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. Readers would also be interested to know where the information is in the cited paper.

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

*Project Code: H079820

cited information is from in the cited paper. For instance, if a citation is made to refer 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.

We want to develop a reading tool for readers of scientific papers. This reading tool will be able to provide the reader with information about the cited papers, and depending on whether the citations are general or specific, the contents of the cited paper will highlighted accordingly, so that reader may make a quick reference to the cited paper and not interrupt the current reading.

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

2 Literature Review

2.1 Helping Users Make Relevance Judgements About Cited Documents

WORK IN PROGRESS

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 Modeling 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 k 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 Tackling The Problem

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 task 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, in the interlink data, a link like `X98-103==>X96-1049` says that the paper `X98-103` cites `X96-1049`.

3.3.2 Baseline

The initial idea was to naively compare the search query (a citing sentence from paper `A`) with the search domain (paper `B`) by term matching frequency. Paper `B` will be divided into overlapping fragments of 5 lines, and then each fragment would be compared with the search query. This idea was short-lived as common terms like *to*, *the*, *is*, *in* etc. would get high matching frequency, and the results would be definitely skew towards fragments with high number of such terms.

We extend the initial idea and adopt a common mechanism used by search engines, that is to compare the search query and domain for Cosine Similarity. In our problem, instead of comparing the query with a full document, we compare the query with fragments of the target paper. We first generated the vocabulary set using terms from the entire corpus. This is essential because during the comparison for cosine similarity, both vectors must have equal number of dimensions. Then it is necessary we adopt the `tf-idf`[2] weighting scheme in order to prevent the commons word affecting the comparison results. `tf` simply

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

[2] Term Frequency-Inverse Document Frequency

refers to *term frequency* in a document, while `idf` refers to the inverse of the *document frequency* of a term. The `df` of a term refers to the number of documents in the corpus that contains that term.

$$\texttt{tf-idf}_{t,d} = \texttt{tf}_{t,d} \times \log \frac{N}{\texttt{df}_t}, N = \text{no. of documents in corpus} \quad (1)$$

By using this weighting scheme, we are saying that the terms that appear rarely – with high `tf-idf` value – are *important*, whereas common terms that occur often, like *the* and *to*, are not. This is feasible, since in general, terms that appear often are less important, and we do not want non-important terms to affect our results.

We implemented the algorithms for computing `tf-idf` and cosine similarity ourselves instead of using the NLTK package. Partly because of performance. We are able to pre-prepare information such as V, our vocabulary set, `tf` for each term with respect to each document, and `df` for each term in V. In addition, we performed word stemming on the entire corpus. So instead of having the NLTK tool computing these information each time a query is ran, we speed up the computation process by removing this repetition, since these information remains constant for the same corpus.

The challenge surfaces as one might see that this solution is not context-sensitive, and would fail to return meaningful results on having paraphrased search queries. We seek for alternatives in Section 4.

4 What's Next

4.1 Gold Standard Annotations

Using the same corpus, we plan to gather gold standard annotations for the research papers. We will annotate for each cite as stated in the interlink data, whether each cite is a general or specific cite. We are particularly interested in specific citations, for that is the goal of our project – to determine citation provenance and if possible, show exactly where the information is from. Annotations for specific citations should be done down to the level of line numbers of the research paper. For instance, if the cited information appears between line 23 and 28, the annotation should capture it as `23-28`.

Collecting gold standard annotations is essential for the project. First, it is for evaluating the performance of our proposed solution. The current state of the corpus do not provide such information. Second, in order to introduce a machine learner component, we would require gold standard annotations to *train* on.

4.2 Paraphrasing

It is a common practice for authors, when citing another paper, to paraphrase the original sentences, often to achieve greater clarity for the current context. Sure enough, the terms in the paraphrase would be significantly different from the original sentence written in the cited paper. The baseline we used would simply treat that as a mis-match. Therefore, we have to explore introducing NLP techniques and mechanisms in order to handle paraphrasing.

References