Sheffield and Terabyte-Scale Test Collections

Hideo Joho and Mark Sanderson
h.joho|m.sanderson@shef.ac.uk

Department of Information Studies, University of Sheffield, Western Bank, Sheffield, UK

The information retrieval research group at the Department of Information Studies, University of Sheffield, has been developing a large web test collection – approximately 94 million documents, 1.2TB in size[1] – under an EU-funded project SPIRIT[2]. One of the objectives of the project is to devise evaluation methodologies using as little human effort as possible. This note summarises our experience and thoughts on building very large test collections based on preliminary experiments. Although we understand the importance of addressing the evaluation issues comprehensively (including usability studies), our concerns will focus on the ad-hoc task not others such as known-item or distillation tasks. However we believe that our discussions can be useful for those tasks, too.

We revisited the interactive searching and judging (ISJ) approach of Cormack, Palmer, & Clarke[3] where a person both searches and assesses retrieved documents issuing queries and modifications of queries to locate as many relevant documents per topic as possible. The advantage of the approach as reported in the 1998 paper is that a single assessor using just one retrieval system can locate relevant documents much more efficiently than is achieved in the traditional TREC system pooling based approach, although at the expense of missing a portion of relevant documents. A form of ISJ is used by TDT assessors[4] and the method has been discussed at the NTCIR evaluation workshop[5]. Such an approach will more than likely be needed in order to build relevance judgements in the proposed terabyte collection discussed at the workshop.

Our initial work has focussed on conducting experiments to determine if the ISJ technique works consistently well across different searchers and different retrieval systems. Using a methodology that exploits previous TREC submission data, we measured the effectiveness of ISJ across 17 distinct searcher-system pairs. The ISJ approach was found to work well at locating relevant documents quickly and accurately for 15 such pairs, a strong indication of the utility of the method. As a part of the study, a new way of measuring the accuracy of a test collection relative to an ideal was also produced. The measure computes the precision and recall of a new test collection in locating significant differences between retrieval systems measured by the collection.

The potential drawback of the interactive searching and judging approach is the possibility of bias in the choice of relevant documents: introduced by either the searching strategy of the assessor or by the retrieval algorithm of the system used by the assessor. Such bias might create test collections that favour certain types of retrieval systems over others. We plan to investigate the potential for bias through a number of careful analyses of how ISJ judgements in TREC collections differ from the main judgements produced by TREC NIST assessors. The potential for bias in ISJ is often cited as a reason for not adopting the method when building test collections. Studying the level of bias is important before ISJ is accepted as a proven technique.

[1] The collection is a copy of one held by Ed Fox and Virginia Tech, who in turn obtained it from Charlie Clarke who ran a large web crawl at the University of Waterloo.

[2] Contract No. IST−2000−26162; http://www.geo-spirit.org/

[3] Cormack, G.V., Palmer, C.R., Clarke, C.L.A. (1998) Efficient construction of Large Test Collections, in *the proceedings of 21st ACM SIGIR Conference*, 282 - 289

[4] Kuriyama, K., Kando, N., Nozue, T., Eguchi, K. (2002). Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop, *Information Retrieval*, 5, 41-59.

[5] Fiscus, J. G. and Doddington, G. R. (2002) Topic Detection and Tracking Evaluation Overview In: *Allan, J (Ed) Topic Detection and Tracking: Event-based Information Organization*, 17-31, Kluwer Academic Publishers.