The TREC Terabyte Retrieval Track

Charles Clarke
University of Waterloo
claclark@plg.uwaterloo.ca

Nick Craswell
Microsoft Research
nickcr@microsoft.com

Ian Soboroff
NIST
ian.soboroff@nist.gov

The Terabyte Retrieval Track of the Text REtrieval Conference (TREC) provides an opportunity to test retrieval techniques and evaluation methodologies in the context of a terabyte-scale corpus. Given the size of the corpus, the track also provides a vehicle for participants to investigate query and indexing speeds. This brief summary outlines track activities to date and previews our plans for TREC 2005. For complete information, the reader should consult the full version of this report [1].

A proposal for the Terabyte Track was developed during a SIGIR 2003 workshop and was accepted by the TREC program committee for inclusion in TREC 2004. For the initial year of operation, we decided to base the track on a crawl of the "gov" domain, since we believed that this would generate roughly a terabyte of data and would provide us with a realistic setting, in which both links structure and anchor text could be productively exploited.

A crawl of gov in early 2004 produced a collection of 25 million pages (426GB), including the extracted text from PDF, Word and postscript files. While this collection (the "GOV2" collection) is considerable smaller than our initial target of 100 million pages, we believe it contains a substantial fraction of the crawlable pages existing in the gov domain at that time.

For 2004, the track task was classic adhoc retrieval, a task which investigates the performance of systems searching a static set of documents using previously unseen topics. Assessors at NIST created 50 new topics for the task, and the TREC organizers provided other resources needed to manage the task. Each participating group was permitted to submit up to five experimental runs, with each run consisting of the top 10,000 documents for each topic. For each run, we also asked groups to report the average query time, indexing time, index size, and hardware configuration.

The groups used a surprising variety of indexing and retrieval techniques to accomplish the task. Many groups divided the collection across a cluster of machines and searched the collection in parallel. Retrieval formulae included Okapi BM25, the cosine measure and language modeling techniques. Several groups also took advantage of Web-specific methods, such as link analysis, anchor text and document structure. The fastest reported average query times ranged under 100ms; the fastest reported indexing times speeds ranged over 200GB/hour.

The track will continue to use the GOV2 collection for at least one more year, providing a total of 100 topics over this collection. TREC 2005 may also include a known-item retrieval task. In future years, we may expand the collection to a full terabyte (or more) and add additional tasks, as dictated by the interests of the participants and the availability of resources.

References

[1] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 Terabyte Track. In *Proceedings of the Thirteenth Text REtrieval Conference*, Gaithersburg, MD, November 2004. NIST Special Publication 500-261. See trec.nist.gov.