Towards Grid-Based Information Retrieval

Gregory B. Newby*
Arctic Region Supercomputing Center
University of Alaska Fairbanks

Abstract

The IRTools software toolkit was used in TREC 2004 for submissions to the Web track and the Terabyte track. Terabyte track results were not available at the time of the due date for this Proceedings paper. While Web track results were available, qrels were not. Because we discovered a bug in the MySQL++ API that truncated docid numbers in our results, we will await qrels to reevaluate submitted runs and report results.

This year, the Terabyte track dictated some changes to IRTools in order to handle the 430+GB of text (about 25M documents). The main change was to operate on chunks of the collection (272 separate chunks, each containing one of the Terabyte collections' subdirectories). Chunks were generated in parallel using the National Center for Supercomputing Application's cluster, Mercury (dual Itanium systems). Up to about 40 systems were used simultaneously for both indexing and querying. Query merging was simplistic, based on the cosine value with Lnu.Ltc weighting.

Use of the NCSA cluster, and other experiments with commodity clusters, is part of work underway to enable information retrieval in Grid computing environments. The site http://www.gir-wg.org has information about Grid Information Retrieval (GIR), including links to the published Requirements document and draft Architecture document. The GIR working group is chartered by the Global Grid Forum (GGF) to develop standards and reference implementations for GIR.

TREC participants are urged to consider getting involved with Grid computing. Computational grids offer a very good fit for the needs of large-scale information retrieval research and practice.

This brief abstract for the proceedings will be replaced with a complete analysis of this year's submissions for the full conference paper. Meanwhile, Newby (2004) provides a profile of IRTools, which is generally applicable to this year's submissions.

References

Newby, Gregory B. 2004. "Document Structure with IRTools." In: Voorhees, Ellen (Ed.). NIST Special Publication 500-255:The Twelfth Text REtrieval Conference (TREC 2003). Gaithersburg, Maryland: NIST. pp. 568-577.

---

* 909 Koyukuk Dr. Fairbanks AK 99775. newby@arsc.edu or http://www.arsc.edu/~newby .
The research described here was funded in part by National Science Foundation grant #0352029.