PIRCS: a Network-Based Document Routing and Retrieval System

K.L. Kwok

Computer Science Department
Queens College, City University of New York
Flushing, NY 11367

PROJECT GOALS

Our objective is to enhance the effectiveness and efficiency of ad hoc and routing retrieval for large scale textbases. Effective retrieval means ranking relevant answer documents of a user's information need high on the output list. Our text processing and retrieval system PIRCS (Probabilisitc Indexing and Retrieval -Components- System) handles English text in a domain independent fashion, and is based on a probabilistic model but extended with the concept of document components as discussed in our last year's site report. Our focus for enhancing effectiveness remains on three areas: 1) improvements on document representation; 2) combination of retrieval algorithms; and 3) network implementation with learning capabilities. Using representation with more restricted contexts such as phrases or subdocument units help to decrease ambiguity in both retrieval and learning. Combining evidences from different retrieval algorithms is known to improve results. Viewing retrieval in a network helps to implement query-focused and document-focused retrieval and feedback, as well as query expansion.

Efficiency of retrieval concerns time, space and cost issues of a system. These become important as the data one deals with grows larger and larger. Our focus for efficiency is to reduce time and space requirements in our system without sacrificing flexibility, robustness and our achieved retrieval effectiveness.

RECENT RESULTS

During 1993, we participated in TREC2 handling the full 2GB textbase provided. We redesign our system in two aspects: 1) on-demand network creation for retrieval and learning - this eliminates full 'inverted file' creation saving space and reducing 'dead time' between a collection is acquired and made searchable, and provides for fast learning capability; 2) 'subcollections within a master' file design - this enables us to handle very large collections in an incremental and robust fashion, yet retaining retrieval ranking flexibility as if all items are in one single large file.

Experiments in retrieval effectiveness shows that query training from past known relevant documents in a routing environment can improve average precision over all recall points by about 40% compared with no learning. Short relevant documents are the quality items for training; they are efficient and effective. A ranking of the relevants and choosing the best 'n' appears not necessary. Breaking documents into subdocuments improves retrieval for lengthy items such as those from the Federal Registry, and facilitates choices of quality items for learning. Terms with high document frequencies are necessary for good representation and performance. A choice of Zipf high frequency cut-off of 50,000 appears a good compromise between efficiency and effectiveness. Our item self-learning procedure to initialize edge weights works well as attested by our ad hoc retrieval results.

PLANS FOR THE COMING YEAR

We plan to employ 'local matching' such as sentence-sentence comparisons to improve the precision of our retrieval. Better learning samples from the relevants will be explored for routing experiments. Additions to our two-word adjacency phrase dictionary will be generated from the new collections in TREC2. Methods to enhance ad hoc retrieval will also be investigated.