The CSIRO Enterprise Search Test Collection

Peter Bailey
CSIRO, Australia
peter.bailey@csiro.au

Nick Craswell
Microsoft, UK
nickcr@microsoft.com

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Arjen P. de Vries
CWI, The Netherlands
arjen@cwi.nl

Abstract

This article describes a new TREC Enterprise Track search test collection – CERC. The collection is designed to represent some real-world search activity within the enterprise, using as a specific example the Commonwealth Scientific and Industrial Research Organisation (CSIRO). It has a deep crawl of CSIRO's public-facing information, that is very similar to the crawl of a real-world search service provided by CSIRO. The search tasks are based on the activities of CSIRO Science Communicators, who are CSIRO employees that deal with public-facing information. Topics and judgments are tied to the Science Communicators in various ways, for example by involving them in the topic development process. The overall approach is to enhance the validity of the test collection as a model of enterprise search, by tying it to real-world examples.

1 Introduction: Enterprise Test Collections

Effective information retrieval is important for the efficient operation of an organisation. According to estimates from industry analysts, knowledge workers spend 15-25% of their time on non-productive information-related activities [2] and up to 10% of staff costs are wasted because "employees simply can't find the right information to do their jobs" [3]. Enhancing enterprise search could mean that search sessions are shorter, freeing up time for other activities. It could also mean fewer failed search sessions, therefore a reduction in wasteful activities such as reproducing work that has already been done.

To address the enterprise search problems via a test collection, it is important for the test collection (documents, topics and judgments) to be representative of real-world enterprise search. This is one of our main concerns. Another case where test collections were tied to a specific real-world activity was the TREC Web Track [4] which had, for example, 'navigational' topics and judgments. These could only be answered effectively by exploiting structure in the web document collection.

To build a realistic enterprise search test collection, we first choose an organisation (enterprise). This presents two potential problems. First, each enterprise is different, so it is

Number:	14	7	25	31	15	8
Bucket:	10^0	10^1	10^2	10^3	10^4	10^5

Table 1: Size distribution of hosts by number of pages, bucketed in powers of 10.

not clear whether a collection built on a single enterprise teaches us about enterprise search in general. However, this is the case for other types of test collection; it is important to reproduce results on different collections. There may be more cross-collection variation in enterprise search, but this remains to be seen. The second potential problem is that real-world enterprise search involves confidential information, some of which is only available to people who work at that organisation, and some of which is only available to a subset of employees. Usually this information is unavailable for building a test collection, and even if an organisation's internal documents are made publicly available, some users of the test collection might have objections to using documents that were meant to be private. Therefore enterprise test collections are usually based on public-facing documents of an organisation.

Enterprise Track experiments in 2005 and 2006 were based on the World Wide Web Consortium (W3C) test collection [1]. The W3C crawl comprises documents of multiple types including email archives, web pages and wiki. Experiments were email search and expert search. CERC was introduced in TREC 2007, to allow the evaluation of enterprise search techniques on a new test collection. CERC is also the first Enterprise Track corpus developed with the involvement of real users (CSIRO Science Communicators, see Section 2.2).

2 CSIRO Enterprise Research Collection

The CSIRO Enterprise Research Collection (CERC) (`http://es.csiro.au/cerc/`) is an information retrieval test collection, and consists of a document collection, topic descriptions, and relevance judgments for documents and experts. It was first used in the 2007 Enterprise Track.

2.1 Documents

The crawl is of *.csiro.au (public) websites conducted in March 2007. The crawl has 370 715 documents, with total size 4.2 gigabytes. The crawl is of outward-facing pages of CSIRO, and is meant to be similar to the crawl used in CSIRO's own search engine. In fact, the same crawler technology that CSIRO uses was used to gather the CSIRO documents (`http://www.funnelback.com/`).

The crawl has 100 hosts and Table 1 shows the distribution of host sizes, with fourteen hosts having only one page each and 8 hosts of size 10 000–100 000. The largest host is `anic.ento.csiro.au` with 72 358 pages. The corpus has approximately 7.9 million hyperlinks, and 95% of pages have one or more outgoing links containing anchor text.

The 2007 Enterprise Track also requires participants to extract email addresses, for use in Expert Search experiments. One participating organisation extracted email addresses of 3678 individuals, with 38% of documents containing at least one mailto field.

2.2 Science Communicators

Before describing the test collection's topics and judgments, we describe a key aspect of the CSIRO test collection: our use of science communicators. A science communicator's role in CSIRO is to enhance CSIRO's public image and promote the capabilities of CSIRO by managing information and interacting with industry groups, government agencies, professional groups, media and the general public.

Science Communicators read and create the outward-facing web pages of CSIRO (as opposed to internal documents). Therefore they were a natural choice when thinking of which users are a good match for our outward-facing crawl.

The primary method for involving Science Communicators was asking them to do topic development. A general email was sent to all science communicators, calling for them to create topics in their area. This yielded 25 usable topics from 9 science communicators from multiple CSIRO divisions. Being short of the standard 50 topics, we then approached one of these communicators and employed them to produce a further 25 topics.

2.3 Topics and Judgments

The 50 topics for use in two types of experiment. One is document search, where the scenario is that no overview page exists for CSIRO projects in a certain topic, so a Science Communicator is trying to build such a page. The retrieval system should list 'key references': pages that should be linked to by a good overview page. For example, a project's homepage in csiro.au is probably a key reference to be included in an overview page. Another way of thinking of the document search task is that the results list could be shown to users who are visiting CSIRO's site, if the overview page has not yet been created.

Each topic description has a query, a description of the information need, some examples of key reference URLs (on average 4 per query) and a short list of key contacts for expert search. Because we are considering cases where an overview page might be created, the topic should not be too narrow. For example, we could imagine a science communicator building an overview page for 'enterprise search' but not an overview page dedicated to 'using BM25 on anchor text in enterprise search'. The latter has too narrow an audience, and creating overview pages of all CSIRO activities at that granularity would require a massive number of overview pages. To jog the memory of science communicators and give them examples of good general queries, we showed them a list of the most popular queries from CSIRO's real public search site.

The same topics are also used for expert search. The relevant experts are 'key contacts' who could also be listed on an overview page. Anecdotally, from a science communicator's point of view, we also believe that an expert search system that lists a handful of key contacts is more valuable than one listing an undifferentiated list of broader contacts. On average, the number of key contacts per topic was 3 (with a minimum of 1 and a maximum of 11).

For expert search, we did no further judging, using the experts listed in the topic as our ground truth. For document search we used community judging.

3 Summary

We have crawled documents from an enterprise to get a corpus snapshot that is largely identical to what CSIRO's real search service would use. We have employed science communicators

to list topics of interest to them, and for each topic to list some key experts and pages. Using these topics, the TREC 2007 Enterprise Track has a document search experiment and an expert search experiment. As future work, we hope to further validate our experiment in the eyes of science communicators, by asking their opinion of our participant judgments and possibly our notion of what makes a good search results list.

Acknowledgements

Thanks to Peter Thew from CSIRO for carrying out the crawl of csiro.au using the Funnelback v6.0 web crawler, and to the CSIRO science communicators who provided topics for the collection.

References

[1] Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC-2005 Enterprise Track. *TREC 2005 Conference Notebook*, pages 199–205, 2005.

[2] Feldman and Sherman. *The High Cost of Not Finding Information.* IDC Technical Report #29127, 2003.

[3] Butler Group Report. *Enterprise Search and Retrieval.* Butler Group, October 2006.

[4] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press, 2005.