Mining Opinion Polarity Relations of Citations

Scott S. Piao, Sophia Ananiadou, Yoshimasa Tsuruoka, Yutaka Sasaki and John McNaught
School of Computer Science
University of Manchester
United Kingdom

November 24 2006

1   Introduction

Opinion mining has been receiving increasing attention recently, and various approaches have been suggested for mining sentiment information, such as mining attitudes or opinions about a topic or product etc. However, as far as we know, little work has been reported on citation opinion mining (COM). By COM, we refer to the process of identifying authors opinions towards the works they cite, such as positive/negative attitudes or approval/disapproval. We contend that such information is useful for semantic information retrieval and text mining, particularly for users who wish to search for documents taking a positive or negative stance towards a specific previous work. In this paper, we propose a system which is based on existing semantic lexical resources and NLP tools, aiming to create a network of opinion polarity relations between documents and citations. This is a web-based system which allows users to access the citations collected from documents and retrieve those documents linked to each of the citations with different opinion polarity relations, namely approval, neutral or disapproval relations. Various approaches will be tested including detecting semantic orientation of subjective words in the context of citations and machine learning using manually annotated data. In particular, we will explore the use of semantic lexicons for this task.

2   Related work

Recently, opinion mining has emerged as an important research area cutting across a number of topics such as information retrieval, NLP and text mining etc. Research in this area covers several topics including the learning of semantic orientation of words/terms, sentiment analysis of documents, analysis of opinions and attitudes towards certain topics or products etc. Previous works closely related to our current work include Wilbur et al. [8], Teuful et al. [7] and Kim and Hovy [3].

Wilbur et al. (2006) suggest that factual information mining is not sufficient. There is a range of non-factual qualitative information which affects the reliability and validity of factual information. For example, when authors discuss individual facts, some may be speculative, e.g. we assume that X can be linked to Y while others are assertive, e.g. our result shows X is linked to Y. In their annotation scheme, they identified five qualitative dimensions that characterise a broad range of scientific sentences: Focus (scientific vs. general), polarity (positive vs. negative), certainty, evidence, direction/trend (increase / decrease in certain measurement).

Kim and Hovy [3] proposed a method for identifying opinion with its holder and topic in news media. They specifically address the issue of how an opinion holder and a topic are semantically related to an opinion bearing word in a sentence. Their method proceeds in three phases: a) identifying an opinion bearing word, b) labelling semantic roles related to the word and c) finding the holder and the topic of the opinion word among the labelled semantic roles. This is a similar

method to the one proposed in this paper. However, they are different in that we are trying to determine the author's attitude of approval/disapproval towards the cited papers in the given article.

Various semantically annotated lexical resources have been developed and made available [2]. A major semantic lexicon in early days is the General Inquirer [6] (http://www.wjh.harvard.edu/ inquirer/), in which 1,915 words and 2,291 words are tagged as positive and negative respectively. Recently, more large-scale semantic lexicons have been constructed, including WordNet (http://wordnet.princeton.edu/), FrameNet (http://framenet.icsi. berkeley.edu/), Lancaster UCREL semantic lexicon [5] and SentiWordNet [1] etc. We will explore various lexical resources for our work.

While our work is based on the previous works, it has a different aim, i.e. to mine the opinion polarity relations between documents (we refer to academic papers in this case), and their citations. As far as we know, little research has been carried out on this topic.

3 Opinion mining of polarity relations between documents and their citations

As we mentioned in the previous section, most of the previous works in the area of opinion analysis/mining have focussed on authors'/holders' opinions towards facts and topics. However, we find that opinion polarity relations between documents and their citations, such as authors attitudes of approval/disapproval towards the works they cite, can be useful for semantic information retrieval and text mining. Our assumption stems from the observation that, besides using citations as a background for their current work, authors very often take positive/negative stances towards the works of others which they cite. For example, an author may approve of a previous work and cite it as a supportive evidence for his/her own statements or points, or cite it as a negative example to be criticized in his/her article, as shown in the following sample (from PLOS http://www.plos.org/).

> The PSI-BLAST program [<xref ref-type="bibr" rid="pbio-0030181-b18">18</xref>, <xref ref-type="bibr" rid="pbio-0030181-b45">45</xref>] *is much more sensitive than* a regular BLAST search due to the use of PSSM.

For someone who wishes to search for documents expressing approval or disapproval of a specific previous work from a large collection of documents, such opinion polarity relational information between a given document and those citing it can be useful. Obviously, it is not practical to manually analyze such information for a large amount of documents, and hence we need an automatic means of miningopinion polarity relational information.

A system model is designed for this purpose, as illustrated in Figure 1. As shown in the figure, our approach employs semantic lexicons and NLP tools to collect and map citations to form a network knit by opinion polarity relations. This approach proceeds as follows:

- Collect cited papers from a collection of academic articles (PLOS journal papers are used as sample data).

- Extract sentences containing citations of the papers from the collection.

- Link each cited work to the papers containing the citation via the sentences extracted above.

- Determine the opinion orientation of the subjective words in the context of the citations (approval, neutral, disapproval) using semantic lexical resources.

- Map the subjective words and their sentiment orientation to the citations using syntactic parse information of the sentences with the Enju parser [4] (http://www-tsujii.is.s.u-tokyo.ac.jp/enju/), which creates predicate-arguments links.
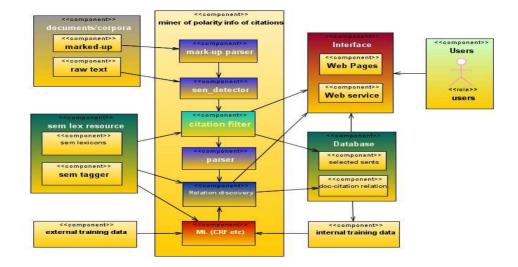
Figure 1: System for mining opinion polarity relations of citations

Figure 2: Citation frequency list linked to citation distributions

- Based on the sentiment analysis of the sentences, induce opinion polarity relations between the papers and citations.

- Create a network of papers knit by the opinion relations and save it in a database.

- Provide a search interface for the system.

Semantic lexical resources form a core component of this system which provides clues for subjective word polarity orientation. As a starting point, we will test Lancaster's UCREL semantic lexicon (http://www.comp.lancs.ac.uk/ucrel) and the General Inquirer lexicon. However, we will including more lexical resources in future.

This system is still in its early stages of development, but a prototype implementing a number of the steps has been developed. Currently, the prototype is capable of collecting cited papers with citation frequencies (in terms of number of documents) from a given collection of PLOS articles and extracting a citation distribution list for each of the cited papers i.e. a list of sentences containing the given citation. Figure 2 shows a snapshot displaying a citation frequency list and a citation distribution list. In this figure, the left-hand webpage displays the list of cited papers extracted from a collection of PLOS articles, and the right-hand one displays the sentences containing the citations found in the paper collection.

After completion, this system would enable users, for a given document, to retrieve other documents which cite it with certain sentiment orientation. For example, the user can retrieve

academic papers that approve of an earlier work by following the positive sentiment links of citations. The documents and sentences will be classified into the categories of approval, neutral and disapproval relations with respect to the paper they cite. All of the information will be accessible via a web interface, as illustrated by Figure 1.

4   Conclusion

In this paper, we briefly described a system model of citation opinion polarity mining which is under development. Our work is motivated by our assumption that such opinion polarity relations between documents and their citations are useful information for advanced semantic information retrieval and text mining. Although much research has been conducted in opinion mining area, as far as we know, little work has been done on opinion mining regarding citations. If an efficient system can be developed for mining opinion polarity relations of citations, it will be helpful in enhancing current text mining systems which currently are generally based on factual information. In particular, considering the huge amount of biomedical literature, this work will bring benefits to biomedical text mining tools for further refining search results.

Acknowledgements

References

[1] Andrea Esuli and Fabrizio Sebastiani. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 193–200, Trento, Italy, 2006.

[2] N. Ide. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006, February 19-25, 2006, Proceedingse*, volume 3878 of *Lecture Notes in Computer Science*, Mexico City, Mexico,, 2006. Springer.

[3] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia, 2006.

[4] Yusuke Miyao and Jun'ichi Tsujii. Probabilistic disambiguation models for wide-coverage hpsg parsing. In *Proceedings of ACL-2005*, pages 83–90, Ann Arbor, US, 2005.

[5] Scott Piao, Dawn Archerand Olga Mudraya, Paul Rayson, Roger Garside, Tony McEnery, and Andrew Wilson. A large semantic lexicon for corpus annotation. In *Proceedings from the Corpus Linguistics Conference Series on-line e-journal*, volume 1, Birmingham, UK, 2006.

[6] P.J. Stone, D.C. Dunphy, M.S. Smith, and D.M. Ogilvie. *The General Inquirer: A computer Approach to Content analysis.* MIT Press, Cambridge, MA, 1966.

[7] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of EMNLP-06*, Sydney, Australia, 2006.

[8] W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(356), 2006.