Meta-data for Distributed Text Retrieval

Grace Crowder *
crowder@cs.umbc.edu

Charles Nicholas
nicholas@cs.umbc.edu

Computer Science and Electrical Engineering Department
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
crowder,nicholas@cs.umbc.edu
June 4, 1997

In an agent-based information retrieval architecture, we use intelligent agents to route queries and new documents to appropriate locations. Users interact with "user agents", which send queries to information providers, or "back-ends" for processing. These back-ends are individually responsible for some portion of the corpus. The corpus as a whole can be regarded as distributed, heterogeneous, dynamic (in the sense that new documents may arrive at any time), and possibly quite large.

In order for such systems to scale, some designated "mediators" or brokers must assist in the routing of queries. The brokers, in turn, need meta-data from each of the back-ends in order to make these routing decisions. N-gram centroids are an attractive form of meta-data for mediated agent-based information retrieval systems. We use n-gram centroids in our CARROT system, where CARROT stands for "Cooperating Agent-based Routing and Retrieval of Text [3].

As the corpus grows, the size of the meta-data may become an issue, in terms of communication as well as storage costs. We therefore need forms of meta-data which are both concise and effective for routing. CARROT uses n-gram representations of text in two ways: as terms in a back-end information

*This work was supported in part by the United States Department of Defense.

retrieval system and as meta-data produced by every back-end for use by the broker. The first back-end information provider to be adapted for use in CARROT was Telltale [5], which is an n-gram based information retrieval system. We have since added Witten, Moffat, and Bell's mg system [7], with a stripped-down version of Telltale to generate n-gram based meta-data.

The goal of n-gram reduction is to keep those n-grams which provide the best information for making routing decisions, while keeping the size of the meta-data object as small as possible. In our experiments with the documents in a single corpus, we have investigated two reduction techniques so far. We find that by discarding singleton n-grams, i.e. n-grams that occur exactly once in any one document, we are able to reduce the size of the corpus centroid by as much as 60% for some small test collections, without significant change in effectiveness as measured using standard precision-recall measurements. The other reduction technique we've looked at, which is based on Cohen Highlights [2], calculates a novelty score for each n-gram and retains only those n-grams that distinguish a given document from the rest of the corpus.

Our current work involves describing a corpus using both the average, or centroid n-grams to capture language information and deviates, or novel n-grams, to capture content information. In building the average set of n-grams we set upper and lower thresholds to discard those n-grams that occur too [in]frequently to be meaningful. The novel n-grams for a corpus are those n-grams that occur more frequently in a corpus than would be expected given the collection of corpora being managed by a broker.

We have also started work on a reduction technique based on latent semantic indexing, in which the principal components of the term-document matrix (where terms are n-grams) are used to select those n-grams that "explain" the most variance in the corpus [1].

We've evaluated these reduction techniques using traditional precision and recall measures [6], the size reduction in the resulting meta-data object, and the time needed to generate the meta-data. We also evaluated routing effectiveness by constructing confusion matrices. Confusion matrices measure how often a query (or new document) is routed to the "correct" corpus [4]. Our test data is drawn from a number of sources, including Project Gutenberg, the TREC corpus, and other corpora available from the Linguistic Data Consortium.

Our contribution to the discussion of networked information retrieval

systems is in the exploration of language-independent, corruption-tolerant, automatically-generated, scalable meta-data for use in making routing decisions.

References

[1] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. SIAM Review, vol. 37, No. 4, pp. 573-595, December 1995.

[2] Jonathan Cohen. Highlights: Language and Domain Independent Automatic Indexing Terms for Abstracting. *Journal of the American Society for Information Science, (JASIS)*, April 1995.

[3] Grace Crowder and Charles Nicholas. Resource Selection in CAFE: an Architecture for Network Information Retrieval *Networked Information Retrieval Workshop at SIGIR'96*. August 1996.

[4] Bradley Kjell and Ophir Frieder. Visualization of Literary Style. *1992 IEEE International Conference on Systems, Man, and Cybernetics.* October 1992.

[5] Claudia Pearce and Charles Nicholas. TELLTALE: Experiments in a Dynamic Hypertext Environment for Degraded and Multilingual Data. *Journal of the American Society for Information Science (JASIS)*, April 1996.

[6] Gerard Salton and Michael J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[7] Ian H. Witten, Alistair Moffat, Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York, 1994.