Overview of the TREC 2007 Enterprise Track

Peter Bailey
Microsoft, USA
pbailey@microsoft.com

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

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

1 Introduction

The goal of the enterprise track is to conduct experiments with enterprise data that reflect the experiences of users in real organizations. This year, the track has introduced a new corpus with the goal to be more representative of real-world enterprise search, by involving actual members of the organization in the topic development process, performing their real work tasks.

2 Collection

The CERC corpus (CSIRO Enterprise Research Collection, (`http://es.csiro.au/cerc/`)) represents the public-facing web of the Australian Commonwealth Scientific and Industrial Research Oganization (CSIRO). Here, we summarize the main characteristics of this corpus; a complete description of the collection is given in Bailey et al. (2007).

2.1 Data

The collection consists of all the `*.csiro.au` (public) websites as they appeared in March 2007. The resulting data set consists of 370 715 documents, with total size 4.2 gigabytes. The web crawler visited the outward-facing pages of CSIRO in a fashion 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 corpus contains approximately 7.9 million hyperlinks, and 95% of pages have one or more outgoing links containing anchor text. One participant extracted email addresses of 3678 individuals, with 38% of documents containing at least one `mailto` field.

2.2 Users

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.

2.3 Tasks and Topics

The 2007 enterprise track defined two tasks: document search and expert search. Both search tasks are grounded in a 'missing overview page' scenario, where the science communicator has to construct a new overview page on the topic of interest, that enumerates the 'key pages' and a few 'key people' of interest. Given this scenario, the document search task models the problem of finding the set S of 'key pages', and the expert search task the problem of locating the 'key contacts' among CSIRO staff.

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. Examples of general queries from CSIRO's real public search site were given for inspiration. 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 who produced another 25 topics to complete the set.

Each topic description has a query and narrative, some examples of key reference URLs (on average 4 per topic) and a short list of key contacts (on average 3 per topic, varying from 1 to 11). The key reference URLs serve as a (admittedly somewhat poor) surrogate for click-log data. Note that both tasks have used the same set of topics.

2.4 Assessments

For document search we used community judging. NIST formed pools and sent them to CSIRO, where the assessment system was hosted. Track participants then judged the pools through the CSIRO system (adapted from the assessment system used in the Million Query track).

The guidelines instructed the assessors to read the query and narrative, and optionally carry out a Web search to learn more about the subject. The guidelines also emphasized that science communicators are web-savvy users – so judgments should take into account that navigational answers and relevant homepages are important results in exploratory search behaviour. Relevance judgments were made on a three-point scale:

2: Highly likely to be a 'key page'.
1: Possible as a candidate for a page in S, or otherwise informative to help build an overview page, but not highly likely.
0: Not a 'key page' as unlikely to be included in S, because, e.g., not relevant, off-topic, not an important page on the topic, on-topic but out-of-date, not the right kind of navigation point, or too informal or too narrow an audience.

After the workshop, we investigated to what extent the people making relevance judgements for the document search task have been exchangeable, comparing assessments made by participants ('bronze' judges) to sampled re-assessments for 33 topics by the topic authors ('gold' judges) and/or other science communicators familiar with the task ('silver' judges). The main finding from the study is that the bronze judges may not be able to substitute for topic and task experts, due to changes in the relative performance of assessed systems, and gold judges are preferred. The full details of this post-TREC study can be found in Bailey et al. (2008).

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

Table 1: Document search results for the automatic run with the highest MAP from each group.

Group	Run	MAP	NDCG	P@20
CAS	DocRun02	0.422	0.743	0.527
York	york07ed4	0.416	0.730	0.513
Waterloo	uwtbase	0.388	0.707	0.508
RMIT	RmitQ	0.388	0.698	0.471
SJTU	SJTUEntDS02	0.374	0.692	0.475
UvA	uams07bfb	0.369	0.675	0.445
Tsinghua	THUDSFULLSR	0.366	0.701	0.461
UALR	UALR07Ent1	0.357	0.662	0.428
Fudan	FDUBase	0.350	0.664	0.426
OU	ouTopicOnly	0.345	0.646	0.464
Glasgow	uogEDSF	0.337	0.675	0.413
DUT	DUTDST4	0.336	0.644	0.441
Iowa	uiowa07entD2	0.310	0.597	0.413
Hyberdad	QRYBASICRUN	0.246	0.487	0.408
CSIRO	CSIROdsQonly	0.194	0.352	0.378
St. Petersburg	insu2	0.028	0.185	0.041

3 Results

3.1 Document search

Systems return docids for document search. Participants submitted 43 automatic, 15 feedback and 5 manual runs. The pools for document search included the top 75 documents from two runs per participant.

Runs were evaluated on their capability to retrieve the key pages, using traditional retrieval measures including MAP and precision at fixed ranks; NDCG is reported to take into account the graded assessments.

Automatic runs may use the query and narrative fields of the topic, but each participating group had to submit at least one run using the query field only. Table 1 shows the best automatic run from each participating group based on mean average precision. Ordering on descending NDCG instead of MAP gives slightly different results; e.g., University of Waterloo's uwKLD run (using query expansion from pseudo-relevant documents) would come second and beat their best MAP-based uwtbase run, and the Open University's ouNarrAuto run (using the narrative for automatic query expansion) would give better results than the ouTopicOnly baseline. These observed differences seem to suggest that query expansion from documents or the topic narrative is more useful when trying to find the highly relevant documents than when just finding any type of relevant document.

Feedback runs can be thought of as simulating one type of click-based system. Using click logs, it is often possible to identify that we have seen this query before, and that one or two URLs were often clicked. In that case, it would be interesting to take those URLs as relevant and perform relevance feedback. Unfortunately, we do not have CSIRO click logs, but we can use the pages field of the topic, to simulate what would happen in such a case. Feedback runs should use the query and pages fields only (not the narrative field and no manual intervention).

There are at least two methods for evaluating relevance feedback in a way that allows a comparison between feedback and non-feedback runs. The predominant method in IR is to evaluate on the residual collection, that is, feedback documents are removed from all runs and the relevance judgments. In the web search engine community, another method known as

Table 2: Document search results for the automatic or feedback run with the highest MAP from each group, using residual ranking. Feedback runs are labeled with a '*'.

Group	Run	MAP	NDCG	P@20
Waterloo	uwRF*	0.395	0.691	0.479
York	york07ed4	0.386	0.677	0.472
UvA	uams07bfbex*	0.359	0.640	0.461
RMIT	RmitQ	0.357	0.633	0.423
CAS	DocRun02	0.353	0.666	0.457
UALR	UALR07Ent2*	0.344	0.623	0.423
SJTU	SJTUEntDS02	0.337	0.629	0.417
Fudan	FDUBase	0.320	0.591	0.382
Tsinghua	THUDSFULLSR	0.310	0.602	0.390
DUT	DUTDST2	0.298	0.577	0.386
OU	ouTopicOnly	0.296	0.582	0.401
Glasgow	uogEDSCLCDIS*	0.290	0.582	0.368
Iowa	uiowa07entD2	0.276	0.555	0.354
Hyberdad	QRYBASICRUN	0.202	0.413	0.353
CSIRO	CSIROdsQonly	0.127	0.282	0.305
St. Petersburg	insu2	0.024	0.146	0.033

promotion is used — the feedback documents are moved to the top of all rankings, or placed there if they have not been retrieved.

Table 2 summarizes the results using residual-collection evaluation. For these scores, the key pages from the topics have been removed from both the qrels and the run. This allows feedback and non-feedback runs to be compared directly, but the residual-collection scores in Table 2 are not comparable to the scores in Table 1. The overall best run is a feedback run, but the difference from the best automatic run is marginal (less than 1% in MAP). Not all groups submitted feedback runs, and for some groups that did, their feedback runs were worse than their non-feedback runs.

Table 3 reports again results for feedback runs, however this time using promotion evaluation. Here, the key pages are moved to or placed at the top of the ranking. This evaluation is another way to compare feedback and non-feedback runs to each other; by comparing the scores of baseline and feedback runs both with and without promotion, you can see if the feedback is generalizing beyond the feedback documents. The table lists only results for submitted feedback runs (so automatic runs are not included in this ranking). Only for Waterloo, UvA and Glasgow, using feedback information lead to their best results; the other teams submitted non-feedback runs that performed better than their feedback runs.

Manual runs involve humans in the loop at any stage, for example composing queries from the topics, manual term expansion, relevance feedback, or manual combination of results. Although DUT submitted a highly performing manual run (run DUTDST1, with MAP 0.402 and NDCG 0.725), it did not outperform the two best automatic runs (by CAS and York University), nor did it outperform the best feedback run (by University of Waterloo).

The remainder of this section reviews some highlights from the participant papers on their document search activities. Several teams experimented with web retrieval methods based on anchor text or determining a static ranking (e.g., by pagerank or URL length), but the results seem to indicate that the CSIRO data behaves differently from Web data and that these methods are less effective than expected. RMIT mentions the fact that most links originate from the non-content part of the CSIRO pages, i.e., layout structure such as menu bars; SJTU and Tsinghua

Table 3: Document search results for the feedback run with the highest MAP from each group, after promotion of the feedback documents.

Group	Run	MAP	NDCG	P@20
Waterloo	uwRF	0.500	0.787	0.585
UvA	uams07bfbex	0.470	0.750	0.555
UALR	UALR07Ent3	0.449	0.720	0.526
DUT	DUTDST3	0.424	0.696	0.523
Glasgow	uogEDSCLCDIS	0.411	0.714	0.482
Fudan	FDUFeedT	0.399	0.693	0.498
SJTU	SJTUEntDS04	0.387	0.706	0.501
Iowa	uiowa07entD4	0.370	0.672	0.474
CSIRO	CSIROdsQfb	0.256	0.435	0.436

Table 4: Expert ranking scores. The best run in each group according to MAP is shown.

Group	Run	MAP	P@5	P@20
Tsinghua	THUIRMPDD4	0.4632	0.2280	0.0910
SJTU	SJTUEntES03	0.4427	0.2360	0.0910
OU	ouExTitle	0.4337	0.2520	0.0950
CAS	ExpertRun02	0.3689	0.2040	0.0790
CSIRO	CSIROesQnarr	0.3655	0.2240	0.0770
Wuhan	WHU10	0.3399	0.1960	0.0710
Glasgow	uogEXFeMNZcP	0.3138	0.2200	0.0800
UvA	uams07exbl	0.3090	0.2080	0.0790
DUT	DUTEXP1	0.2630	0.1400	0.0580
Fudan	FDUn7e3	0.1788	0.1440	0.0610
Beijing	PRISRR	0.1571	0.0920	0.0440
Twente	qorwnewlinks	0.1481	0.1080	0.0540
Peking	zslrun	0.0944	0.0600	0.0220
Hyberbad	AUTORUN	0.0939	0.0560	0.0330
UALR	UALR07Exp1	0.0200	0.0160	0.0130

made independently the same observation and used the percentage of links to seperate layout from content and weight the latter stronger. Tsinghua reports an improvement using Pagerank and HITS, but the improved results are lower than the Lemur language modelling baseline without static weighting reported by RMIT. The participants who used the narrative, e.g. for query expansion, report improved effectiveness over their baseline systems.

3.2 Expert search

Expert finding systems participating in the 2007 enterprise track had to return email addresses to identify candidate experts. Since no canonical list of candidate experts could be made available, the track required participants to extract the email addresses of the 'key people' from the data. Participants submitted 45 automatic, 4 feedback and 6 manual runs.

The evaluation results, summarized in Table 4, measure the quality of the ranked list of people using traditional retrieval measures including MAP and precision at fixed ranks.

Tables 6 and 5 summarize the results of the feedback and manual runs. For expert search, the best runs are manual runs, but notice how many automatic runs have outperformed the

Table 5: Expert ranking scores of feedback runs.

Group	Run	MAP	P@5	P@20
CSIRO	CSIROesQpage	0.3660	0.2040	0.0670
Iowa	uiowa07entE1	0.2828	0.1640	0.0710
Twente	feedbackrun	0.2371	0.1480	0.0650

Table 6: Expert ranking scores of manual runs.

Group	Run	MAP	P@5	P@20
OU	ouExNarrRF	0.4787	0.2720	0.0990
OU	ouExNarr	0.4675	0.2680	0.0980
DUT	DUTEXP3	0.3404	0.1840	0.0680
DUT	DUTEXP2	0.3324	0.1920	0.0640
DUT	DUTEXP4	0.1876	0.1000	0.0440
UALR	UALR07Exp3	0.1840	0.1320	0.0360

other manual and the feedback runs.

We again highlight some findings from studying the participant papers. Most participants use some form of two-stage model. Several teams (e.g., SJTU, UvA) retrieved homepages of the identified candidate names to aid in the expertise assessment. Proximity between candidate mentions and query terms seems an important factor in SJTU, Glasgow and OU results. Both CAS and Twente experimented with query-specific graphs of expert-document pairs, but results are not yet conclusive. What we can however conclude from this year's experiments is that the lack of candidate list has complicated the task significantly when compared to previous years. Almost all participants have used template matching to identify candidates from email occurrences in the corpus, sometimes including sophisticated heuristics to circumvent anti-spam measures and to exclude general group email addresses from consideration. Several participants report however that they had missed about half of the candidates that were found relevant in the assessments (with correspondingly lower effectiveness).

To validate the outcome of the experiments, we asked one science communicator to look into the highly-ranked non-relevant responses, and classify those as follows:

E: Expert, but not key contact

K: Knowledgable, but not expert

N: Not knowledgable or expert

S: Science Communicator

U: Unknown status

None of these responses has been reconsidered as a 'key contact' missing from the topic definition. For three topics authored by this science communicator, we found that the systems identified five different science communicators (S) as the experts. Two of the ranked experts were deemed knowledgeable staff members but not experts (K), and four clearly not knowledgeable (N). The remaining twenty-eight highly-ranked non-relevant responses had unknown expertise (U).

We conclude from this minor investigation that the generic methods of expert identification are not taking into account the context of the situated task - science communicators created the topic set, and would not have nominated themselves as the key contact.

4 Summary

The third year of the enterprise track has introduced the CERC corpus (CSIRO Enterprise Research Collection). The data consists of a crawl of the public-facing web of the Australian Commonwealth Scientific and Industrial Research Oganization (CSIRO). The track involved CSIRO's science communicators in the topic development process, with the goal to model accurately the search activities of real members of the enterprise.

The newly introduced document search task is motivated by a 'missing overview page' scenario, where a search is conducted to find a set of 'key pages' related to the topic in question; for example, to assist the science communicator to create the missing overview page. The topics provided a small number of example 'key pages' to facilitate experiments with relevance feedback strategies.

The expert search task follows naturally from the missing page scenario, where the 'key contacts' among CSIRO staff should be identified. As opposed to previous years, the 2007 expert search task did not provide a pre-defined list of candidates, and fewer experts were expected per topic. The expertise judgments originate from the topic authors themselves, and encode inside knowledge. For example, highly-ranked non-relevant candidate experts for some topics turned out to be science communicators and other knowledgeable people that are not seen as experts.

References

P. Bailey, N. Craswell, I. Soboroff, and A.P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference*, Singapore, July 20–24 2008. To appear.