The MultiText Retrieval System

Gordon V. Cormack[1] Charles L. A. Clarke[2] Christopher R. Palmer[3] Robert C. Good[1]
[1]Department of Computer Science, University of Waterloo
[2]Department of Electrical and Computer Engineering, University of Toronto
[3]School of Computer Science, Carnegie Mellon University
mt@plg.uwaterloo.ca
http://multitext.uwaterloo.ca

The ultimate aim of MultiText is to develop techniques to provide access by content for a significant fraction of all text available electronically. To this end we consider only algorithms and data structures that can be implemented efficiently on cheap commodity hardware, and that can be scaled without performance penalty to arbitrarily large collections by the mere addition of hardware. Our objectives oblige us to consider in addition the issues of multiple concurrent users, on-line update, continuous availability, and distributed data.

The data model for MultiText is unique in that the collection is not divided *a priori* into units such as documents, paragraphs, sentences, or lines. Rather, the text is organized as a linear stream of words, and structural components like those named above are delimited by markup symbols [2]. As far as the retrieval system is concerned, markup symbols are simply words and may be used as such in a query. An important aspect of this data organization is that the system imposes no particular schema on the data — data with many different formats can be stored, searched, and retrieved within the same collection [3].

Queries in MultiText are satisfied by substrings or passages within the document, rather than by documents, paragraphs, lines, etc., which have no particular meaning to the system. Specifically, the system finds the *shortest* passages that satisfy a particular query. That is, a passage is returned only if it satisfies the query and does not contain a shorter passage that also satisfies the query. This approach can be considered to yield the most general solution to the query [1]. The MultiText query language, GCL, expresses boolean queries as well as containment relationships (`contained in`, `containing`, `not contained in`, and `not containing`). For example, one could express directly in GCL the query: ***Find the titles of documents containing all of "Cormack", "Clarke"", "Palmer" and "Good" in the author field.***

Classical information retrieval is effected in MultiText using a few search terms and the ranking technique *Cover Density Ranking* [4]. The objective here is to find the documents most likely relevant to the user's information need as expressed by the search terms. To this end we order the documents in the following way: First, documents containing more of the search terms are considered more likely to be relevant than those containing fewer. (Documents are simply passages delimited by markup like <doc> and </doc>.) Second, within documents containing the same number of search terms, those containing these terms within a shorter passage (i.e. closer together) are considered more likely to be relevant. This simple technique provides excellent efficiency, and very good precision and recall as measured by traditional information retrieval standards [5].

At SIGIR 99 we demonstrate the MultiText retrieval system using the 100 GB TREC-7 Very Large Corpus [6] loaded onto a pair of single-processor Pentium II - 350 workstations with a total purchase price of approximately $5K.

SIGIR '99 8/99 Berkley, CA, USA
© 1999 ACM 1-58113-096-1/99/0007...$5.00

References

[1] Clarke, C. L. A., and Cormack, G. V. Shortest substring retrieval and ranking. *ACM Transactions on Information Systems* (1999). To appear.

[2] Clarke, C. L. A., Cormack, G. V., and Burkowski, F. J. An algebra for structured text search and a framework for its implementation. *The Computer Journal 38*, 1 (1995), 43–56.

[3] Clarke, C. L. A., Cormack, G. V., and Burkowski, F. J. Schema-independent retrieval from heterogeneous structured text. In *Fourth Annual Symposium on Document Analysis and Information Retrieval* (Las Vegas, Nevada, April 1995), pp. 279–289.

[4] Clarke, C. L. A., Cormack, G. V., and Tudhope, E. A. Relevance ranking for one to three term queries. In *Fifth RIAO Conference* (Montreal, June 1997), pp. 388–400. A version of this paper will appear in *Information Processing and Management*, 1999.

[5] Cormack, G. V., Clarke, C. L. A., Palmer, C. R., and To, S. S.-L. Passage based refinement. In *Sixth Text REtrieval Conference (TREC-6)* (Gaithersburg, Maryland, November 1997), National Institute of Standards and Technology (NIST), United States Department of Commerce, pp. 303–319.

[6] Cormack, G. V., Palmer, C. R., van Biesbrouck, M., and Clarke, C. L. A. Deriving very short queries for high precision and recall. In *Seventh Text REtrieval Conference (TREC-7)* (Gaithersburg, Maryland, November 1998), National Institute of Standards and Technology (NIST), United States Department of Commerce.