Automated Citation Indexing of Government Documents

1 Introduction

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

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* .