Automated Citation Indexing of Government Documents

Ben Fisher Bruce Desmarais

October 22, 2015

Abstract

We extract citations from government documents.

Intended Journals *Journal of Informetrics* and *Scientometrics*

1 Introduction

Assessing the impact of scientific research in applied contexts is challenging, but necessary in evaluating the returns to society from investing in basic science. The measures used to assess the scientific importance of basic research are largely derived from citation in the bibliographies of scientific papers. One of the major challenges in assessing impact outside of the scientific community is that applied users of basic science rarely produce systematic and publicly available bibliographies. One exception is the recent focus on patents (Huang, Huang & Chen 2014, Liaw, Chan, Fan & Chiang 2014, Huang, Yang & Chen 2015, Wong & Wang 2015). One area of real-world impact of science that is highly relevant to all members of society is the use of science in the creation of government policy (NRC 2012). Indeed, the technical basis of government regulations has grown so salient in recent decades that regulations are regularly challenged in court on the basis of the science underpinning major provisions (Ellig & Fike 2013, Morrall & Broughel 2014). As such, while developing policies, government officials regularly create publicly available and heavily vetted documents in which they present the research on which a regulation is based. If we can systematically extract and organize this information, we can directly assess the impact of scientific research on public policy.

1.1 Tasks

- Why we would want to do this automatically (BD)

2 Literature Review

The development of citation indexing has primarily focused on indexing scholarly citations in the confines of large-scale projects like Garfield (1964)'s Science Citation Index, which eventually morphed present Web of Science or Lawrence, Giles & Bollacker (1999)'s Citeseer database. The process by which proprietary services like Web of Science or Google Scholar actually extract citations is opaque. The Citeseer project employs a heuristic, rule-based approach to identify citation strings within text. For example, the ParsCit parser developed by Councill, Giles & Kan (2008), which Citeseer currently uses, identifies the bibliography of a paper by searching for "bibliography" or synonyms like "references". If one of these is found past the first 40% of the paper, then the parser extracts all of the text past that line through the end of the paper or the appearance of another section heading, such as "appendix". The citation strings are then separated using regular expressions.

There has been limited publicly available work done on improving automated citation indexing beyond the Citeseer project. Papers on citation metadata extraction generally rely on citation datasets downloaded from Citeseer or manually compiled datasets, rather than developing new methods for parsing documents themselves (Anzaroot & McCallum 2013). Additionally, the automated citation indexing methods used by Citeseer and similar databases is tailored for extracting scholarly citations, making the application of existing algorithms problematic for extracting heterogeneous citations from government reports.

2.1 Tasks

- Review "altmetrics" lit (BD)
- read huang2014 and Liaw2014 to develop notes about framing and presenting the contribution (BF)
- Consider practical alternatives/extensions of Parscit (BD)

3 RIA Description

4 Research Design

We use 105 RIAs whose citations have already been extracted by hand as a test sample. The structure of these RIAs is very heterogeneous. They vary in length from a few dozen pages to hundreds of pages. References may appear in a reference section, in footnotes, or be scattered throughout the text. The citations of interest include both scholarly citations and non-scholarly citations, such as existing federal regulations, court cases, and treaties.

To assess how our own approach compares to existing citation indexing programs, we run the ParsCit program on the manually coded sample of RIAs. ParsCit was originally developed for indexing scholarly citations from journal articles for the Citeseer project (Councill, Giles & Kan 2008). In addition to extracting citations, the program also extracts citation metadata, such as author, article title, and journal. We do not expect it to handle metadata extraction for non-scholarly citations. However, it may be able to extract

text containing non-scholarly citations, which could then be manually coded. This would greatly speed up the current coding process, which requires research assistants to read the entire document for citations. If ParsCit performs well enough at extracting all citation types, then there is no need to develop new indexing software for this project.

We use four different string distance metrics to measure similarity between the titles of manually collected citations and those of the citations collected by ParsCit: Levenshtein, Jaccard, Sorenson, and partial string similarity. Levenshtein distance is measured by calculating the number of insertions, deletions, or substitutions needed to turn one string into another string [CITE Levenshtein 1966]. To calculate the Jaccard distance, the number of characters shared by two strings is divided by the total number of characters and subtracted from 1 [CITE Jaccard 1901]. The Sorenson distance measure is very similar to the Jaccard measure. Instead of looking at the number of shared characters, the Sorenson measure calculates the number of shared bigrams. This number is multiplied by 2 and then divided by the total number of bigrams in each string and subtracted from 1 [CITE Sorenson 1948]. Except for Levenshtein distance, all measures fall on a scale of 0 to 1, with 0 indicating an exact match and 1 indicating two completely different strings. The Levenshtein measure is normalized to the same scale for the purposes of comparison. The final measure, partial string similarity, uses substring matching and measures the percentage of characters in string a that appear in string b.

To determine which distance measure works best as a classifier in this case, we first hand code whether the best match from the ParsCit-collected citations suggested by each

distance measure for each hand coding whether the suggested match is actually a match. The best distance measure is that which provides the most correct matches. We then determine a cutoff point for the best distance measure that which best divides matches from non-matches and apply it to the whole sample.

4.1 Tasks

- See which distance metrics perfectly predict a match at a given distance (BF)
- Assess the cost of hand-coding our test data (BF)
- Assessing the precision and recall of parscit (BF)
- See how this changes over the RIA date (BF)

5 Results

We treat the values for each distance measure as pseudoprobabilities, which we use to generate Receiver Operating Characteristic (ROC) curves. Figure X displays ROC curves for each distance measure. The best measure is the one that has the greatest area-under-the-curve (AUC). In this case, partial string similarity is the best. We determine the ideal cutoff point by choosing the point that maximizes the sum of sensitivity and specificity. In this case, the best threshold is .15. Potential matches with a score below .15 are considered matches, while those above are considered non-matches.

Roughly 41% of the hand-coded citations are also present in the ParsCit-extracted citations. Table 1 displays the number of matched citations by the major citation types[1]. The major advantage of ParsCit should be it's ability to pick up scholarly journal citations. However, it actually performs quite poorly, picking up less than half of the scholarly citations present in the hand-coded dataset. This may be due to the fact that ParsCit was designed under the assumption that each document contains a reference section, which is not always the case with RIAs. Table 2 displays the number of matches and total citations for RIAs with bibliographies and RIAs without, while Table 3 displays matches by citation types for RIAs with bibliographies only.

Table 1: Hand-coded citations matched by type

Type	Matched	Total
Article	105	188
Book	66	131
Court Case	0	25
Database	147	349
Government Report	522	1308
Hearing	20	45
Magazine/Newspaper	10	13
NGO Report	243	622
Other	224	640
Reference Manual	21	64
Regulation	54	296
RIA	31	80
Scholarly Journal	576	1209

[1] not all citations fell into one of these categories

Table 2: Hand-coded citations matched by presence of bibliography

Bibliography?	Matched	Total
Yes	1788	3300
No	227	1713

Table 3: Hand-coded citations matched for RIAs with bibliography

Type	Matched	Total
Article	87	143
Book	43	80
Court Case	0	2
Database	111	208
Government Report	337	712
Hearing	16	20
Magazine/Newspaper	5	5
NGO Report	190	376
Other	206	397
Reference Manual	13	29
Regulation	32	69
RIA	21	54
Scholarly Journal	430	787

5.1 Tasks

- Meet with Lee's student (Kyle?) to see how we can improve

6 Conclusion

6.1 Tasks

- Why the world needs this? (BD)

References

Anzaroot, Sam & Andrew McCallum. 2013. "A New Dataset for Fine-Grained Citation Field Extraction." *Proceedings of the 30th International Conference on Machine Learning* .

Councill, Isaac, C. Lee Giles & Min-Yen Kan. 2008. "ParsCit: An Open-Source CRF Reference String Parsing Package." *LREC* .

Ellig, Jerry & Rosemarie Fike. 2013. "Regulatory Process, Regulatory Reform, and the Quality of Regulatory Impact Analysis." *Mercatus Center at George Mason University Working Paper* (13-13).

Garfield, Eugene. 1964. "Science Citation Index - A New Dimension in Indexing." *Science* .

Huang, Mu-Hsuan, Hsiao-Wen Yang & Dar-Zen Chen. 2015. "Increasing science and technology linkage in fuel cells: A cross citation analysis of papers and patents." *Journal of Informetrics* 9(2):237 – 249.
URL: http://www.sciencedirect.com/science/article/pii/S175115771500019X

Huang, Mu-Hsuan, Wei-Tzu Huang & Dar-Zen Chen. 2014. "Technological impact factor: An indicator to measure the impact of academic publications on practical innovation." *Journal of Informetrics* 8(1):241 – 251.
URL: http://www.sciencedirect.com/science/article/pii/S1751157713001132

Lafferty, John, Andrew McCallum & Fernando C.N. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." *Proceedings of the 18th Inernational Conference on Machine Learning* .

Lawrence, Steve, C. Lee Giles & Kurt Bollacker. 1999. "Digital Libraries and Autonomous Citation Indexing." *Computer* .

Liaw, Yi-Ching, Te-Yi Chan, Chin-Yuan Fan & Cheng-Hsin Chiang. 2014. "Can the technological impact of academic journals be evaluated? The practice of non-patent reference (NPR) analysis." *Scientometrics* 101(1):17–37.
URL: http://dx.doi.org/10.1007/s11192-014-1337-0

Morrall, John F & James Broughel. 2014. "The Role of Regulatory Impact Analysis in Federal Rulemaking." *Available at SSRN 2501096* .

NRC. 2012. Using Science as Evidence in Public Policy. Technical report National Research Council: The National Academies Press Washington, DC: .

Wong, Chan-Yuan & Lili Wang. 2015. "Trajectories of science and technology and their co-evolution in BRICS: Insights from publication and patent analysis." *Journal of Informetrics* 9(1):90 – 101.
URL: http://www.sciencedirect.com/science/article/pii/S1751157714001084