Improving Full Text Search With Text Mining Tools

Scott Piao, Brian Rea, John McNaught, and Sophia Ananiadou

National Centre for Text Mining
School of Computer Science
The University of Manchester
Manchester UK
{scott.piao,brian.rea,john.mcnaught,sophia.ananiadou}@manchester.ac.uk

Key words: Information Retrieval, Full Text Search, Term extraction, Termine, Document clustering, Natural Language processing

Today, academic researchers face a flood of information. Full text search provides an important way of finding useful information from mountains of publications, but it generally suffers from low precision, or low quality of document retrieval. A full text search algorithm typically examines every word in a given text, trying to find the query words. Unfortunately, many words in natural language are polysemous, and thus many documents retrieved using this approach are irrelevant to actual search queries. A variety of techniques have been used to mitigate this problem, such as controlled vocabularies, keywords (manually provided by authors or indexers) and phrase search, however, these techniques have their own limitations, such as low coverage of controlled vocabularies.

In our work, we attempt to improve full text search with text mining tools. we have developed a document search engine[1] based on Apache Lucene [2] at the UK National Centre for Text Mining (NaCTeM)[3] for the INTUTE Repository Search (IRS) Project [4], in which we attempt to improve the full text search by using text mining tools, particularly terminology extraction and document clustering tools.

We observe that one of the main reasons for the low precision of full text search algorithms lies in the fact that they match the query words indiscriminately against all words in a document, whether they reflect the topic of the

[1] It has been deployed as a web demonstrator at url: `http://www.nactem.ac.uk/nactem_irs/doc_search`

[2] Apache Lucene is a high-performance, full-featured text search engine library written in Java. See website: `http://lucene.apache.org`

[3] For information about NaCTeM, see website: `http://www.nactem.ac.uk`

[4] This project is funded by UK JISC (`http://www.jisc.ac.uk`). For further details, see `http://www.nactem.ac.uk/intute`

document or not. We address this problem by constraining the search to those words or terms representative of the documents. We use a terminology extraction tool, named Termine[5], which has been developed for text mining in NaCTeM. Employing the C-value termhood metric, it identifies candidate terms which are representative of a given document. This tool was integrated in the document indexing process, such that the document contents are represented by the term sets extracted by the tool. Given a query, the search is carried out on the terms. In other words, our search is for the documents in which the query occurs as a key term or part of such terms. For example, if users use query "monthly", our algorithm would only retrieve those documents in which the query word is a domain concept term or part of such terms, such as "monthly breast self-examination". In addition to improving the quality of full text retrieval, the extracted concept terms are also used to address the information overlook problem by guiding users to navigate those documents which share domain concepts, helping users to discover the underlying semantic network of documents.

Our initial evaluation of the terminology tool shows that it is effective for reducing the number of irrelevant documents retrieved while increasing the porportion of "interesting" documents. For example, when we tried the query "help", compared with standard full text search, the term-based search reduced the number of retrieved documents from 3,470 to 135. Meanwhile, when we checked the top ten documents retrieved, the number of "interesting" documents, in which this word is used in domain terms such as *self-help manual*, increased from 2 to 8.

Although our tools need further improvement and evaluation, our experiment demonstrates that, compared with the baseline full text approach, the term-based approach is capable of filtering out a significant amount of noise documents while increasing the number of "interesting" documents aming the top documents retrieved.

Furthermore, we also incorporated a document clustering package, Carrot2[6], which employs the LINGO algorithm to provide the functionality that helps overcome the information overlook problem. We use this tool to cluster top 200 retrieved documents on the fly, grouping documents under human readable labels. For example, for a query "environment", our search engine cluster the retrieved documents into groups under labels such as *Learning Environment*, *Virtual Research*, *Software Design* etc, which reflect different aspects from which the topic "environment" is addressed.

Our tool will be further improved with the development of a more efficient document clustering tool based on Termine and a customised visualisation tool.

[5] See: Frantzi, Katerina, Ananiadou, Sophia, Mima, Hideki: Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal of Digital librarie, 3(2), pp. 117132 (2000)

[6] See Carrot2 website: `http://www.carrot2.org`