SESSION 12: INFORMATION RETRIEVAL

Donna Harman
National Institute of Standards and Technology
Building 225/A216
Gaithersburg, MD 20899

Research in information retrieval is enjoying renewed interest by many different communities. Commercial retrieval systems, which in the past have concentrated on Boolean pattern matching methodologies, are beginning to look into more sophisticated search methods, including complex statistical and/or natural language processing systems. This has spurred new interest in research in information retrieval in this community and also in the academic communities. Technology is being additionally pushed by the current emphasis on electronic communication, including digital libraries and the interlinking of offices to allow office automation. The availability of electronic text records in huge (and exponentially-growing) quantities, and the rapidly-expanding Internet access by potential users, is a third factor in promoting research.

ARPA has contributed to this increased interest by sponsoring a new test collection for information retrieval. The widespread availability of the TIPSTER collection has allowed research on large-scale, real-world retrieval problems. This has not only opened up new areas of research that were not discovered using the smaller test collections, but has provided proof that the more complex retrieval systems do indeed scale up to handle realistic text collections.

The first paper in this session, "Overview of the Second Text Retrieval Conference (TREC-2)", by Donna Harman, illustrates the use of this collection in a massive cross-system evaluation. The TREC-2 conference, held in August of 1993, compared results from 31 different retrieval systems working with the TIPSTER collection. These systems used many different approaches to retrieval, including manually constructed patterns, automatically constructed statistical queries that were input to statistical retrieval systems, and natural language approaches to information retrieval. The paper discusses the TIPSTER test collection, the evaluation methods used in TREC, and the results from the conference.

The next two papers in the session represent systems that appeared in TREC-2. The first of these papers discusses a mostly statistical system and the second of these papers discusses a system using natural language processing techniques.

The paper "Learning from Relevant Documents in Large Scale Routing Retrieval", by K.L. Kwok and L. Grunfeld, discusses experiments performed using a routing or filtering paradigm. This type of information retrieval assumes that users have a standing request for information, such as in an electronic dissemination service or an intelligence operation. There exists training information in the form of previously-seen documents considered relevant, and this training information is used to produce better queries. This paper discusses in detail the problems of learning from full-text relevant documents, which range in length from a short paragraph to many hundreds of pages. This problem is compounded by the availability of large numbers of such relevant documents. Many experiments were performed to discover the optimal method of selecting which (and what parts) of documents to use for training, and the results are given in the paper.

The next paper, "Document Representation in Natural Language Text Retrieval", by Tomek Strzalkowski, discusses experiments performed using mostly the adhoc retrieval paradigm. In this case the documents are known in advance, but the information is requested on an "adhoc" basis. There is no training data, and systems are often required to deal with short user requests that might not map well onto the terminology used in the documents. One way around this problem is to automatically transform the user query into a linguistic structure that is expanded to better map into the document collection. This paper presents a series of experiments in automatically locating useful linguistic fragments of documents to match against such a modified user query. One of the main issues dealt with here is the correct term weighting for these fragments.

Information retrieval is not limited to the matching of textual material; two of the papers in the session deal with speech retrieval systems. The first of these papers describes a modification of traditional information retrieval methods to handle speech, whereas the second paper uses traditional speech recognition technology with information retrieval as the application.

The paper, "Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors", by Peter Schauble and Ulrike Glavitsch, deals with retrieval of speech (speech "documents"). Their retrieval system uses phonetically motivated subword units as opposed to complete words for indexing of speech. The use of subwords as index terms means that the system can be used against either speech or text, and that techniques traditionally used in text retrieval can be modified for use with speech. The production of these subwords is dependent on current speech recognition technology, which is known to be error-prone. This paper presents some experiments using simulated speech recognition errors against well-known information retrieval test collections (textual) to see what effects these errors have on retrieval performance.

The second of these papers, "Speech-Based Retrieval using Semantic Co-Occurrence Filtering", by Julian Kupiec, Don Kimber, and Vijay Balasubramanian, uses a standard hidden Markov model as input to a text retrieval system. The issue in this paper is how to deal with the very large (generally unrestricted) vocabulary size that is normal for most text retrieval applications. Speech input using large vocabularies (and possibly many different speakers) is likely to produce many inaccurate words so that a direct phonetic dictionary lookup would not be reasonable. The paper presents a method using an n-best word selection to locate the user's query words, and then uses co-occurrence of these n-best query word lists to locate relevant documents.

The final paper in this session, " '(Almost)' Automatic Semantic Feature Extraction from Technical Text", by Rajeev Agarwal, does not deal directly with information retrieval, but with the production of data that would be useful in an information retrieval system. Some of the newer retrieval systems use knowledge bases to supplement (or replace) the document indices. This paper deals with natural language processing methods that allow faster acquisition of such information. The methods discussed also allow faster porting of all types of natural language systems into new domains by providing machine-aid to the building of semantic knowledge.