Automated Citation Indexing of Government Documents

Ben Fisher Bruce Desmarais

November 11, 2015

Abstract

We extract citations from government documents.

Intended Journals

- *Journal of Informetrics*
- *Scientometrics*
- *PLoS ONE altmetrics collection*
- *WWW Workshop on Big Scholarly Data*
- *PNAS*

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.

Our goal is to assess the degree to which a publication in a peer-reviewed scientific journal has been directly used in benefit-cost analyses conducted by policymakers. These benefit-cost analyses, termed "Regulatory Impact Analyses" (RIAs) in the US federal government, directly inform the process of crafting government regulations. It is now standard procedure in dozens of national governments, and many sub-national governments such as the US states, for policymakers to conduct a benefit-cost analysis to inform the provisions

in government regulations (Hahn & Tetlock 2007). By indexing these analyses for citations to the scientific literature, we construct a measure of the degree to which an individual paper or journal, has directly informed the process of crafting government regulations. We build upon a dataset introduced by (Desmarais & Hird 2014) in which citations from 104 RIAs covering major regulations passed by the US federal government during the period 2008–2012. Though feasible for small collections of benefit-cost analyses, hand-indexing would be prohibitively expensive for larger corpora, which could easily number in the tens of thousands if they covered longer spans of time, other nations, or subnational governments. We use the dataset introduced by (Desmarais & Hird 2014) to train and evaluate procedures for automating citation indexing for RIAs. We then apply automated indexing to an expanded corpus of RIAs from the US federal government, and use these citations to construct a regulatory policy impact factor.

1.1 Tasks

- Write "outline of the paper" paragraph (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.

Citation indexing can be seen as a special case of token tagging in text (i.e., part of speech (POS)) tagging. Consider a document in which every word (or token) has a tag from a given tag set (e.g., POS = {noun, verb, adjective,...,}). The word "run", for example, would likely be tagged as a verb, but could also be tagged as a noun (e.g., "I am going on a run"). Citation indexing could be seen as a two-part token tagging task. First, within a document, we need to tag tokens regarding whether they are within a citation instance, with adjacent sets of tokens in a citation instance representing a single citation. Second, after extracting citations, tagging the tokens within the citation index as title, author, etc. For our purposes, we really only need a tagger that performs well at the first task, as the

second task can be performed by hand or with the assistance of other citation indexes (e.g., Google Scholar). The Stanford POS tagger can be used to train a tagger using a custom tag set (Toutanova, Klein, Manning & Singer 2003). We could use a grep for scientific cite titles in our hand-coded data to tag tokens as starting a cite title, within a cite title, ending a cite title and not in a cite title. Other tags may be useful, but whatever the tag set, we could train the Stanford Tagger and test extraction rules for automatically indexing cites.

2.1 Tasks

- Review "altmetrics" lit (BD)
- read huang2014 and Liaw2014 to develop notes about framing and presenting the contribution (BF)

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. The source documents are in PDF format, which we convert to plain text for the purposes

of analysis using Apache Software's PDFBox program.

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, Sorensen, 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 (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 (Jaccard 1901). The Sorensen distance measure is very similar to the Jaccard measure. Instead of looking at the number of shared characters, the Sorensen 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 (Sørensen 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.

We can think of our use of Levenshtein distance to label a match as a diagnostic test with X% accuracy. Given the standard testing adjustments (Pewsner, Battaglia, Minder, Marx, Bucher & Egger 2004), we can say that ParsCit identifies A-B% of all citations in the data.

Steps: Get Parscit, find CiteULike, calculate precision and recall

To better assess the performance of ParsCit, we filter the results through CiteULike, an open source citation database website, by checking whether a citation title supplied by ParsCit is found using CiteULike's search engine. The matching procedure is the same as the one for the hand-coded dataset, using Levenshtein distance with a cutoff point. Using this in conjunction with the hand-coded data, we calculate precision and recall metrics

for ParsCit. Performance is poor on both metrics. Precision is .07, while recall is .2. A review of some of the scholarly citations from the hand-coded dataset suggests that not all scholarly articles are on CiteULike, which may explain the results.

4.1 Tasks

1. Make sure data we are using on articles comes from WebofScience via DOI download. BF send comma-separated list of DOIs to BD. BD return DOI-based data frame to BF.

2. Normalize journal impact factors by volume of articles published (BF). Get data on journals from JCR. For each journal, try to download number of articles published and impact factor data for each year going back as far as you can.

3. Calculate precision after CiteULike matches.

4. Calculate five year impact factor (BF). For time T, the number of cites to that journal between T and T+4 (verify this shouldn't be T+5).

5. Create journal ranking scatterplot in hand-coded and parsCit coded (BF).

6. Calculate average difference in rank between hand-coded and Parscit coded (BF).

7. Need to figure out if cite records from ParsCit can pull DOIs from the web (BF).

8. What is the rate of article recovery if we use CiteULike matched records rather than ParScit matched records (i.e., re-do Figure 4 throwing out parscit-discovered records that cannot be found on CiteULike)? (BF)

9. Get big trove of RIAs (BD)

5 Results

We treat the values for each distance measure as pseudoprobabilities, which we use to generate Receiver Operating Characteristic (ROC) curves. Figure 1 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 and Levenshtein distance perform roughly the same, while Jaccard and sorensen measures do worse. We use Levenshtein distance due to its superior performance and wide use. Using the Levenshtein measure, 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 .6. Potential matches with a score below .6 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.

We also look at how these metrics change over time. The sample of RIAs we use

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

here covers 2008 through 2012. Figure 2 shows the average accuracy of ParsCit for all citation types by year when compared to the hand coded data. Figure 3 and Figure 4 look at scholarly citations specifically. The former shows the total number of scholarly citations from the hand coded data by year, while the latter shows the percentage of these that were actually picked up by ParsCit for each year. ParsCit maintains a fairly consistent accuracy of around 50%. One caveat is that this is all under the same presidential administration, and these metrics may change with a change in president and consequent policies.

Figure 5 shows journal impact factor statistics for the full sample of scholarly citations and various subsets of the data. These subsets include citations found by ParsCit, citations not found by Parscit, citations from RIAs with bibliographies, and citations from RIAs without bibliographies. Overall, the average impact factor across the different samples is fairly similar, with the found by ParsCit and no bibliography samples slightly higher on average. Table 4 and Table 5 show the top journals for the full sample of scholarly citations and the sample of those found by ParsCit respectively. The two are quite similar and, and with one exception, consist entirely of scientific and health journals.

5.1 Tasks

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

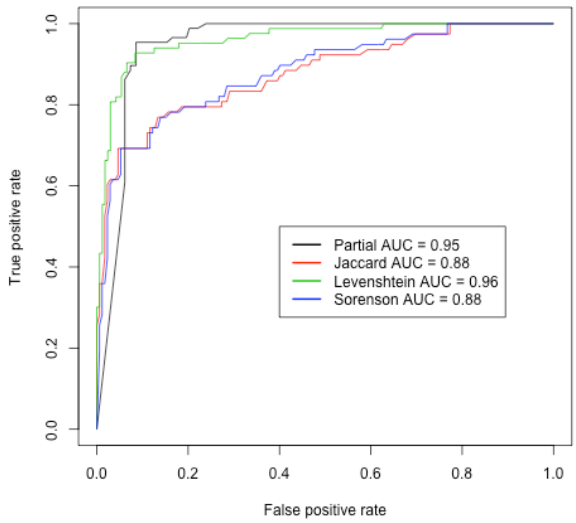

Figure 1: ROC Curves

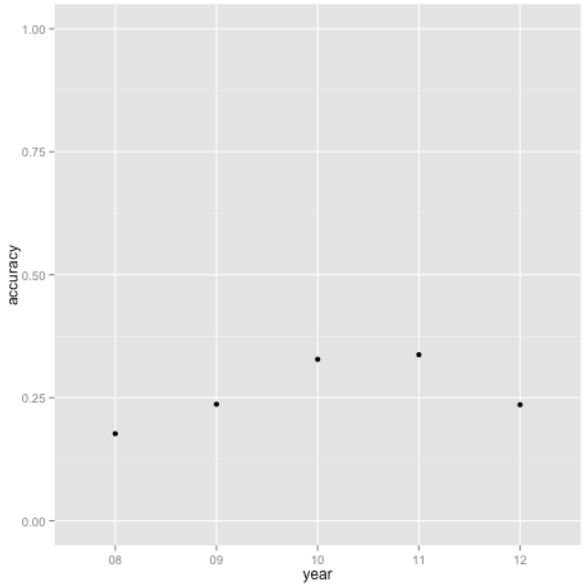

Figure 2: ParsCit Accuracy by Year

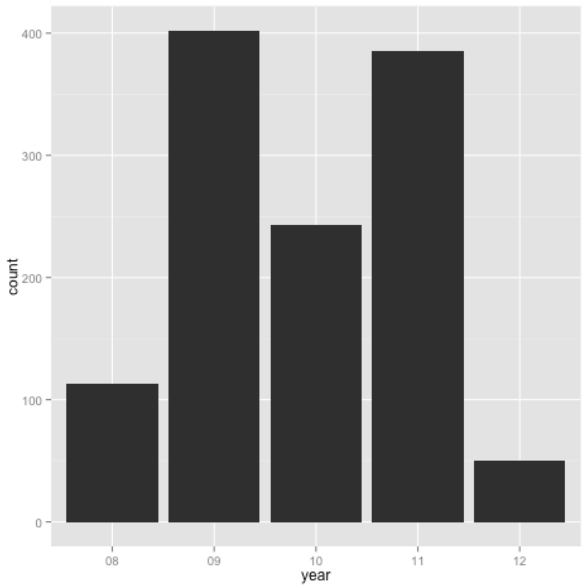

Figure 3: Scholarly Citations by Year

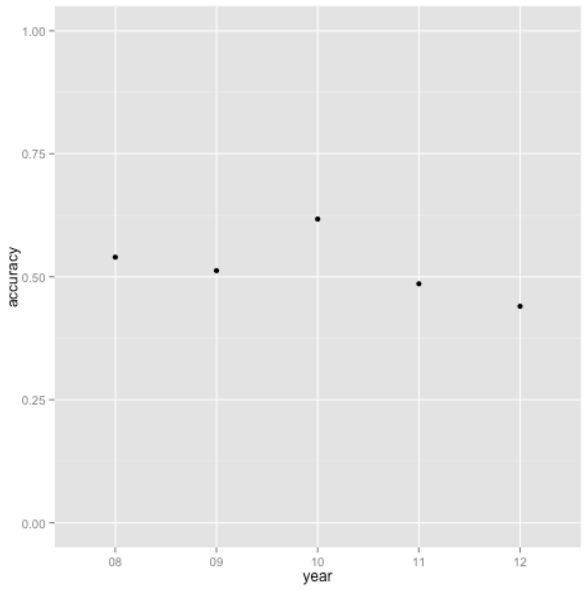

Figure 4: ParsCit Scholarly Accuracy by Year

Table 1: Hand-coded citations matched by type

Type	Matched	Total
Article	96	188
Book	65	131
Court Case	1	26
Database	115	334
Government Report	462	1308
Hearing	16	45
Magazine/Newspaper	10	13
NGO Report	217	622
Other	183	640
Reference Manual	14	67
Regulation	46	296
RIA	28	80
Scholarly Journal	626	1209

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

Bibliography?	Matched	Total
Yes	1788	3340
No	227	1750

6 Conclusion

6.1 Tasks

- Why the world needs this? (BD)

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

Type	Matched	Total
Article	76	143
Book	45	80
Court Case	0	2
Database	92	208
Government Report	299	712
Hearing	13	20
Magazine/Newspaper	5	5
NGO Report	176	376
Other	166	398
Reference Manual	7	32
Regulation	35	69
RIA	20	54
Scholarly Journal	470	787

Table 4: Top Journals from Hand Coded Sample

Journal	Cites	Impact Factor
Epidemiology	62	6.196
Environmental Science and Technology	56	5.330
Environmental Health Perspectives	35	7.977
American Journal of Respiratory and Critical Care Medicine	33	12.996
Energy Policy	30	2.575
Environmental Research	30	4.373
American Economic Review	3.673	
Atmospheric Environment	25	3.281
Health Affairs	24	4.966
Journal of Environmental Economics and Management	24	2.394

Table 5: Top Journals from ParsCit Sample

Journal	Cites	Impact Factor
Epidemiology	30	6.196
Environmental Science and Technology	27	5.330
Environmental Health Perspectives	25	7.977
Atmospheric Environment	20	3.281
Journal of the American Medical Association	19	35.289
Environmental Research	30	4.373
American Journal of Respiratory and Critical Care Medicine	33	12.996
New England Journal of Medicine	14	55.873
Circulation	12	15.073
American Journal of Public Health	12	4.552

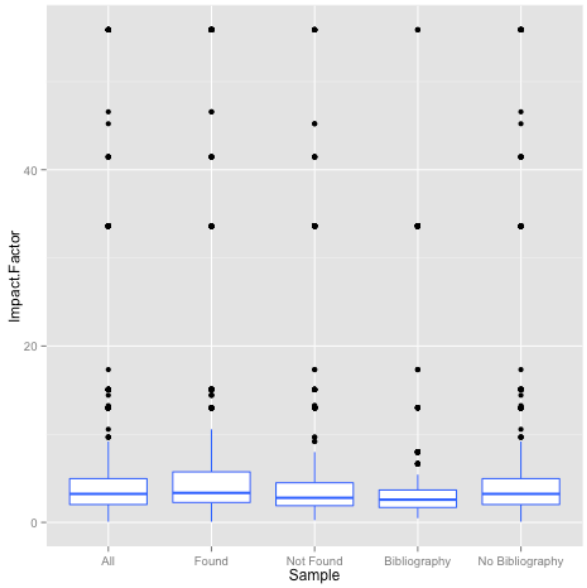

Figure 5: Impact Factor by Sample

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* .

Desmarais, Bruce A. & John A. Hird. 2014. "Public policy's bibliography: The use of research in US regulatory impact analyses." *Regulation & Governance* 8(4):497–510.
URL: http://dx.doi.org/10.1111/rego.12041

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* .

Hahn, Robert W & Paul C Tetlock. 2007. "Has economic analysis improved regulatory decisions?" *AEI-Brookings Joint Center Working Paper* (07-08).

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

Jaccard, Paul. 1901. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura." *Bulletin de la Société Vaudoise des Sciences Naturelles* .

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* .

Levenshtein, Vladimir. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* .

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: .

Pewsner, Daniel, Markus Battaglia, Christoph Minder, Arthur Marx, Heiner C Bucher & Matthias Egger. 2004. "Ruling a diagnosis in or out with ?SpPIn? and ?SnNOut?: a note of caution." *BMJ* 329(7459):209–213.

Sørensen, Thorvald. 1948. "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and its Application to Analyses of Vegetation on Danish Commons." *Kongelige Danske Videnskabernes Selskab* .

Toutanova, Kristina, Dan Klein, Christopher D Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics pp. 173–180.

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