The SMART Information Retrieval Project

C. Buckley, G. Salton, J. Allan

Department of Computer Science
Cornell University
Ithaca, NY 14853

PROJECT GOALS

The primary goal of the SMART information retrieval project at Cornell University remains, as it has for the past 30 years, investigating the effectiveness and efficiency of automatic methods of retrieval of text. In recent years this has expanded to include retrieval of parts of documents in response to both user queries (passage retrieval) and parts of other documents (automatic hypertext links). The emphasis of SMART has always been on purely automatic text retrieval — starting from an arbitrary piece of natural language text from the user and matching against automatically indexed documents — and this continues.

RECENT RESULTS

Under this rather broad goal, we've performed a number of investigations this past year. These include:

- Local/global matching: Looking at the effect of determining an overall global similarity between query and document, and then requiring that some small local portion of the document (paragraph or sentence) focuses in on the query. The overall performance level of local/global matching for the TREC 1 workshop was quite good, though it appears the local requirement only gains about 10% improvement over a pure global match.

- Phrases: Examining methods for both statistical phrase selection and phrase weighting. For TREC 1, SMART's statistical phrases gained 5 to 9% over our single term methods.

- Learned Features of Terms: In cooperation with Norbert Fuhr, we've been looking at learning good term weights based upon characteristics of a term rather than history of how that term itself behaves. This enables us to come up with good term weights based upon much less information than conventional weight learning techniques. This did very well for TREC 1: tied at the top of the automatic ad-hoc category with the local/global approach above.

- Efficiency and Effectiveness Trade-offs: A number of tradeoffs were also examined at TREC 1. Major conclusions were
  - Retrieval effectiveness can be very reasonably traded for retrieval efficiency by truncating the retrieval appropriately.
  - Massive stemming of words to their root forms has efficiency benefits and costs, but offers no significant effectiveness gains.
  - Document indexing can be sped up significantly, at a large cost in disk space.

- Evaluation: Examining evaluation measures suitable for TREC. We supplied the TREC 1 evaluation routines, and have designed several other measures that may be used for TREC 2.

- Automatic Hypertext: Local/global matching was used to automatically construct hypertext links between articles of a 29 volume encyclopedia.

- Passage Retrieval: Local/global matching was used again to retrieve appropriate scopes of encyclopedia articles in response to a query.

- SMART System: A new publicly-available release of SMART (for research purposes only) was finished in June. This release provides support for multi-gigabyte databases.

PLANS FOR THE COMING YEAR

We'll be continuing with most of the investigations above in the coming year. We'll use automatic learning techniques to help combine local and global similarities, and to help weight phrases. Local/global matching will be used heavily in the TREC routing environment to regain precision after query expansion techniques. Passage retrieval and automatic document linkage will be extended to automatically form a coherent summary reading pattern for a topic. The SMART system itself will be revamped to enable very large distributed databases to be searched effectively and efficiently.