On Computing Text-based Similarity in Scientific Literature

Seok-Ho Yoon
Dept. of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791, Korea
bogely@hanyang.ac.kr

Sang-Wook Kim
Dept. of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791, Korea
wook@hanyang.ac.kr

Ji-Soo Kim
Dept. of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791, Korea
kimjisu29@hanyang.ac.kr

ABSTRACT

This paper addresses computing of similarity among papers using text-based measures. First, we analyze the accuracy of the similarities computed using different parts of a paper, and propose a method of *Keyword-Extension*, which is very useful when text information is incomplete.

Categories and Subject Descriptors: I.5.3 [Clustering] Similarity measures

General Terms: Measurement, Reliability

Keywords: Scientific Literature, Text-based Similarity Measure

1. INTRODUCTION

As the number of people who are interested in scientific literature grows, there have been a number of research efforts on this area. One of the most important issues is to compute similarities among papers because it is used as a basic component in several advanced functions such as clustering, recommendation, and ranking [1].

Previous similarity measures are categorized into two classes: text-based and link-based similarity measures. The text-based similarity measure considers the number of terms in common between two papers while the link-based similarity measure takes common citations among two papers into account. In this paper, we address the text-based measure to accurately compute the similarity among papers.

A paper is composed of three parts: the title, abstract, and body. The similarities between a pair of papers could be significantly different when computed by using different parts, since sets of terms in these parts differ. Thus, we need to understand which similarity reflects well the *actual* similarity between two papers and what weights should be assigned to the parts if we compute the final similarity by combining terms obtained from more than a part.

Typical services for literature retrieval such as CiteSeer, Google Scholar, and MS Libra provide text information on the title, abstract, and body of papers by crawling and parsing the original paper files. However, they do not provide the full text information on the body due to the copyright problem. The abstract information is also frequently missed due to the limitations of crawling and parsing. In summary, complete text information is unavailable in practice. This causes low accuracy of similarities computed by using text-based measures.

In this paper, we analyze the accuracy of the similarities obtained by using the three parts, and find good weights for the parts when combining terms obtained from more than a part to compute the final similarity. Also, we propose a method of *Keyword-Extension*, which is so useful in case text information is incomplete.

2. FINDING THE WEIGHTS

In this section, we present the accuracy of similarities computed by using the terms from each of the title, abstract, and body. For experiments, we used the scientific literature database crawled from Libra[1], which has 1,071,973 papers and 2,473,636 citations inside. We employed the vector model with TF-IDF for computing text-based similarities.

In order to evaluate the accuracy, we did the following. We selected twenty sub-chapters in a textbook of data mining [1] and extracted reference papers (124 papers) from each sub-chapter. For each reference paper (*query paper*), we retrieved its k most similar ones from all the reference papers by using the similarity computed with each part. Then, we compared the k papers with the reference papers in the sub-chapter that the query paper belongs to. We performed the same process with all the reference papers in the textbook except for those absent in our database [2].

Figure 1 shows the precision of the results obtained using different parts. The result using the abstract shows the best precision. It is surprising that using the body performs much worse than the others, even though the body provides a lot of information. Nevertheless, many of terms in the body seem to be unrelated to the main issue dealt with in the paper. Thus, the body is inappropriate to be used as features of the paper. On the other hand, while the title surely contains the only terms most related to the main issue of the paper, some important terms are easily missed due to its limited length. The result indicates that the abstract provides (relatively) sufficient and necessary terms that are important for representing the main issue of the paper, thereby outperforming the others.

We note, however, the result does not imply that the title and body are useless in computing text-based similarity. Thus, we examined the accuracy of similarities computed by combining the terms from multiple parts. For the body, we already observed it contributes to low accuracy in similarity computation because it contains a large number of unimportant terms, and thus excluded it in further investigation.

[1]http://academic.research.microsoft.com

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

We compared the accuracy of similarities with different weight combinations for the title and abstract. The result showed that, on average, the result with the combination of 0.3:0.7 showed the best precision, outperforming the result using only the abstract around 5%. This implies that the combination of terms from the title and abstract is meaningful for computing the similarity of papers accurately.

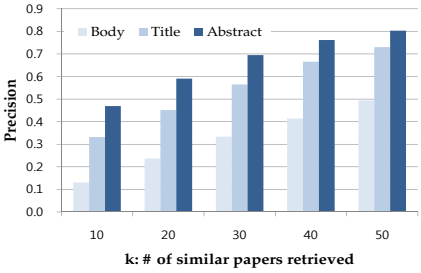

Figure 1: Accuracy of similarities using different parts.

3. KEYWORD EXTENSION

In Section 2, we verified that the abstract provides very important terms in computing text-based similarity among papers. However, abstracts are not always available in practice due to the limitations of crawling and parsing. For some papers, abstracts are partly available; For some other papers, abstracts are completely unavailable. These papers would suffer from insufficient terms related to their issue in similarity computation, which leads to low accuracy of similarities. Therefore, finding of additional good terms is necessary for improving the accuracy of similarities with which those papers are involved.

Authors of a paper cite such papers that are thought to be highly related to the main issue of the paper. Thus, if two papers are involved in a citation relationship, we can expect that they would have similar terms in their titles and abstracts. Based on this observation, we propose a method called *Keyword-Extension* that extends the term set of paper P by including the terms in titles and abstracts of all the papers that are in the citation relationship with paper P. For example, suppose paper B cites paper C and is also cited by paper A. As a term set of paper B, we will use not only the terms from paper B but also the terms from papers A and C. This simple extension could successfully solve the problem of incomplete terms.

We verified the effect of *Keyword-Extension* on the accuracy by comparing *Keyword-Extension* using the incomplete abstract and the two methods using the incomplete and complete abstract without the application of *Keyword-Extension*. We simply set the ratio of weights of terms from the original paper, cited papers, and citing papers as 1:1:1. Also, we set the ratio of weights of terms from the title and abstract as 0.3:0.7.

Figure 2 shows the result. We see that the accuracy of incomplete abstract *with Keyword-Extension* increases dramatically up to 3.3 times compared to that of incomplete abstract *without Keyword-Extension*. Moreover, our *Keyword-Extension* shows almost the same accuracy as the result using the *complete* abstract without *Keyword-Extension*. The result indicates that *Keyword-Extension* successfully makes up for the terms of the papers with incomplete information.

We also compare our *Keyword-Extension* with typical link-based similarity measures of Bibliographic coupling, Co-citation, and Amsler. Figure 3 shows the result. We observe that our *Keyword-Extension* significantly outperforms the link-based similarity measures.

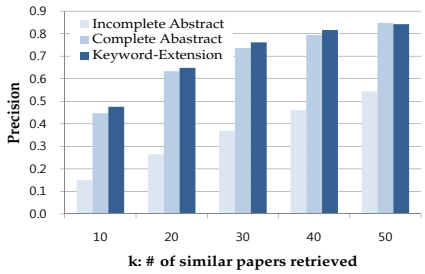

Figure 2: Accuracy of Keyword-Extension and two methods without Keyword-Extension.

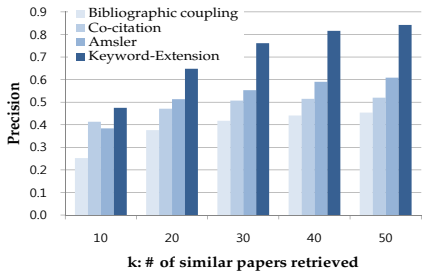

Figure 3: Accuracy of Keyword-Extension and link-based similarity measures.

4. CONCLUSIONS

We analyzed the accuracy of the similarities using different parts in a paper, suggested good ratio of weights for the title and abstract. Also, we proposed *Keyword-Extension*, which is so useful in case text information is incomplete. Via a series of experiments, we verified the effectiveness of *Keyword-Extension*.

5. ACKNOWLEDGMENTS

This work was supported by NHN Corp and by NRF (Grant No. 2008-0061006). Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

6. ADDITIONAL AUTHORS

Additional author: Won-Seok Hwang (Dept. of Electronics and Computer Engineering, Hanyang University, Seoul, 133-791, Korea, email: `hws23@agape.hanyang.ac.kr`)

7. REFERENCES

[1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques (2nd Edition)*. Morgan Kaufmann, San Francisco, 2006.

[2] S. Yoon, S. Kim, and S. Park. A link-based similarity measure for scientific literature. In *Proc. of Int'l. Conf. on World Wide Web*, pages 1213–1214, April 2010.