An Overview of MultiText

Charles L. A. Clarke[2] Gordon V. Cormack[1] Christopher R. Palmer[1]

[1] Department of Computer Science, University of Waterloo, Canada
[2] Department of Electrical and Computer Engineering, University of Toronto, Canada
mt@plg.uwaterloo.ca http://multitext.uwaterloo.ca

The research focus of the MultiText project is the development and prototyping of scalable technologies for distributed information retrieval systems. The MultiText system is based on the network-of-workstations architecture shown in figure 1. The system is composed of several major components: The *index engines* maintain the index file structures and provide search capabilities. The *text servers* are specialized by document type and provide retrieval capabilities for arbitrary text passages specified at the word level. Finally, the *marshaller/dispatcher* interacts with clients and coordinates query and update activities.

Research issues are addressed in the context of this distributed architecture. Issues of concern to the MultiText Project include data distribution, load balancing, fast update, compression, fault tolerance, document structure, relevance ranking, and user interaction. Support for document structure is a particular feature of the MultiText system. The system can support multiple document formats within a single integrated database and provide specific support for structure inherent in each document type. The MultiText query language, GCL, provides facilities for directly referencing document structure and allows queries to reference equivalent structural elements across differently formatted documents.

Ranking in the MultiText system is based on passage retrieval, with the score of a passage based on its length and the score of a document based on the score of the passages contained within it. As well as ranking full documents, the method allows ranking of arbitrary document components. Scores do not depend on collection-wide statistics, making the ranking method particularly suitable for use in a dynamic distributed environment.

In order to implement the structural retrieval capabilities of GCL, the MultiText system supports a full positional index using inverted list file structures. Text in the database is concatenated into a single term sequence, essentially one large document, and is addressed by term position. Markup is used to identify documents and other structural elements at query time. For each specific term (e.g. "the") the index stores a sorted list of the positions in the sequence where the term occurs.

The index provides two basic access operations. Given a specific term and a position in the term sequence, one operation returns the location of the first occurrence of the term after the specified position; the other returns the location of the last occurrence before the specified position. The GCL query language is implemented in terms of these basic operations. The index structures used by MultiText efficiently support the operations, minimizing the number of disk accesses and allowing large portions of the index lists to be skipped during query processing.

On a network of four inexpensive PC's, costing less than US$10,000, the current version of MultiText can search an index for 100GB of text and return the top 20 documents in less than a second on average, with good retrieval effectiveness. We continue to work to improve performance. Strategies for prefetching and managing multiple query streams are of particular interest.

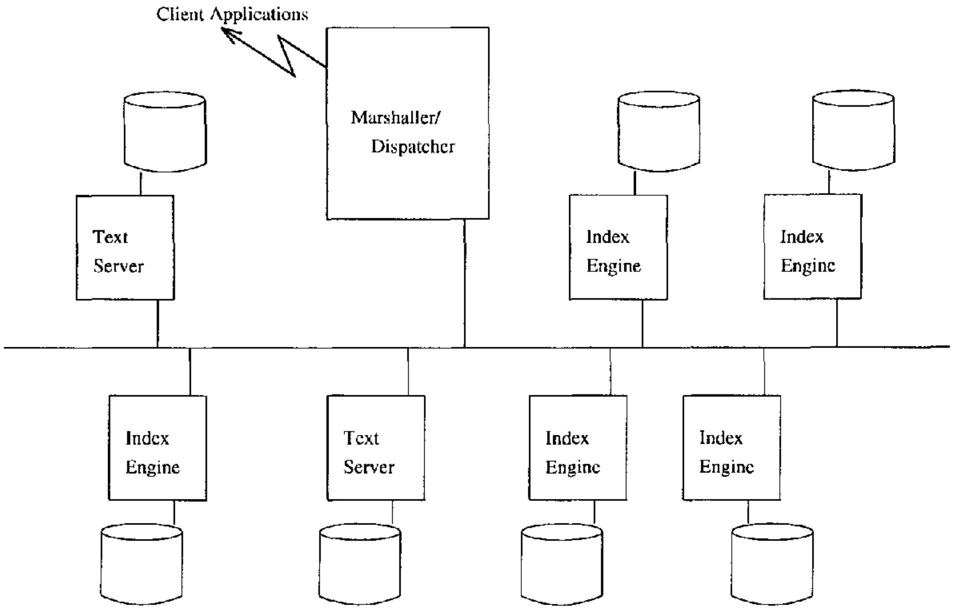

Figure 1: Architecture of the MultiText system

MultiText Bibliography

C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. Fast inverted indexes with on-line update. Technical Report CS-94-40, University of Waterloo Computer Science Department, November 1994.

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

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

C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. Shortest substring ranking. In *Fourth Text REtrieval Conference (TREC-4)*, pages 295–304, Gaithersburg, Maryland, November 1995.

R. C. Good, G. V. Cormack, C. L. A. Clarke, and D. J. Taylor. A robust storage systems architecture. In *Eighth International Conference on Computing and Information*, June 1996.

C. L. A. Clarke and G. V. Cormack. Interactive substring retrieval. In *Fifth Text REtrieval Conference (TREC-5)*, Gaithersburg, Maryland, November 1996.

C. L. A. Clarke, G. V. Cormack, and E. A. Tudhope. Relevance ranking for one to three term queries. In *Fifth RIAO Conference*, pages 388–400, Montreal, June 1997.

G. V. Cormack, C. L. A. Clarke, C. R. Palmer, and S. S. L. To. Passage-based refinement. In *Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, Maryland, November 1997.