2011 11th IEEE International Conference on Data Mining Workshops

Poll: A Citation Text Based System for Identifying High-Impact Contributions of an Article

Lalith Polepeddi, Ankit Agrawal, Alok Choudhary
Department of EECS
Northwestern University
2145 Sheridan Rd
Evanston, IL 60201
USA
lpolepeddi@u.northwestern.edu, {ankitag, choudhar}@eecs.northwestern.edu

***Abstract*—The body of scientific literature is growing yearly, presenting new challenges in accurate retrieval of relevant publications. Citation sentences stand to be a useful way to concisely represent the main contributions of a publication. In this paper, we present Poll, a prototype of an academic search engine which utilizes citation sentences to indicate the most important contributions of a cited publication.**

Keywords-information retrieval; academic search; citation sentences

I. INTRODUCTION

The advancement of scientific research and its widespread dissemination has resulted in a huge amount of scientific research and process data, representing the discussions and outcomes of complete and in-progress research projects, which include but are not limited to, research papers, technical reports, discussion forums, mailing lists, and software descriptions. There is a huge amount of data and potential knowledge buried in these sources, which could be analyzed to discover many important features valuable not only for scientific discovery, but also for making the discovery process more effective, efficient, and productive for the researchers.

Academic publications are probably the most formal and popular form of such scientific research data. Information retrieval from academic publications is a challenging task. The large and continually growing body of scientific literature creates difficulties in navigating and accurately finding relevant publications. In 2010 alone, there were 920,674 biomedical publications in the PubMed database, and the number of publications has been increasing by an average factor of 1.06x each year for the past decade. As more journals utilize the web as their primary publishing medium and digitize their archival content, new challenges are presented in finding relevant publications and distinguishing their individual contributions from previous work.

Academic search engines have had to scale in order to manage the growing body of literature. There are several search engines currently in existence, most of which are useful within specific domains. PubMed provides access to over 10 million unique publications on biomedical topics. IEEE Xplore, ACM Digital Library, and CiteSeerX are search engines primarily for technical literature in engineering. Google Scholar and Microsoft Academic Search are multidisciplinary search engines that cover literature from a broad range of disciplines.

These search engines approach the problem of academic search much like general web search. A typical query is constructed from keywords capturing a specific publication or a general topic of interest. Literature is served back to the user based on matches to indexed keywords, ranking algorithms, and other features as proxies for relevance. This approach is problematic for academic search because the results served are highly dependent on the initial keywords used. It is subsequently time-consuming for the user to ascertain with confidence whether the results are both relevant in their content and authoritative in their contributions. This gives way to multiple levels of indirection wherein the user must comb through the literature by back-tracing through bibliographic references to find explanations of a paper's relevance and importance to the field.

Furthermore, this approach places responsibility on the user to read through a publication and correctly interpret its main points. This is problematic because it is time-consuming and error-prone to comprehend a paper, especially for non-experts.

Acquiring a meaningful handle on a publication should not unnecessarily consume time or require domain-expertise. Instead, it should rely on organizing key pieces of information. Scientific publications are interlinked by citations. This connectivity provides considerable structure on top of the content of each publication, such as citation network topology and citation text. In this paper, we take advantage of citation text to produce a global summary of a publication's main contributions. We also present a prototype of an academic search engine which makes use of the semantic information present in citation text. We chose our system name to be Poll because the search results it yields represents the community's "polled" consensus on a publication's contributions. Poll is designed to enable quick and accurate comprehension of a scientific publication's

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.136

IEEE computer society

main contributions. The prototype is available at http://info.eecs.northwestern.edu/~lpolepeddi/poll/.

Section II gives an overview of our goals in designing Poll. In Section III, we describe the semantic value present in citation text. To test the utility of citation text for academic search, we implemented a search engine (Section IV and V). In Section VI, we discuss future directions for this work.

II. Design Goals

Our central goal is to find the main points of a publication quickly. Citation text is particularly useful to do this because it represents human-curated, peer-reviewed summaries of a publication's content. Our system maps citation text to the publication(s) it attributes, thereby computationally organizing key pieces of information in favor of manual back-tracking through references and keyword hunting.

Another important goal is to ensure this mapping is up-to-date. Maintaining a current citation network is critical to find newly established semantic relationships between publications. Maintaining a comprehensive archive of these related semantic content is important to us because we think some of the most interesting research will involve mining this data. This content is otherwise difficult to obtain because they are distributed across several individual publications.

Our final design goal is to improve the quality of academic search engines. We believe that academic search engines should tell stories. A publication is meaningful not in isolation, but in context of other related publications. Gleaning this context manually is time-consuming and error-prone because it involves back-tracking through references. Academic search engines generate citation networks by virtue of crawling publications, and this network can be leveraged to serve up the unique contributions of a given publication as reported by other related publications.

III. Scientific Research Network

Most publications include a literature survey to define the context of their presented work. In theory, every claim made in a publication must be substantiated with attribution to previous work. A paper derives its credibility in part from well-chosen citations. Authors pick and choose the best content from prior work and summarize the main contributions of that work with one or two citation sentences. The citation text offer human curated micro-reviews of the cited paper's contributions. The relationship between papers and their link text is described in Figure 1.

Current academic search engines incorporate citation structure prominently into their ranking algorithms in the form of citation counts. Citation text is not noticeably featured. We believe there is additional semantic information in citation text about an article's unique contributions that can help improve academic search. To achieve this goal, we have built a crawling and mapping system that keeps a database of publications and the citation text pointing to them. The database will searchable from our search engine Poll.

IV. Scientific Research Network

Poll follows a pipeline architecture, illustrated in Figure 2. First, the system checks a list of journals to discover newly published articles. The full text each article is retrieved, passed through a pre-processing phase which normalizes text into a standard form, and subsequently stored. Full text is tokenized into sentences, and sentences containing citations are mapped to their corresponding referenced source. Most of Poll is implemented in Python.

A. Journal Server

The Journal Server discovers newly published articles. The dataset for the current system prototype is the PubMed Central Open Access Subset, a collection of articles in the biomedical sciences provided by the US National Institutes of Health's National Library of Medicine (NIH/NLM). While articles in the Open Access Subset are still protected by copyright, they have been made available under a Creative Commons or similar license that allows for more liberal use. As of 10/14/2011, there are 355,670 publications archived in the Open Access Subset. As most journals require paid subscriptions to access full-text publications, we chose the Open Access Subset for our system prototype to demonstrate the concept of utilizing citation text.

B. Crawler

Fetching newly published articles is done in parallel by a mulit-threaded crawler. A publication is assigned a unique ID, its title, authors, journal, date published, PubMed Identifier (PMID), PubMed Central Identifier (PMCID), full text content when available, and URL are retrieved, and subsequently stored into a repository. We use MySQL for Poll's repository. Tables are indexed on title, PMID, and PMCID for fast retrieval.

C. Mapper

The mapper performs a number of functions. It parses citation sentences from a publication's full text. Full text is tokenized into sentences, and citation sentences are extracted based on the presence of anchor links pointing to the publication's bibliography. Citation sentences are mapped to

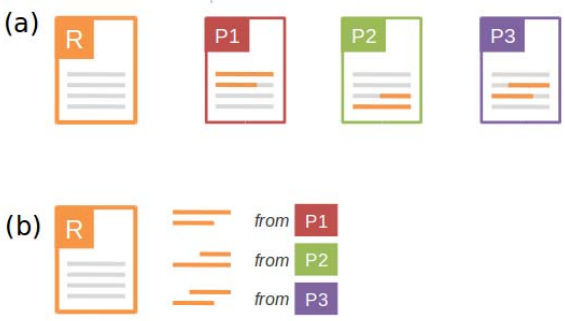

Figure 1. (a) Bibliographic reference *R* is cited by three publications *P1*, *P2*, and *P3*. Each of these publications summarize the relevance of *R* in the form of their citation sentences. These citation sentences can be aggregated and presented as the major contributions, or main points, of *R* (b).

their corresponding bibliographic reference and stored. This mapping is displayed via the Searcher component of the system when a user submits a query for a paper of interest.

V. Implementation

We implemented a search engine that uses publications from the PubMed Central Open Access Subset in order to test the usefulness of link text in search (available at http://info.eecs.northwestern.edu/~lpolepeddi/poll/). A user queries the system using keywords capturing a general field or specific publication of interest. To answer the query, the system returns a list of publications that contain one or more of the search terms in their titles. The user selects a publication of interest and is returned with a list of citation text from citing work. Each citation text is listed with the title and authors of its corresponding publication.

Figure 3 gives a sample of results returned for the publication "The human disease network" [9]. The queried publication's citation, title, and authors are given at the top of the page, while its summary as cited by other publications are given under "Main Points." The citing publication and authors are given under each citation summary. A user can skim through these sentences and quickly get a reasonably good understanding of the paper's contributions.

VI. Conclusion and Future Work

In this paper, we mapped citation sentences to their cited publications as a way to quickly grasp the main points of those publications. We implemented Poll, a prototype of an academic search engine, to test the value of citation text in academic search.

In our implementation, we saw several opportunities to improve search quality. Currently Poll is limited to data available in the PubMed Central Open Access Subset. Although there are millions of papers available in PubMed, their text may not be used for research purposes due to stringent copyright restrictions. We were only able to use the ~355,000 papers deposited in the Open Access subset since they have been made available under a Creative Commons or similar license. Poll relies on a rich citation network topology, specifically a rich backlinking structure, and its usefulness improves as it incorporates more publications into this network. The main difficulty in implementing our system was an inability to access orders of magnitude more publications due to copyright restrictions. In order achieve our design goal of maintaining a comprehensive archive of related semantic content, Poll will need to access a broader range of articles from publishers.

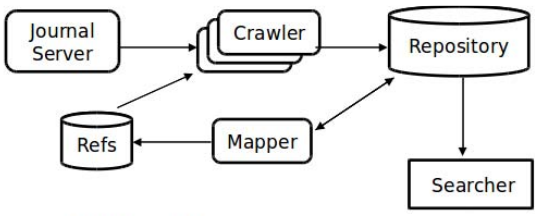

Figure 2. High level Poll architecture.

Poll will likewise need to support heterogeneity in citation formatting. PubMed Central is fairly consistent in using in-line parenthetical citations with hyperlinks to the corresponding reference, so citation sentences are easy to identify. However, citation formats can vary across different publishers, so Poll must be able to handle this variance in order to reliably detect citation sentences. Sugiyama et al developed a supervised classifier to detect citation sentences in publications that could be an option to incorporate into Poll [10].

There are also opportunities for improving the results yielded to the user. Currently Poll's mapper extracts only the sentences with in-line parenthetical citations. In doing so, some of the sentences served in the results read as if they are out of context. For example, the second result sentence in Figure 3 reads, "From this perspective, metabolism-related diseases are of special interest because high-quality molecular interaction maps exist for human cell metabolism (15, 16)..." It is unclear what perspective this sentence is referring to. The result sentence would likely be clarified if the preceding sentence was also included. Therefore, including the sentences immediately preceding and following a citation sentence may provide more context and make for a more coherent result.

The second result sentence in Figure 3 also refers to four papers, and it is unclear which of these four relates to "The human disease network." Applying a formatting signal, such as selectively boldfacing the respective portion of the multi-citation sentence, would be useful to clarify which citation summary is a main point of the queried publication.

Ranking algorithms may also be applied to Poll's results to present the citation sentences in order of relevance. Based on our observations, a high citation count doesn't necessarily mean that the publication has more relevant citing sentences. Yet citation count figures prominently into the ranking algorithms used by existing academic search engines. We see an opportunity to improve ranking based on a publication's total factual content in addition to citation index. Applying ranking algorithms will also organize results for sensible consumption by the user. If a reference is cited by several articles, it will have several citations sentences and the user will have to go through a lot of text. By including sentences immediately preceding and following a citation sentence, this problem would be compounded. Ranking citation sentences will be a useful step before presenting them to the user.

Acknowledgment

We would like to thank Sanchit Misra for system administration support. We would also like to thank Ed Sequiera for discussions about PubMed Central copyright policies, and the PubMed Central team for making the Open Access Subset available. This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, and DE-SC0005340.

REFERENCES

[1] W. Lehnert, C. Cardie, and E. Rilofl, "Analyzing research papers using citation sentences," Proceedings of the 12 Annual Conference Of The Cognitive Science Society, pgs. 511-S18, 1990.

[2] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, and D. Radev, "Blind men and elephants: What do citation summaries tell us about a research article?," Journal of the American Society for Information Science and Technology, 2008.

[3] P.I. Nakov and A.S. Schwartz, M.A. Hearst, "Citances: Citation Sentences for Semantic Analysis of Bioscience Text," Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics, 2004.

[4] A. Schwartz and M. Hearst, "Summarizing key concepts using citation sentences," Proceedings of the HLT-NAACL BioNLP Workshop on Linking National Language Processing and Biology, pgs. 134–135, 2006.

[5] V. Qazvinian and D.R. Radev, "Scientific paper summarization using citation summary networks," Proceeding COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics, 2008.

[6] V. Qazvinian, D.R. Radev, and A. Özgür, "Citation summarization through keyphrase extraction," Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, 2009.

[7] H. Yu, S. Agarwal, and N. Frid, "Investigating and annotating the role of citation in biomedical full-text articles." Bioinformatics and Biomedicine Workshop, 2009.

[8] A. Ritchie, S. Teufel, and S. Robertson, "Using Terms from Citations for IR: Some First Results," Proceedings of the European Conference on Information Retrieval (ECIR), pgs. 211-221, 2008.

[9] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabási. "The human disease network," Proceedings of the National Academy of Sciences, pgs 8685-8690, 2007.

[10] K. Sugiyama, T. Kumar, Min-Yen Kan, and R.C. Tripathi, "Identifying citing sentences in research papers using supervised learning," Dept. of Comput. Sci., Nat. Univ. of Singapore, Singapore, Singapore, pgs 67 - 72, 2010.

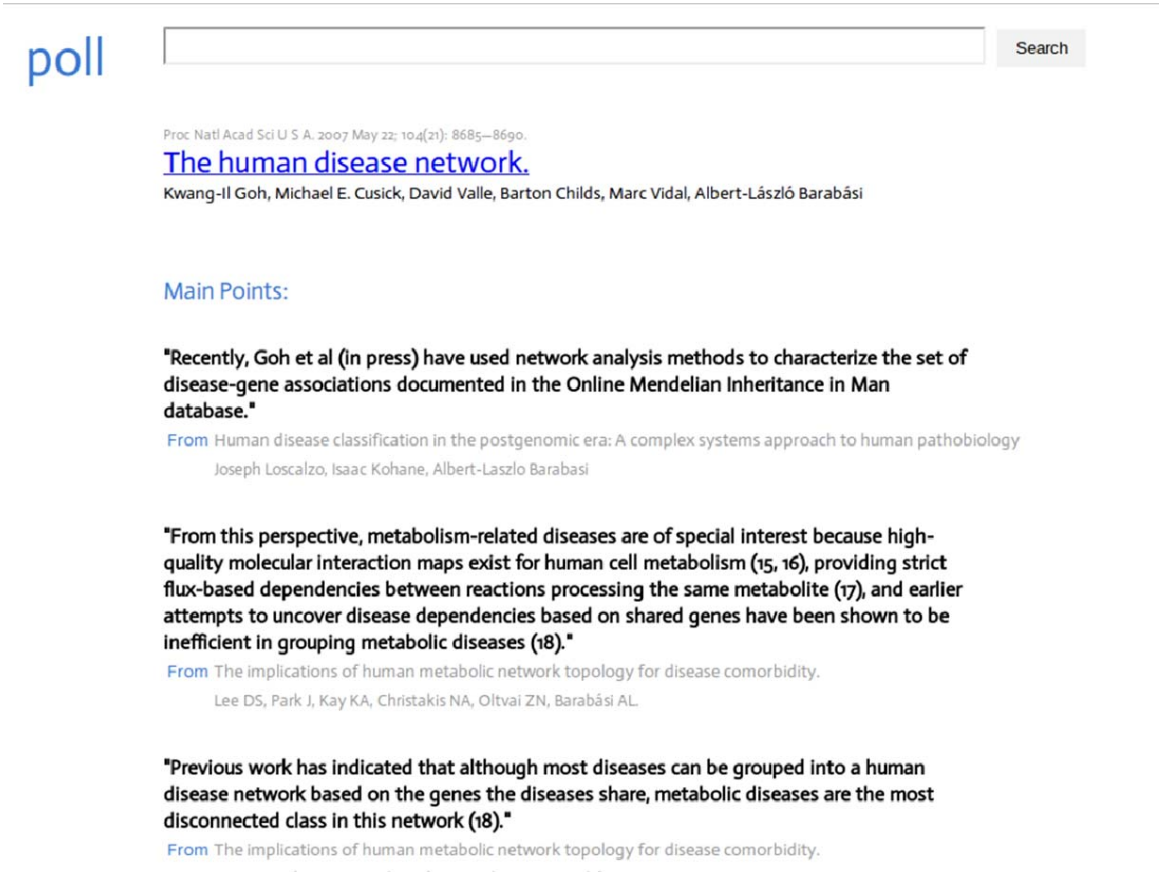

Figure 3. Sample results from Poll for the publication "The human disease network."