ClueWeb09 and TREC Diversity

Charles Clarke
University of Waterloo, Canada

Abstract

The TREC Web Track explores and evaluates Web retrieval technologies. The TREC 2009 Web Track included both a traditional adhoc retrieval task and a new *diversity task*. The goal of this diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. Both tasks will continue at TREC 2010, which will also include a new *Web spam task*. The track uses the ClueWeb09 dataset as its document collection. This collection consists of roughly 1 billion web pages in multiple languages, comprising approximately 25TB of uncompressed data crawled from the general Web during January and February 2009.

For TREC 2009, topics for the track were created from the logs of a commercial search engine, with the aid of tools developed at Microsoft Research. Given a target query, these tools extracted and analyzed groups of related queries, using co-clicks and other information, to identify clusters of queries that highlight different aspects and interpretations of the target query. These clusters were employed by NIST for topic development. For use by the diversity task, each resulting topic is structured as a representative set of subtopics, each related to a different user need. Documents were judged with respect to the subtopics, as well as with respect to the topic as a whole.

In 2009, a total of 18 groups submitted runs to the diversity task. To evaluate these runs, the task used two primary effectiveness measures: α-nDCG as defined by Clarke et al. (SIGIR 2008) and an “intent aware” version of precision, based on the work of Agrawal et al. (WSDM 2009). Developing and validating metrics for diversity tasks continues to be a goal of the track. For TREC 2010, we will report a number of additional evaluation measures that have been proposed over the past year, including an intent aware version of the ERR measure described by Chapelle et al. (CIKM 2009).

Nick Craswell from Microsoft serves as the track co-coordinator. Ian Soboroff is the NIST contact. The ClueWeb09 collection was created through the efforts of Jamie Callan and Mark Hoy at the Language Technologies Institute, Carnegie Mellon University. More information may be found on the track Web page: http://plg.uwaterloo.ca/~trecweb/2010.html.

Bio

Charles Clarke is a professor in the David R. Cheriton School of Computer Science at the University of Waterloo, Canada. He has published on a wide range of topics within the area of information retrieval, including papers related to evaluation, efficiency, ranking, parallel systems, security, question answering, document structure, and XML. He was a Program Co-Chair of SIGIR 2007 and General Co-Chair of SIGIR 2003. From 2004 to 2006 he was the coordinator of the TREC Terabyte Retrieval track. Since 2009 he has been a co-coordinator of the TREC Web Track. He is a co-author of the book *Information Retrieval: Implementing and Evaluating Search Engines (*MIT Press, 2010). He has previously held software development positions at a number of computer consulting and engineering firms. In 2006 he spent a sabbatical at Microsoft, where he was involved in their search engine development effort.