The TREC-5 Database Merging Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD
ellen.voorhees@nist.gov

There are many times when users want to search separate text collections as if they were a single collection. For example, computer networks can provide access to a variety of corpora that are owned and maintained by different entities. Instead of issuing search commands to each of the databases in turn and manually collating the individual results, users prefer a mechanism for performing a single, integrated search. In other cases, reliability and efficiency concerns may dictate that databases that are under the same administrative control should be physically separate. Again, users want to issue a single search request that returns an integrated result. The database merging track investigates methods for combining the results of separate searches into a single, cohesive result.

1 The Task

The initial running of the database merging track occurred in TREC-4. To foster participation by allowing as many different types of merging strategies as possible, the task in the TREC-4 track was left very open: the data was split into ten collections (corresponding to each source on each TREC disk used in the ad hoc task) and participants were free to produce a merged result any way they saw fit.

The task in TREC-5 was somewhat more focussed. The track used the same topics as the ad hoc task (topics 251–300), and the same documents as the ad hoc task (the documents on TREC disks 2 and 4). (This allowed the track to contribute to the ad hoc relevance assessment pools, and to use those pools to evaluate the runs.) The documents on the two disks were partitioned into 98 different databases, with each partition containing documents from a single source[1]. Participants were required to produce a ranking of the documents for each topic *without* searching every database for every topic. That is, merging strategies that routinely search all available databases were specifically excluded from the track.

Track participants could submit up to two merging runs, and were required to submit a comparable ad hoc run (all documents in a single collection) to provide a baseline for comparisons.

Database merging consists of two sub-problems: *resource discovery* and *result combination*. Resource discovery is deciding which of the set of available databases should be searched for the current query; result combination is producing one ranked list of documents from the results of the sources searched. The decisions to significantly increase the number of databases over the TREC-4 task and to exclude methods that that always search all databases were made to focus this running of the track on the resource discovery subproblem.

[1] The databases were defined by a script created by the group at the University of Massachusetts at Amherst. Contact Ellen Voorhees at NIST (ellen.voorhees@nist.gov) for a copy. Category B participants used WSJ90, WSJ91, and WSJ92 as separate databases.

2 Participants

Three groups participated in the TREC-5 track. See the respective papers by these groups elsewhere in the proceedings for more details regarding their results.

Université de Neuchâtel: The Université de Neuchâtel group used TREC-5 to investigate a retrieval model based on logistic regression that treats data fusion (combining different search schemes) and database merging (combining distributed information services) to be different facets of the same problem. They submitted two category B database merging runs using both long and short topics (*UniNE0* and *UniNe9*).

Australian National University: This group used the resource discovery emphasis of the track to examine the specific problem of selecting network servers. In addition to retrieval effectiveness, their work examined the efficiency measure of the number of servers that needed to be contacted to produce the result. They submitted three category A runs: *anu5mrg0*, their baseline ad hoc run; *anu5mrg1*, a run that used historical data to pick servers; and *anu5mrg7*, a run that used lightweight probes to pick servers.

FS Consulting: This group used their database merging track entry to measure the effectiveness of their document scoring algorithms when searching across multiple databases. They found the document scoring algorithm to be stable for widely varying numbers of databases. FS Consulting submitted one category A database merging run, *fsclt3m*, which is comparable to their ad hoc submission, *fsclt3*.

3 Future of the Track

Unfortunately, the database merging track has proven to be a high overhead track for participants. Despite generally high interest in the problem addressed by the track, the track has attracted few participants, likely because of the amount of data manipulation it requires as compared to other TREC tracks. The track will be suspended for at least a year while a simpler track design is sought.