Text Retrieval and Routing Techniques Based on an Inference Net

W. Bruce Croft, Principal Investigator

Department of Computer Science
University of Massachusetts
Amherst, MA 01002

PROJECT GOALS

The TIPSTER detection project at the University of Massachusetts is focusing on information retrieval and routing techniques for large, full-text databases, including Japanese. The project approach is to use improved representations of text and information needs in the framework of a probabilistic inference net model of retrieval.

In this project, retrieval (and routing) is viewed as a probabilistic inference process which "compares" text representations based on different forms of linguistic and statistical evidence to representations of information needs based on similar evidence from natural language queries and user interaction. New techniques for learning (relevance feedback) and extracting term relationships from text are also being studied.

Some of the specific research issues we are addressing are morphological analysis in English and Japanese, word sense disambiguation in English, the use of phrases and other syntactic structure in English and Japanese, the use of special purpose recognizers in representing documents and queries, analyzing natural language queries to build structured representations of information needs, learning techniques appropriate for routing and structured queries, probability estimation techniques for indexing, and techniques for automatically building and using thesauri. A retrieval system based on the inference net approach, INQUERY, has been built for studying these issues and distribution.

RECENT RESULTS

The results generated in the TIPSTER and TREC evaluations have been generally successful in that the probabilistic approach implemented in INQUERY has achieved the highest average levels of precision. We have also found that capturing a good description of the information need, either automatically or with some simple user input, has a significant impact on effectiveness. In particular, describing paragraph-level structure between concepts has been effective, whereas phrase-level structure has (somewhat surprisingly) only resulted in small improvements. Special purpose recognizers, such as for company names, have resulted in small improvements. Early experiments with automatic expansion of the queries using associated words found through analysis of the corpus have begun to show promise. Relevance feedback techniques used for the routing situation where there are many previous relevance judgements have advanced to the point where the average performance of the fully automatically generated queries can compete with the best manually-adjusted queries. With regard to Japanese, we have found that retrieval using character-based indexing of documents combined with word-based segmentation of the queries results in performance that is as good as word-based indexing of the documents. This is significant since character-based indexing is much more efficient than current word segmenters.