CP4101 HYP Interim Report
Citation Provenance*

Heng Low Wee
U096901R

April 1, 2012

1 Introduction

To cite previously published scientific papers is a necessary and common practice among researchers. It gives credit and acknowledgement to original ideas, and to those researchers who did significant amount of work in that particular field of research, and more importantly, to uphold intellectual property. A reader of such research papers often encounter these citations made by the authors in various sentences throughout the paper. Often enough, if a reader wishes to gain a better understanding of the current context, it is necessary to follow through these citations and read up on these cited papers.

1.1 The Problem

Readers would read up on the cited papers to gain a better insight on the current topic of discussion. However, as frequent readers might find, most citations are only *mentions*. They do not directly refer to some particular section of the cited paper, for example, to make reference to the evaluation results made by the authors of the cited paper. Instead, they are general citations. These citations are important. However since it is not immediately clear where the cited information is from, a reader has to invest additional time to read through the entire cited paper before being able to find out what the critical information is. The difficulty is increased when the cited document is a research journal, when the amount of content is comparably huge. The author might cite specifically using the journal's volume number, issue and page number, yet the problem remains. This makes little contribution to a reader's understanding of the current topic, and it is also time-consuming.

1.2 Our Goal

In general, we wish to improve the reading experience of scientific & research documents, from the various fields of research. We want to provide information about the provenance of a citation made in a paper. Readers will be informed of where exactly the cited information is from in the cited paper. For instance, if a citation is made to refer

*Project Code: H079820

to the evaluation results computed by the authors of the referenced paper, the reader would be *guided* to the particular section of that paper. By *guiding*, we refer to having the section with the critical information highlighted, and so the reader can quickly read up on it, before resuming on the current paper. If the citation was a general citation, the reader will possibly be referred to the Abstract section of the paper.

With the above in mind, we investigate these related works.

2 Literature Review

2.1 Helping Users Make Relevance Judgements About Cited Documents

3 Progress So Far

3.1 Analysing The Problem

All citations could be classified into 2 categories: General, and Specific citations. A general citation is essentially a *mention*, and the author is making reference to a paper as a whole, usually because the cited paper is of the same field of study. A specific citation is more direct. It refers to a particular section, paragraph, or even line of the cited paper. For example, this often happens in the case where the author wishes to refer to some development by the cited authors, or to make reference to the evaluation results in order to compare performance. In the field of Computer Science, it is often authors make reference to some particular computer system, or some computing algorithms, and to compare performance on speed & accuracy on solving problems. Our problem can be reduced to determining whether a citation is General or Specific. If a citation is general, the reader can be directed to the Abstract section of the cited paper. If a citation is specific, the reader can be directed to that specific paragraph respectively.

3.2 Modelling The Problem

The main question in this problem is *Where does this information come from in that cited paper?* We can say that we already know the list of references cited by a paper. We have this information from a paper's References or Bibliography section. Conveniently, we also know that any sentence that contains a citation must refer to one in the list of references. For this project, we work on based on the fact that we are already made known of the list of references.

In web search engines, an user enters a search query, and a search engine would use this query to search within its search domain – millions of web pages – and then display the best matching web pages as compared to the search query. That would be equivalent to having a search query for an entire corpus of research papers. Our problem can also be modeled as a searching problem, but a reduced version as compared to web search engines.

Consider reading a paper, `A`. We know the citations made by `A`, and the cited papers are listed in the References section of `A`. From this our search domain for this paper would be the contents of the list of cited papers. We reduce the search domain further when we are investigating a particular citation in `A`, say `A` cite the paper `B`. Now, for this citation, the scope of search would be a sub-domain – contents of paper `B`. So instead of searching for the best matching document in the corpus, we are searching for the best matching *fragment* in `B`. For now we define our search accuracy to the level of a *fragment*, where a *fragment* consists of N sentences of the paper. The search query comes from `A`, the sentence in which the citation is located. The surrounding words to the citation would serve as the list of words in the search query.

3.3 Our Approach

3.3.1 Training Corpus

At this initial stage, we picked the ACL Anthology Reference Corpus[1] (ACL-ARC). The ACL-ARC consists of publications of the Computational Linguistics field. Note that in general, we wish to perform this citation provenance tasks on all publications from all fields of research. This corpus is chosen as a start, because it provides the interlink data that conveniently informs us of the citation-links between the papers in the corpus. For instance, a link like `X98-103==>X96-1049` says that the paper `X98-103` cites the other.

4 What's Next

References

[1] Zhong, Z., and Ng, H. T. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations* (Morristown, NJ, USA, 2010), ACL '10, Association for Computational Linguistics, pp. 78–83.

[1] http://acl-arc.comp.nus.edu.sg/