Overview of the TREC-2005 Enterprise Track

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 — intranet pages, email archives, document repositories — that reflect the experiences of users in real organisations, such that for example, an email ranking technique that is effective here would be a good choice for deployment in a real multi-user email search application. This involves both understanding user needs in enterprise search and development of appropriate IR techniques.

The enterprise track began this year as the successor to the web track, and this is reflected in the tasks and measures. While the track takes much of its inspiration from the web track, the foci are on search at the enterprise scale, incorporating non-web data and discovering relationships between entities in the organisation.

Obviously, it's hard to imagine that any organisation would be willing to open its intranet to public distribution, even for research, so for the initial document collection we looked to an organisation that conducts most if not all of its day-to-day business on the public web: the World Wide Web Consortium (W3C). The collection is a crawl of the public W3C (*.w3.org) sites in June 2004. It is not a comprehensive crawl, but rather represents a significant proportion of the public W3C documents. It comprises 331,037 documents, retrieved via multithreaded breadth-first crawling. Some details of the corpus are in Table 1.

The majority of the documents in this collection are email, and thus the tasks this year focus on email. Note that the documents are not in native formats, but are rendered into HTML.

There are two tasks with a total of three experiments:

- Email search task: Using pages from `lists.w3.org`.
  - Known item experiment: 125 queries. The user is searching for a particular message, enters a query and will be satisfied if the message is retrieved at or near rank one. There were an additional 25 queries for use in training.
  - Discussion search experiment: 59 queries. The user is searching to see how pros and cons of an argument/discussion were recorded in email. Their query describes the topic, and they care both whether the results are relevant and whether they contain a pro/con. There were no training queries, and indeed no judgements prior to submission.

Table 1: Details of the W3C corpus. Scope is the name of the subcollection and also the hostname where the pages were found, for example lists.w3.org. The exception is the subcollection 'other' which contains several small hosts.

Type	Scope	Size (GB)	Docs	avdocsize (KB)
Email	lists	1.855	198,394	9.8
Code	dev	2.578	62,509	43.2
Web	www	1.043	45,975	23.8
Wiki web	esw	0.181	19,605	9.7
Misc	*other*	0.047	3,538	14.1
Web	people	0.003	1,016	3.6
	all	5.7	331,037	18.1

- Expert search task: 50 queries. Given a topical query, find a list of W3C people who are experts in that topic area. Finding people, not documents, based on analysis of the entire W3C corpus. Participants were provided with a list of 1092 candidate experts for use on all queries. There were 10 training queries.

2 Email search task

This task focuses on searching the 198,394 pages crawled from lists.w3.org. These are html-ised archives of mailing lists, so participants can treat it as a web/text search, or they can recover the email structure (threads, dates, authors, lists) and incorporate this information in the ranking. Some participants made their extracted information available to the group.

In the known item search experiment, participants developed (query, docno) pairs that represent a user who enters a query in order to find a specific message (item). Of the 150 pairs developed, 25 were provided for training and 125 were used for the evaluation reported here. Results are in Table 2. The measures for this task were the mean reciprocal rank (MRR) of the correct answer, and the fraction of topics with the correct answer somewhere in the top 10 ("Success at 10" or S@10). Also reported is the fraction of topics that found the correct answer anywhere in the ranking (S@inf). In recent Web Track homepage finding experiments, it was possible to find the correct homepage with $MRR > 0.7$ and $S@10 \simeq 0.9$. Known item email search results are quite good for a first year, being about 0.1 lower on both metrics.

Nearly every group took a different approach at integrating the email text with email metadata and the larger thread structure. To give some examples, University of Glasgow (`uog`) combined priors for web-specific features — anchor text, titles of pages — with email-specific priors — threads and dates in messages and topics [7]. Microsoft Cambridge (`MSRC`) used their fielded BM25 with message fields, text, and thread features [4]. CMU (`CMU`) mixed language models for individual messages, message subjects, threads, and subthreads, and used thread-depth priors [8]. While the initial results are encouraging, it's clear that with this many types of data to balance, more work remains to be done.

Run	MRR	S@10	S@inf
uogEDates2	**0.621**	0.784	0.920
MSRCKI5	0.613	0.816	**0.952**
covKIRun3	0.605	0.792	0.896
humEK05t3l	0.604	0.808	0.912
CMUnoPS	0.601	0.816	0.912
CMUnoprior	0.598	**0.824**	0.912
qdWcEst	0.579	0.792	0.920
priski4	0.551	0.728	0.896
KITRANS	0.536	0.728	0.880
WIMent01	0.533	0.784	0.912
csiroanuki5	0.522	0.776	0.888
UWATEntKI	0.519	0.712	0.888
csusm2	0.510	0.712	0.792
qmirkidtu	0.367	0.600	0.768
LPC5	0.343	0.480	0.504
PITTKIA1W8	0.335	0.496	0.808
LMplaintext	0.326	0.544	0.704
DrexelKI05b	0.195	0.376	0.624

Table 2: Known item results, the run from each of the 17 groups with the best MRR, sorted by MRR. The best in each column is highlighted. (An extra line was added to show the run with best S@10.)

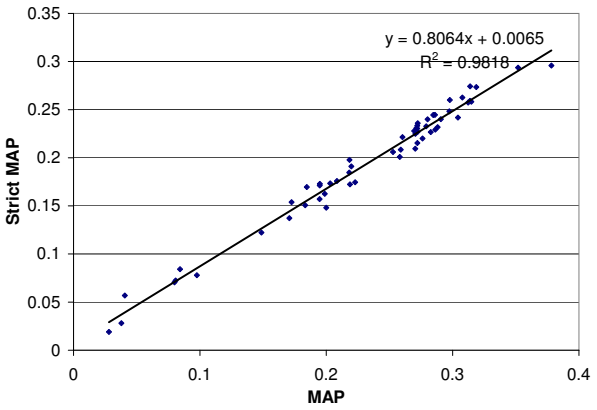

Figure 1: MAP for the 57 discussion search runs, calculated by conflating the top two (MAP) or bottom two (Strict MAP) judging levels.

In the discussion search experiment, participants developed topic descriptions and performed relevance judgements as described in Section 4. There are three types of answers: irrelevant, relevant without pro/con statement (also called "partially relevant") and relevant with pro/con statement. Table 3 shows discussion search results where any document that is not judged irrelevant is relevant (conflating the two positive judging levels). Interestingly, the top two runs are significantly better than the rest on our main measure mean average precision (MAP). For TITLETRANS, this is primarily due to the influence of a single topic [6]. The table also reports several other measures: R-precision (precision at rank R, where R is the number of relevant documents for that topic), bpref [2], precision at ranks (5, 10, 20, 30, 100, 1000), and reciprocal rank of the first relevant document retrieved.

Table 4 shows similar results if we now conflate the lower two judging levels, giving a 'strict' evaluation that only counts documents that include a pro/con statement as relevant. The overall rankings of systems are nearly identical, with a Kendall's tau of 0.893. Figure 1, shows a scatter plot, with the two types of MAP being strongly correlated.

The common focus of most groups in the discussion search subtask was how to effectively exploit thread structure and quoted material. University of Maryland (TITLETRANS in table 3 and 4) explored expanding documents using threads and the trade-off between reinforcing quoted passages and removing them altogether, with mixed results [6]. University of Amsterdam (ToNsBs) applied a straightforward language model with a filter to eliminate non-email documents [1]. Microsoft Research's (MSRC) best-performing run used only textual fields from the messages and no static features (year of message, number of parents in the thread) [4]. So it seems that these results represent mostly topic-relevance retrieval effectiveness, and we have not yet found definitive solutions to discussion search.

An important point raised by the University of Maryland team is that some of the topics did not necessarily lend themselves to pro/con arguments on the subject. Additionally, while the relevance judgements do indicate whether a pro/con argument is present in the message, we did not collect whether the argument was for or against the subject. They also found that some topics were not only more amenable to pro/con discussions, but also exhibited greater agreement between assessors. For the 2006 track, we plan to focus more closely on the topic creation process.

3 Expert search task

In the expert search task, participants could use all 331,037 documents in order to rank a list of 1092 candidate experts. This could involve creating a document for each candidate and applying simple IR techniques, or could involve natural language processing and information extraction technologies targeted at different document types such as email. Results are presented in Table 5.

For this year's pilot of this task, the search topics were so-called "working groups" of the W3C, and the experts were members of these groups. These ground-truth lists were not part of the collection but were located after the crawl was performed. This enabled us to dry-run this task with minimal effort in creating relevance judgments.

Top-scoring runs used quite advanced techniques:

THUENT0505 This run makes use of all w3c web part information and Email lists (the list part) together with inlink anchor text of these files. Text content are

Run	MAP	r-prec	bpref	P@5	P@10	p@20	p@30	P@100	P@1000	RR1
TITLETRANS	**0.3782**	**0.4051**	**0.3781**	0.5831	0.5000	0.4246	0.3712	**0.2427**	**0.0469**	0.7637
ToNsBs350F	0.3518	0.3769	0.3588	0.5729	**0.5407**	**0.4449**	**0.3768**	0.2147	0.0439	**0.7880**
UwatEntDSq	0.3187	0.3514	0.3266	0.5153	0.4831	0.4034	0.3610	0.2244	0.0415	0.6860
csiroanuds1	0.3148	0.3597	0.3310	0.5593	0.5102	0.4051	0.3469	0.2037	0.0416	0.7292
MSRCDS2	0.3139	0.3583	0.3315	**0.5864**	0.5169	0.4127	0.3475	0.1966	0.0428	0.7423
irmdLTF	0.3138	0.3461	0.3318	0.5254	0.4797	0.4169	0.3729	0.2183	0.0409	0.7249
prisds1	0.3077	0.3393	0.3294	0.5797	0.4966	0.3881	0.3277	0.1815	0.0381	0.6617
du05quotstrg	0.2978	0.3431	0.3163	0.5288	0.4712	0.3881	0.3362	0.2047	0.0417	0.6793
qmirdju	0.2860	0.3202	0.3017	0.5119	0.4695	0.3788	0.3226	0.1976	0.0421	0.7026
LMlam08Thr	0.2721	0.3062	0.2884	0.3932	0.3746	0.3263	0.2887	0.1819	0.0412	0.5678
PITTDTA2SML1	0.2184	0.2494	0.2333	0.3864	0.3271	0.2712	0.2288	0.1339	0.0290	0.4759
MU05ENd5	0.2182	0.2655	0.2530	0.4407	0.3831	0.3136	0.2893	0.1819	0.0381	0.6121
NON	0.0843	0.1305	0.1082	0.2576	0.2237	0.1771	0.1508	0.0869	0.0087	0.4123
LPC1	0.0808	0.0981	0.0907	0.2237	0.1746	0.1305	0.1062	0.0544	0.0072	0.3670

Table 3: Discussion search: Evaluation where judging levels 1 and 2 are ‘relevant’. Lists the run with best MAP from each of the 14 groups, sorted by MAP. The best in each column is highlighted.

Run	MAP	r-prec	bpref	P@5	P@10	p@20	p@30	P@100	P@1000	RR1
TITLETRANS	**0.2958**	0.3064	**0.3381**	0.3661	0.3356	0.2797	0.2429	**0.1531**	**0.0279**	0.5710
ToNsBs350F	0.2936	**0.3065**	0.3286	0.4068	**0.3763**	**0.2907**	0.2407	0.1292	0.0256	0.6247
MSRCDS2	0.2742	0.2892	0.3043	**0.4339**	0.3661	0.2864	0.2282	0.1200	0.0253	**0.6376**
UwatEntDSq	0.2735	0.2990	0.3086	0.3593	0.3220	0.2669	0.2373	0.1388	0.0250	0.5612
prisds1	0.2626	0.2803	0.2977	0.4000	0.3407	0.2695	0.2232	0.1136	0.0237	0.5234
du05quotstrg	0.2600	0.2837	0.2883	0.3864	0.3356	0.2576	0.2226	0.1246	0.0246	0.5436
irmdLTF	0.2592	0.2712	0.2852	0.3966	0.3407	0.2881	**0.2514**	0.1464	0.0247	0.5890
csiroanuds1	0.2583	0.2854	0.3000	0.3864	0.3492	0.2712	0.2243	0.1253	0.0253	0.5791
qmirdju	0.2446	0.2750	0.2841	0.3492	0.3153	0.2568	0.2085	0.1236	0.0248	0.5673
LMlam08Thr	0.2153	0.2442	0.2409	0.2576	0.2390	0.2068	0.1836	0.1149	0.0254	0.4369
PITTDTA2SML1	0.1978	0.2072	0.2165	0.2949	0.2508	0.1907	0.1565	0.0868	0.0176	0.4110
MU05ENd5	0.1847	0.2262	0.2309	0.3322	0.2627	0.2136	0.1989	0.1214	0.0230	0.5518
NON	0.0842	0.1285	0.1099	0.1864	0.1678	0.1280	0.1040	0.0568	0.0057	0.3061
LPC1	0.0724	0.0872	0.0811	0.1661	0.1220	0.0873	0.0723	0.0369	0.0050	0.3012

Table 4: Discussion search: Strict evaluation, where only judging level (includes a pro/con statement) is considered relevant. Lists the run with best MAP from each of the 14 groups, sorted by MAP. The best in each column is highlighted.

reconstructed and formed description files for each candidate person. Structure information inside web pages was also used to improve performance. Words from important pages are emphasised in this run. Bigram retrieval was also applied [5].

MSRA054 The basic model plus cluster-based reranking. (The basic model, 1) a two-stage model of combining relevance and co-occurrence 2) the co-occurrence model consists of body-body, title-author, and title-tree submodels 3) a back-off query term matching method which prefers exact match, then partial match, and finally word-level match.) [3]

This suggests that there were gains in effectiveness to be had via leveraging the heterogeneity of the dataset and the 'information extraction' flavor of the task. On the other hand, some groups (including THU and others) did notice that the search topics were W3C working groups, and took advantage of this fact by mining working group membership knowledge out of the collection. Thus, these results should be considered preliminary pending a more realistic expert search data set.

4 Judging

Since each known item topic is developed with a particular message in mind, that message is by definition the only answer needed, so no further relevance judging is required. However, in a corpus with significant duplication, it may be necessary to examine the pool for duplicates or near-duplicates of the item, as in the Web and Terabyte tracks. This year, because we do not believe that duplication is such a problem in `lists.w3.org`, we decided to expend effort in duplicate identification, so each query has exactly one answer.

Similarly, there was no judging required for the expert search task. This is because we used working group membership as our ground truth, as described in Section 3.

For the discussion search task, the judging was more involved. Because it is an ad hoc search task, it needs true relevance judgments, but the technical nature of the collection meant that NIST assessors would not be ideal topic creators or relevance judges. Instead, track participants both created the topics and judged the pools to determine the final relevance judgments.

In response to a call for participation in April, thirteen groups submitted candidate topics for the discussion search and known item tasks. For the known item search task, the topics included the query/name for the page and the target docno. For discussion search, the topic included a "query" field (equivalent to the traditional "title" field) and a "narrative" field to delineate the relevance boundary of the topic. In all, 63 topics were submitted, and NIST selected 60 topics for the final set.

Judging was done over the internet using an assessment system at CWI. Each topic was assigned to two groups, the group who authored the topic (the primary assessor) and another group (the secondary assessor). Secondary assessment assignments were made so as to balance authors across judging groups and to somewhat limit overall judging load. The topics and judging groups are shown in table 6. One group created three topics (24, 27, and 46) but did not submit any runs or respond to requests to help judge; their topics were reassigned to groups A, B, and C respectively as primary judges. Groups M and N did not contribute topics but did submit runs and agreed to help judge as secondary assessors. The pools were intentionally kept small to reduce the judging burden on sites. Three runs from each group were pooled to a depth of 50, and the final pools contained between 249 and 865 documents (mean 529).

Judging began in August and ran through early October, and was extremely successful, with all but three topics fully judged by their primary assessor, and 52 by the secondary assessor. The official qrels set consists of the primary judgments for 56 topics, and the secondary judgments for the remaining topics (26, 53, and 57). No relevant documents were found by the primary assessor for topic 4, and so we have left this topic out. This qrels set contains 31,258 judgments: 27,813 irrelevant, 1,441 relevant non-pro/con (R1) and 2,004 relevant pro/con (R2) messages. Median per topic was 14 for R1 and 20 for R2.

At the time of this writing, we have done some examination of the affects of assessor disagreement, by comparing the ranking of systems according to the primary and secondary judgments. For this experiment, we considered the 48 topics for which judgments exist from both assessors (and again dropping topic 4). Comparing the rankings of systems using each set of judgments yields a Kendall's tau of 0.763, which is less than the level of 0.9 taken to indicate "essentially identical", but still signifi-

Run	MAP	r-prec	bpref	P@5	P@10	P@20	P@30	P@100	P@1000	RR1
THUENT0505	**0.2749**	**0.3330**	0.4880	**0.4880**	**0.4520**	**0.3390**	**0.2800**	0.1142	0.0114	**0.7268**
MSRA054	0.2688	0.3192	**0.5685**	0.4080	0.3700	0.3190	0.2753	0.1306	**0.0131**	0.6244
MSRA055	0.2600	0.3089	0.5655	0.3920	0.3580	0.3150	0.2733	**0.1308**	**0.0131**	0.5832
CNDS04LC	0.2174	0.2631	0.4299	0.4120	0.3460	0.2820	0.2240	0.0942	0.0094	0.6068
uogES05CbiH	0.1851	0.2397	0.4662	0.3800	0.3160	0.2600	0.2133	0.1130	0.0113	0.5519
PRISEX3	0.1833	0.2269	0.4182	0.3440	0.3080	0.2530	0.2087	0.1026	0.0103	0.5614
uams05run1	0.1277	0.1811	0.3925	0.2720	0.2220	0.2000	0.1753	0.0944	0.0094	0.4380
DREXEXP1	0.1262	0.1743	0.3409	0.3120	0.2500	0.1760	0.1467	0.0720	0.0072	0.4635
LLEXemails	0.0960	0.1357	0.2985	0.2000	0.1860	0.1530	0.1213	0.0628	0.0063	0.4054
qmirex4	0.0959	0.1511	0.2730	0.2360	0.1880	0.1390	0.1233	0.0534	0.0053	0.4189

Table 5: Expert search results, the run from each of the 9 groups with the best MAP, sorted by MAP. The best in each column is highlighted. (An extra line was added to show the run with best P@100.)

Group	Authored topics							Assigned topics					Total
A	7	8	33	41	52	24		12	25	48	60		10
B	4	37	43	51	60	27		13	26	49			9
C	6	11	20	34	48	46		14	37	50			9
D	9	19	58					1	15	27	38	51	8
E	3	15	23	31	35			2	16	28	39	52	10
F	5	10	14	16	36			3	17	29	40	53	10
G	1	2	25	26	53			4	18	41	54		9
H	39	40	50	56				5	19	30	42	55	9
I	18	30	45					6	31	36	43	56	8
J	12	32	47	55	57			7	20	44	46		9
K	22	29	38	42	49			8	21	32	45		9
L	13	17	21	28	44	54	59	9	22	33	57		11
M								10	23	34	47	58	5
N								11	24	35	59		4

Table 6: Topic assignments for relevance assessment. “Authored topics” were created by that group. “Assigned topics” were assigned to that group by NIST for judging.

cantly correlated ($p < 2.2 \cdot 10^{16}$). We intend to look more closely at this data to see if particular topics or assessors cause more variation in the ranking.

5 Conclusion

This year participants made heavy use of email structure and combination of evidence techniques in email search and expert search with some success, but there remains much to learn. In future enterprise search experiments it would be nice to further our exploration of novel data types such as email archives, and of novel tasks such as expert search. This might include incorporation of a greater amount of real user data (perhaps query and click logs) to enhance our focus on enterprise user tasks.

For discussion search, we plan to approach topic creation with more care. Specifically, next year's topics will more closely target pro/con discussions, and we may ask assessors to label messages as either pro, con, both, or can't tell.

This year's foray into community-developed topics and relevance judgments marked a significant change for TREC, although such is the practise in other forums such as INEX. It has been a very successful experience, and we intend to continue collection development this way next year.

Task details for this year are maintained on the track wiki, at `http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page`.

Acknowledgements

We are grateful to the World Wide Web Consortium for allowing us to make a snapshot of their site available as a research tool. We also thank the University of Glasgow for hosting a Terrier-based search interface to the W3C collection for topic development. Lastly, we thank the participants of the 2005 enterprise track for helping to create the test collection.

References

[1] Leif Azzopardi, Krisztian Balog, and Maarten de Rijke. Language modeling approaches for enterprise tasks. In Voorhees and Buckland [9].

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 25–32, Sheffield, UK, July 2004. ACM Press.

[3] Yunbo Cao, Jingjing Liu, Shenghua Bao, and Hang Li. Research on expert search at enterprise track of TREC 2005. In Voorhees and Buckland [9].

[4] Nick Craswell, Hugo Zaragoza, and Stephen Robertson. Microsoft Cambridge at TREC-14: Enterprise track. In Voorhees and Buckland [9].

[5] Yupeng Fu, Wei Yu, Yize Li, Yiqun Liu, Min Zhang, and Shaoping Ma. THUIR at TREC 2005: Enterprise track. In Voorhees and Buckland [9].

[6] Jimmy Lin, Eileen Abels, Dina Demner-Fushman, Douglas W. Oard, Philip Wu, and Yejun Wu. A menagerie of tracks at maryland: HARD, enteprise, QA, and genomics, oh my! In Voorhees and Buckland [9].

[7] Craig Macdonald, Ben He, Vassilis Plachouras, and Iadh Ounis. University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with terrier. In Voorhees and Buckland [9].

[8] Paul Ogilvie and Jamie Callan. Experiments with language models for known-item finding of e-mail messages. In Voorhees and Buckland [9].

[9] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2004.