Automated Citation Indexing of Government Documents

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 patents [CITES]. 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 [CITE SCIENCE IN POLICY REPORT]. 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 [CITES]. 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.

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.

3 RIA Description

4 Modeling Description

We use 104 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.

5 Results

6 Conclusion

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* .

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

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* .