ITEC 810 Project Proposal:
A Platform for Citation Analytics

Robert Dale

26th February 2009

SUMMARY

In the research world, there are various measures used to assess the impact of a given publication. Real world limitations deriving from the infeasibility of manually analyzing large quantities of data mean that coarse-grained aggregated metrics (such as journal impact factors) are used in preference to finer-grained, and more accurate, measures such as citation counts. Tools such as Microsoft's Live Search Academic and Google Scholar provide the basis of an infrastructure that can go beyond coarse-grained metrics, but the primary focus of these tools so far is on finding publications, not on assessing the citing relationships between publications. The aim of this project is to build a platform for what we call **citation analytics**, a new area that will integrate work on information extraction with work on digital libraries to build sophisticated technologies for citation linkage, tracking, and assessment.

1 PROJECT DESCRIPTION

1.1 Background

There is a burgeoning interest across the world in measuring the impact and quality of the work of researchers. This has led to increasing scrutiny of existing aggregated metrics such as journal impact factor [Garfield 1955], and a slew of new measurements such as the h-index [Hirsch 2005]. It is widely recognized that the number of times a particular piece of work is cited is an important element in determining its significance; but, faced with the immense quantity of published material now in existence, it is clearly impossible to obtain this data manually. New web-based search facilities such as Microsoft Live Search Academic and Google Scholar provide some help in automating citation counts, but suffer from problems such as difficulty in identifying self-citation, and name matching issues resulting from the ambiguity of underspecified or widely-shared names.

This project aims to leverage our existing experience in information extraction and text analytics work for [Dale et al 2004; Mazur and Dale 2007] to construct a set of tools for what we call **citation analytics**: the high-quality automatic mining of citation patterns in scholarly documents in order to provide fine-grained assessments of research impact. In the last few years, text analytics – the area of natural language processing concerned with extracting useful information from large bodies of text – has begun to demonstrate its practical utility; but these techniques have not yet been applied comprehensively in citation analysis. This project aims to develop a platform for high-quality automatic citation analysis. We already have a number of key components developed in the scope of other projects; the purpose of the proposed project is to allow us to build a proof of concept that will demonstrate how information extraction technologies can be put to use in this increasingly important area.

1.2 Aims, Significance and Expected Outcomes

In recent years, the measurement of the quality of research has become an important issue worldwide, most notably in the UK's RAE, and in Australia's more recently aborted RQF and its replacement, the ERA. In other countries (notably in Asia and the US), tenure committees scrutinize publication lists and try to assess the quality of promotion candidates' research output. Lacking a more fine-grained mechanism, a researcher's impact is measured by the fora in which he or she publishes, rather than on the basis of the intrinsic merit of the publications themselves. Although metrics such as journal impact factor can serve as rough diagnostics, discomfort with ISI's methodology for determining what publications are in-scope has led to competing products such as Elsevier's Scopus, and to the much more inclusive web-based coverage of Microsoft Live Search Academic Search and Google Scholar.

In an increasingly wikified electronic world, we believe that it is automatic tools of the latter type that will deliver the most useful results. Unfortunately, their ability to carry out some of the most fundamental operations required for citation analysis quality control – specifically, detection of duplicates through metadata matching, detection of self-citations, and high-accuracy matching of metadata from reference lists with cited works – is limited. This is an inevitable consequence of their developers' primary focus on providing a search tool with high recall (i.e., the minimization of 'missed data'), leaving the precision of the results (i.e., whether or not the search returns 'false positives') for the user to consider.

The problems are not insurmountable, however. Text analytics tools, which use natural language processing techniques to overcome the ambiguities faced in extracting useful information from text, have been successfully applied in a number of areas, including finance and biotechnology. To date, these techniques have not been applied in the context of citation analysis; our aim is to build a platform that will enable these techniques to be used to deliver automatic high quality and reliable citation analysis.

In the medium to longer term, there is the potential here to produce a paradigm shift in how research impact is measured, moving from aggregated measures such as journal impact factor, through more specific measures based on detailed citation counts, to even the analysis of meaning of the citations themselves: was Jones cited because what she did was of great significance, or because her experiment was a flawed piece of work that requires replacement?

To make these functionalities a reality we need two things. First, we require a technology base that is capable of handling the complexity of electronic documents. In particular, PDF files are designed for rendering on a screen or on paper, not for being scrutinized by sophisticated electronic agents, and so we need to extract the text flow from these in a form that is not sullied by arbitrary aspects of physical rendering such as pagination, headers and footers, and floating material such as figures and tables. Second, we need to integrate and apply natural language processing tools and techniques developed for the text analytics tasks of named entity recognition (so we can identify the wide variety of citation forms that appear in academic works), cross-document name matching (so we can correlate citations across different works), information extraction (so we can develop more a refined analysis of what author A says about author B), and sentiment analysis (so we can determine whether what A says is positive or negative).

No one to date has brought these technologies together for this purpose. The aim of the project proposed here is to integrate these various components in a proof of concept that demonstrates what is possible. We expect this to result in one or more high-quality publications within the relevant communities (natural language processing; digital libraries; bibliometrics).

2 RESEARCH METHODOLOGY AND PLAN

2.1 Approach

The objective is to develop a proof-of-concept citation linkage tool over a period of 12 months. The tool will demonstrate that, for a given document collection, we are able to (a) identify the citations within the body text of each document, (b) link these to the references in the references list, (c) link these references to canonical metadata about the referred-to documents, and (d) in those cases where the referred-to documents exist on the web, directly connect these documents to produce a true web of science. This provides a platform on top of which ever-richer citation-based services can be built; the platform will be made publicly available so that others can develop components that can be plugged in.

In the Centre for Language Technology, we have already put some of the key elements that are required in place. The PhD work of a student in the group, Brett Powley, has provided some of the basic infrastructure required for processing raw electronic documents in PDF form [Powley and Dale 2007]; through our involvement with the Association for Computational Linguistics, the peak international body in the field, we act as a mirror site for the web-hosted ACL's Anthology Research Corpus, a digital collection of some 10000 papers and articles in the area of natural language processing which we intend to use as the base document collection for the proof of concept [Bird et al 2008]; and we have extensive relevant experience in various aspects of named entity recognition and information extraction from our work for the Capital Markets Cooperative Research Centre [e.g, Dale et al 2004] and the Defence Science Technology Organisation [e.g, Mazur and Dale 2007]. The project proposed here integrates and extends these components in a coherent platform. The six steps involved are shown in the task plan below, and elaborated upon in the following; time estimates are elapsed time, based on a part-time research programmer working 50% time.

Task	Description	J	A	S	O	N	D	J	F	M	A	M	J
1	Text Extraction	■	■										
2	Logical Structure Extraction			■	■								
3	Reference Parsing					■	■						
4	Citation Finding							■	■				
5	Web Search									■	■		
6	Metadata Matching											■	■

2.2 Task Plan

Task 1: Text Extraction

This involves extracting the text stream from a text in such a way that artifacts of physical rendering are removed. In our CMCRC work, we have developed a basic tool for doing this in other kinds of documents; this needs to be refined and adapted to work with the ACL ARC Anthology. [2 months]

Task 2: Document Structure Extraction

In order to process a document we have to break it into its parts: we want to separate the reference list or bibliography from the rest of the document, and to separate out the header data (such as title, author and affiliation data). A current undergraduate project demonstrates that good baseline performance can be achieved on with relatively straightforward methods; we will use this work as a starting point, fine tuning and extending it to deal with the target data collection. [2 months]

Task 3: Reference Parsing

Segmenting a bibliographic reference into its constituent fields is an important enabling step for metadata matching. There is considerable prior research we can lean on here (see, e.g., Besagni et al [2003]), so the focus will be on tailoring existing techniques to our specific needs. [2 months]

Task 4: Citation Finding

In order to be able to determine what a document says about a cited work, we have to extract the citing sentence and match the embedded citation to the corresponding entry in the reference list. Here we will integrate some of the work from Powley's PhD work [Powley and Dale 2007]. [2 months]

Task 5: Web Search Interface

The steps above deal with processing a given document; we then have to link this document to those it cites. This means finding either candidate cited documents or their metadata in web-based repositories. Leveraging the connection with Microsoft, we would hopeto make use of an API to Live Search search engine for this purpose. [2 months]

Task 6: Metadata Matching

Finally, having identified a set of candidate cited documents for a given citation, we need to develop fuzzy matching techniques to determine which is the best match. This task shares much in common with what is known as the cross-document co-reference matching task [Gooi and Allan 2004], where the aim is to determine whether two similar or identical company or person names in separate documents refer to the same entity. We have explored aspects of this problem in work with the DSTO, so our aim here is to leverage this by exploring the adaptation and extension the findings from these areas for the purposes of reference metadata matching. [2 months]

REFERENCES

D Besagni, A Belaid, and N Benet [2003] A segmentation method for bibliographic references by contextual tagging of fields. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 84–88 vol.1.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev and Yee Fan Tan [2008] The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morocco.

R Dale [1989] Computer-based Editorial Aids. Pages 12-20 in *Recent Developments and Applications of Natural Language Understanding*, edited by Jeremy Peckham. Kogan Page, London.

R Dale, R Calvo and M Tilbrook [2004] Key Element Summarisation: Extracting Information from Company Announcements. In *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence*, 7th-10th December 2004, Cairns, Queensland, Australia.

E Garfield [1955] Citation indexes to science: a new dimension in documentation through association of ideas. *Science* 122:108-11.

C H Gooi and J Allan [2004] Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2004)*.

J E Hirsch [2005] An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences*, 102(46):16569-16572.

C Matheson and R Dale [1993] BibEdit: A knowledge-based Copy Editing Tool for Bibliographic Information. In E S Atwell (ed) *Knowledge at Work in Universities: Second Annual Conference of the Higher Education Funding Councils' Knowledge Based Systems Initiative*. Cambridge.

P Mazur and R Dale [2007] Handling conjunctions in named entities. *Lingvisticae Investigationes*, 30:1.

B Powley and R Dale [2007] Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification. In *Proceedings of RIAO 2007: the 8th Conference on Large-Scale Semantic Access to Content*, 30 May 30 to 1 June 2007, Pittsburgh, PA, USA.