quartile	4.460859 *** (.9713052)
Adj R-squared = 0.6705 * p < 0.05 *** p < 0.000	

The results from this model show that accounting for distance provides significant improvement in our ability to predict total citation impact. Accounting for the topical similarity between citing and cited articles improves our model fit from 0.54 to 0.67. This appears to occur because some types of citations are better indicators of impact than others. Articles that are cited by papers that feature very dissimilar content (1st quartile) are likely to go on to receive many more citations than those that do not. Citations from papers with middling similarity (2nd and 3rd quartile) provide very little insight into an article's ultimate impact. Finally, citations from topically similar papers also provide a significant signal as to which articles will go on to have wider impact.

4. DISCUSSION

The above results show that accounting for article content along with citations in bibliometric analyses has significant promise. Not only does our model show improved predictive power when we account for both citations and topical similarity, the results also tell an interesting story about knowledge flow.

We see that the most topically distant citations are the best predictors of future impact. There are a number of plausible explanations of why this might be so. Citations from topically distant papers suggest that an article's content has broad appeal.

When research is relevant not just to others working in the same topical areas, but also to those working in diverse fields it is more likely to be cited in future years.

Increased visibility could also explain why citations from topically distant papers might lead to a higher overall citation count. If we assume there are distinct groups of scholars reading topically distant publications, then being cited by a topically distant paper exposes one's article to more readers. Exposing additional readers to the article increases its chance of being cited in the future.

4.1 Future Work

This proof of concept shares preliminary results from a relatively small dataset. It demonstrates that taking text into account in order to provide more information for bibliometric analyses is promising, but more work is required. Next steps will involve scaling up the analysis to include many more articles and many more disciplines.

In addition to increasing the dataset size, we will improve the distance measure by including both longer and shorter keyword phrases in the analysis. This will create much larger keyword vectors. To help deal with these larger vectors, future work will use dimension reduction to generate more accurate topical distance scores.

There is also work remaining to be done in determining how best to interpret and model topical distance scores. We choose to transform the scores into citation counts by distance quartile. This is not the only method that can be used to report or analyze citation distance data. Given that both high distance and low distance citations were significant predictors of impact while medium distance citations were not, the distance distribution appears to be a particularly promising measure for future exploration.

4.2 Bibliometrics and Text

Using keyword extraction and comparison to calculate topical distance between citing and cited papers is only one way that text can be used to advance bibliometrics. Advances in data availability and processing power open a variety of promising avenues for bibliometric methods development.

Topical variation between papers is not the only important dimension that citations vary along. Perhaps most importantly, citations vary in their importance. Some are perfunctory nods to the titans within a field, drawing little directly from the cited paper but including the citation for other reasons. Others build directly upon past work and clearly demonstrate the flow of knowledge. Taking text into account when assessing citation relationships presents a promising way for future researchers to distinguish between different types of citations. Some work has been done in this area [e.g. 11], but it remains a relatively undeveloped field.

Along with attempting to detect different degrees of citation strength, using article text shows promise in distinguishing between positive and negative citations. Although current methods tend to treat all citations as equal, we know that a significant portion of citations are included not to express agreement with the cited work, but rather to disagree with it. Or point out its weaknesses. Textual analysis shows promise in automating detection of citation sentiment [8], thereby improving bibliometric methods, but much work remains to be done.

5. Conclusion

There are myriad ways in which text can be used to improve bibliometrics. In this paper we present our topical distance measure as an example of the potential that using text along with citations has to improve bibliometrics. We show that accounting for the topical distance between citing and cited articles significantly improves our ability to determine which articles will go on to have high impact. The future will certainly bring improvements and further applications of this measure as well as others.

6. REFERENCES

1. Börner, K., Penumarthy, S., Meiss, M., and Ke, W. Mapping the diffusion of scholarly knowledge among major U.S. research institutions. *Scientometrics 68*, 3 (2013), 415–426.

2. Chen, P., Xie, H., Maslov, S., and Redner, S. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics 1*, 1 (2007), 8–15.

3. Garfield, E. Citation Analysis as a Tool in Journal Evaluation. *Science 178*, 4060 (1972), 471–479.

4. Glänzel, W. and Moed, H.F. Journal impact measures in bibliometric research. *Scientometrics 53*, 2 (2002), 171–193.

5. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America 102*, 46 (2005), 16569–16572.

6. Leydesdorff, L. and Bornmann, L. How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology 62*, 2 (2011), 217–229.

7. Piwowar, H. Altmetrics: Value all research products. *Nature 493*, 7431 (2013), 159–159.

8. Sendhilkumar, S., Elakkiya, E., and Mahalakshmi, G. Citation semantic based approaches to identify article quality. *Proceedings of International Conference ICCSEA*, (2013), 411–420.

9. Shiffrin, R.M. and Börner, K. Mapping knowledge domains. *Proceedings of the National Academy of Sciences 101*, suppl 1 (2004), 5183–5185.

10. Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. Atypical Combinations and Scientific Impact. *Science 342*, 6157 (2013), 468–472.

11. Wan, X. and Liu, F. Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, (2014).