# Information Retrieval from Large Textbases

*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 of retrieval and routing operations for large scale textbases. Retrieval concerns the processing of ad hoc queries against a static document collection, while routing concerns the processing of static, trained queries against a document stream. Both may be viewed as trying to rank relevant answer documents high in the output. Our text processing and retrieval system PIRCS is based on the probabilistic model and extended with the concept of document components. Components are regarded as single content-bearing terms as an approximation. Considering documents and queries as constituted of conceptual components allows one to define initial term weights naturally, to make use of nonbinary term weights, and to facilitate different types of retrieval processes. The approach is automatic, based mainly on statistical techniques, and is generally language and domain independent.

Our focus is 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 sub-document units help to decrease ambiguity. 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. It also provides a platform for using other learning techniques such as those from artificial neural networks.

## RECENT RESULTS

During 1992, we participated in TREC1 and experimented with the 0.5 GByte Wall Street Journal collection of the Tipster program. Our results based on precision-recall evaluation compared very favorably with other participants in both ad hoc retrieval and routing environments. Our experimental results support the general conclusion that techniques which work for small collections also work in this large scale environment. Specifically:

- Breaking documents with unrelated stories, or long documents into more uniform length sub-documents at paragraph boundaries, together with Inverse Collection Term Frequency weighting to account for the discrimination power of content terms, is a viable initial term weighting strategy. It is also useful to augment single terms with two-word phrases for representation.

- PIRCS's combination of query-focused and document-focused retrieval works well. Combining them with a soft-boolean retrieval strategy produces additional gains. Our boolean expressions for queries are manually formed.

- Known relevant documents used for feedback learning in our network lead to improvements compared with no feedback. More performance increases are obtained by expanding queries with terms from the relevant feedback documents.

## PLANS FOR THE COMING YEAR

We will enhance our system in both hardware and software in order to handle the two GByte multi-source textbase. We need to segment our network to fit available memory. In document representation, we will test a more powerful initial term weighting method based on document self-learning. We will generate two-word phrases automatically using word adjacency information captured during text processing. We plan to obtain boolean expressions from the well-structured query 'topics' automatically. Because more relevant documents are known, we will experiment with various learning schedules and different learning samples.