Overview of the TREC 2006 Enterprise Track

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

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

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

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 organizations, 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 in TREC 2005 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 organization. As a result, we have created the first test collections for multi-user email search and expert finding.

This year the track has continued using the W3C collection, a crawl of the publicly available web of the World Wide Web Consortium performed in June 2004. This collection contains not only web pages but numerous mailing lists, technical documents and other kinds of data that represent the day-to-day operation of the W3C. Details of the collection may be found in the 2005 track overview (Craswell et al., 2005). Additionally, this year we began creating a repository of information derived from the collection by participants. This data is hosted alongside the W3C collection at NIST.

There were two tasks this year, email discussion search and expert search, and both represent refinements of the tasks initially done in 2005. NIST developed topics and relevance judgments for the email discussion search task this year. For expert search, rather than relying on found data as last year, the track participants created the topics and relevance judgments. Twenty-five groups took part across the two tasks.

2 Email discussion search task

This task focuses on searching the `lists` subcollection, which are 198,394 pages crawled from `lists.w3.org`, the archive of the W3C mailing lists. Each page contains either a single email or a monthly listing. The messages are rendered into HTML, 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.

One can imagine many different kinds of searches in a mailing list archive. We have focused on searching for discussions and arguments about design and development issues within the W3C.

pop-up ads rely upon javascript to "pop up" OnLoad - that is, when the requested document is parsed by the user agent...since the "pop up" is part of the user interface, if a site employing pop-up ads claims conformance to WCAG, then the markup employed in pop-up adds are also subject to WCAG, while control over the popping is addressed by the User Agent Accessibility Guidelines (UAAG)

no matter the source of the content that pops up, if the site which utilizes pop-up ads does not ensure that the pop ups are WCAG compliant, then that site, or the document to which the OnLoad event that causes a new viewport to be generated is attached (if the claim is document-specific) cannot be considered WCAG compliant... for starters, pop-up ads are not rendered by non-javascript-aware browsers, such as lynx, which means that some users do not have access to all of the content on the page/site – regardless of whether that content is useful...

moreover, as david p has pointed out, turning off scripting in order to suppress the generation of pop-up ads is far too draconian a solution –

Figure 1: Part of an email arguing against the usability of pop-up ads. Note that the topic (DS64) is about pop-up ads, software to block them, and their relative advantages and disadvantages.

Over the course of their standards work, many decisions are made, sometimes after considerable and perhaps contentious debate. In the discussion search task, the goal of systems is to find those discussions, and in particular those messages where different sides of the debate are argued.

2.1 Topics and relevance judgments

In the first year of the track, the topics and relevance judgments for the discussion search task were created by the participants. This was not only due to limited resources at NIST, but primarily because it was thought that the technical nature of the collection was not well-matched to NIST assessor expertise. The experience of developing the collection within the community led us to reconsider this assumption, and so this year NIST assessors developed the topics and made the relevance judgments.

NIST developed fifty topics each of which describe a subject of discussion on the W3C mailing lists. These topics range from differences in the P3P 1.0 and 1.1 recommendations to blocking pop-up windows to evaluating color contrast for color-blind users. Participants were to search for on-topic emails that contain a pro or con argument. For example, a message relevant to the pop-up blocking topic with a negative argument is shown in Figure 1.

An important part of this task is developing an understanding of the kinds of searches that people would like to make in this collection. Wu et al. arranged last year's topics into several general categories, and observed that some categories were more amenable to pro/con discussion, and also that some categories had better inter-assessor agreement (Wu et al., 2006). This year, the assessors followed Wu's categories in designing their topics, and tried to ensure that pro/con discussion existed for that topic in the collection.

In addition to judging whether a message was irrelevant, on topic, or contained a pro/con argument, we also asked the assessors to try to note specifically whether the message was pro or con. Sentiment and relevance are denoted in the relevance judgments according to the following scale:

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
THUDSTHDPFSM	**0.2858**	**0.3186**	**0.3007**	0.4261	0.4022	0.3674	0.6415
srcbds5	0.2852	0.3179	0.2979	**0.4478**	**0.4370**	**0.3913**	0.6323
DUTDS3	0.2808	0.3110	0.2958	0.4304	0.4022	0.3522	**0.6483**
UAmsPOSBase	0.2590	0.3054	0.2743	0.4174	0.3826	0.3435	0.6028
york06ed03	0.2482	0.3141	0.2838	0.4348	0.3978	0.3620	0.5900
UMaTDMixThr	0.2316	0.2824	0.2539	0.3609	0.3478	0.3413	0.5051
IIISRUN	0.2269	0.2720	0.2442	0.3609	0.3217	0.3152	0.5328
IBM06JAQ	0.2030	0.2481	0.2337	0.3826	0.3391	0.3315	0.5992
uwTsubj	0.1891	0.2404	0.2136	0.3043	0.2913	0.2696	0.4285
InsunEnt06	0.1223	0.2004	0.1543	0.3304	0.3000	0.2652	0.5391

Table 1: Discussion search results for the run with the highest MAP from each group. Scores are computed where judging levels '2' (contains a pro/con) and above are considered relevant. The best score for each measure is highlighted. DUTDS3 is a manual run.

0: not relevant.
1: relevant, does not contain a pro/con argument.
2: relevant, contains a negative (con) argument.
3: relevant, contains both pro and con arguments.
4: relevant, contains a positive (pro) argument.

A 10% random sample of each topic's pool was drawn and given to a second assessor in order to measure agreement. Agreement within the sample was similar to levels found in last years relevance judgments as reported in (Wu et al., 2006). When judgments were thresholded so we could measure agreement on whether a message was relevant at all or not, we find a Cohen's kappa of 0.4. Agreement on whether a message was pro/con as opposed to relevant or nonrelevant was 0.35. The sample was not large enough to measure agreement on pro or con messages alone. Relevance judgments for retrieval tasks tend to have a kappa of around 0.4 (varying somewhat between collections and assessor groups), so these values while low are not unusual.

2.2 Results

Runs were evaluated on retrieval of messages containing a pro/con sentiment (levels 2 and above) as well as just retrieving relevant messages (levels 1 and above). Table 1 shows the top run from each group according to mean average precision in retrieving pro/con messages. Table 2 shows the top run from each group for topic relevance retrieval.

Figure 2 compares the MAP scores between the two rankings. Overall, the two rankings of the runs are very similar, with a Kendall's tau of 0.9 for MAP. Three runs, DUTDS3, york06ed02, and IBM06JAQ, are more highly ranked at pro/con retrieval than they are at relevant message retrieval. DUTDS3, a manual run (i.e., a person was involved in some stage of the query processing), is the eleventh-highest ranked run by MAP on relevant messages, but the third-highest ranked by MAP on pro/con messages.

The strong tau correlation indicates although the runs are trying to focus on pro/con messages, topic relevance is the dominant factor in their document rankings. We further tried to determine if the relative ranking of pro/con and relevant messages was better or worse than random. For each topic in each run, we removed the nonrelevant retrieved documents, and computed the average precision of the residual ranking with only pro/con documents considered relevant. We call this "pro/con AP", and it equals 1 when all pro/con messages are ranked

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
THUDSTHDPFSM	**0.4083**	0.4204	**0.4264**	0.6520	0.6120	0.5590	0.7702
srcbds5	0.4065	**0.4275**	0.4222	0.6520	0.6100	0.5610	0.7917
DUTDS4	0.3891	0.4048	0.4062	0.6000	0.5780	0.5310	0.7004
york06ed03	0.3782	0.4195	0.4128	**0.6840**	**0.6180**	**0.5690**	**0.8048**
UAmsPOSBase	0.3750	0.3991	0.3943	0.6280	0.5920	0.5350	0.7776
UMaTDMixThr	0.3631	0.3963	0.3863	0.5880	0.5820	0.5470	0.7134
IIISRUN	0.3430	0.3769	0.3678	0.6080	0.5640	0.5130	0.7283
IBM06JILAPQD	0.3310	0.3717	0.3709	0.5800	0.5640	0.5040	0.7677
uwTsubj	0.2927	0.3377	0.3112	0.5000	0.4980	0.4590	0.5819
InsunEnt06	0.1872	0.2612	0.2125	0.4720	0.4520	0.4210	0.7085

Table 2: Discussion search results for the run with the highest MAP from each group. Scores are computed where judging levels '1' (relevant to the topic) and above are considered relevant. The best score for each measure is highlighted.

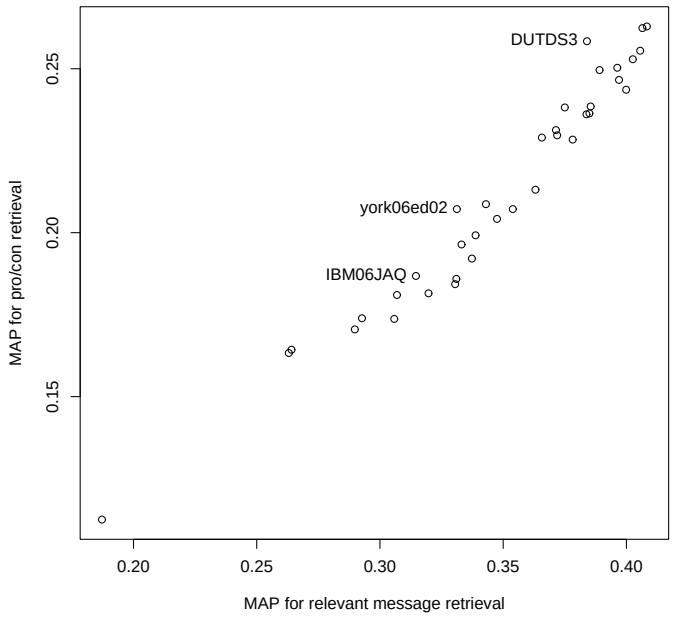

Figure 2: Scatterplot of MAP scores for pro/con and relevant message retrieval. The three labeled runs are ranked more highly for pro/con retrieval than for relevant message retrieval.

ahead of "just relevant" ones. We then generated 1000 random permutations of those pro/con and relevant documents, computing the pro/con AP of each permutation. Sorting the pro/con APs and noting the top and bottom 25 gives a 95% confidence interval on pro/con AP for that number of pro/con and relevant documents. If the actual pro/con AP is above the confidence interval, we would conclude that the run is significantly ordering pro/con documents ahead of relevant ones. Likewise, if the actual pro/con AP is below the interval, we would conclude that the ordering was worse than would be achieved by random shuffling.

Figure 3 presents these results both for each run and for each topic. The top graph shows the number of topics for each run that the actual pro/con AP was above, within, or below the 95% interval. The bottom shows for each topic the number of runs for which their actual pro/con AP was above, within, or below the interval. In each graph, the bar is divided into three sections: the top part counts the topics (or runs) where actual pro/con AP was above the interval, the middle those within the interval, and the bottom those below it. These graphs seem to indicate that most runs do not significantly differentiate relevant and pro/con messages for the majority of topics. Some topics are "easier" in this regard than others, but some are much much harder; note topic 62, where all runs actually ranked the relevant documents ahead of the pro/con ones.

We lastly compared the system ranking for relevant message retrieval to one based on the second assessor's relevance judgments. Since the secondary judgments are only a random sample, we used Yilmaz and Aslam's inferred average precision (infAP) (Yilmaz and Aslam, 2006) measure to estimate average precision for the runs using the sampled judgments. The Kendall's tau correlation of the official MAP ranking to the infAP ranking is 0.695. This is about the same as we saw in last year's judgments, when you consider that the use of a subsample also causes the correlation to be lower. Along with the similarity in agreement measures noted above, this indicates that assessor disagreement for this task is not very different and has about the same effect whether participants or NIST assessors are assessing relevance. We surmise that the lack of familiarity with the W3C and the collection affects all assessor groups strongly.

3 Expert search task

The expert search task is quite different from the traditional TREC search task, in that the goal of the search is to create a ranking of people who are experts in the given topic, rather than relevant documents about the topic. Nick Craswell extracted a canonical list of people addressed in email or on a web page in the W3C collection; this is called the *candidate list*. In response to a given topic, systems return a ranking of candidate experts. In contrast to the email search task, participants may make use of the entire W3C collection. Candidates are pooled and judged for expertise, and the systems are scored using traditional ranked retrieval measures.

The expert search task was the more popular in the track, with 23 groups contributing topics, runs, and/or relevance judgments. There were 91 runs submitted.

3.1 Topics and relevance judgments

In 2005, the enterprise track ran a pilot expert search task where the topics were W3C working groups, and systems were to identify who was part of each working group. The working group truth data came from an official listing of groups and members which was not part of the collection (although some groups were able to find the list by searching the live web). This year, we decided to develop topics for expert search from scratch.

As was done last year for email discussion search, the topics for the expert search task were created and judged by track participants. Twenty groups agreed to help, and each contributed 3-6 topics. Of these, we selected 55 topics for the final set. Once runs were submitted, NIST formed pools and sent them to CWI, where the assessment system was hosted. The topic authors

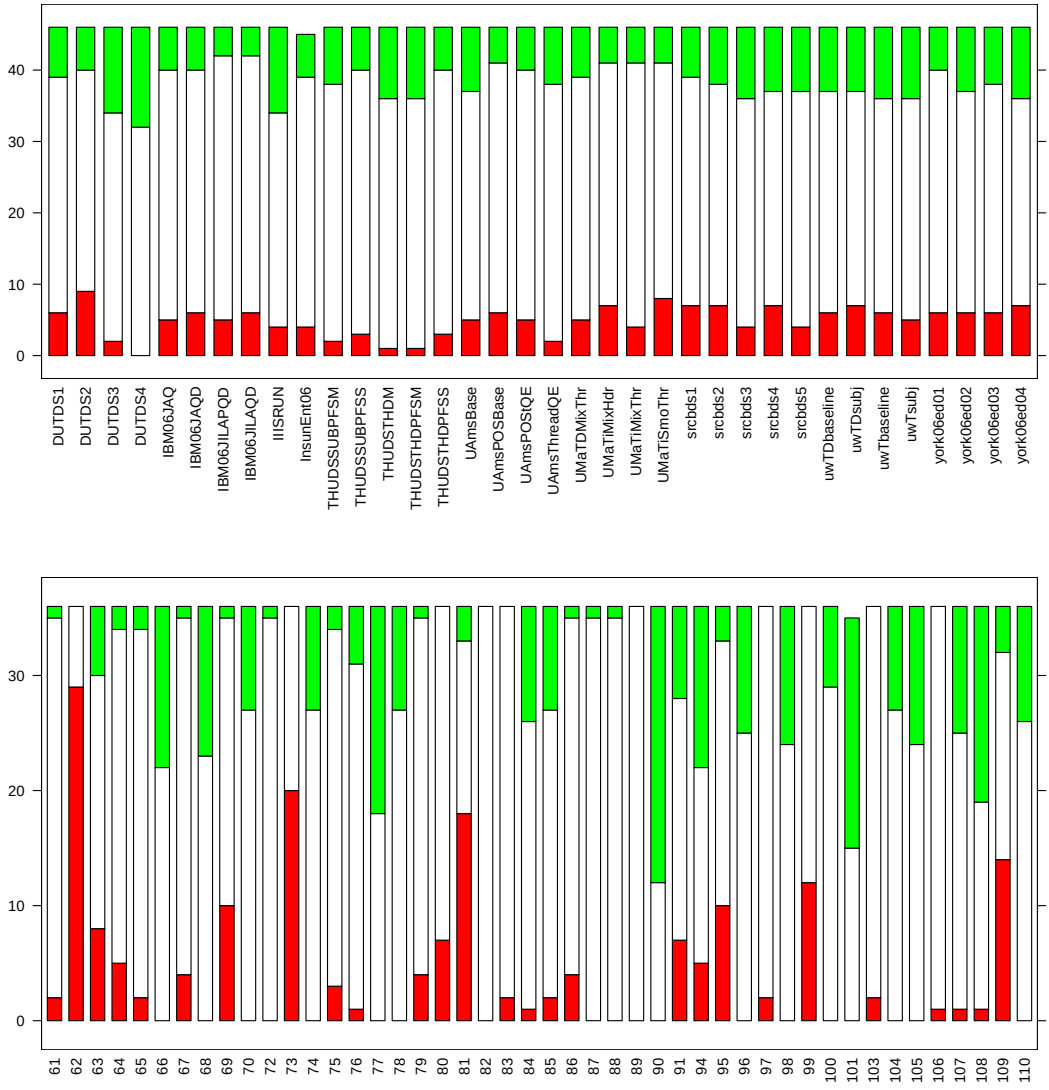

Figure 3: (Top) For each run, the number of topics where actual pro/con AP was better, equivalent to, or worse than random. (Bottom) For each topic, the number of runs whose actual pro/con AP in that topic was better, equivalent to, or worse than random.

Beijing University of Posts & Telecom.	Queen Mary University of London
California State University, San Marcos	Queensland University of Technology
Case Western Reserve University	Robert Gordon University
City University	Shanghai Jiao Tong University
DaLian University of Technology	Tsinghua University
Fudan University	University Amsterdam
University of Glasgow	University of Illinois Urbana-Champaign
IBM	University of Massachusetts
Lowlands Team	University of Waterloo
Open University	University Ulster and
University of Pittsburgh	St. Petersburg State University

Table 3: Groups contributing topics and judgments for the expert search task.

```
<top>
<num> Number: EX51
<title> relationship cardinalities </title>

<desc> Description:
A relevant expert will have knowledge in relationship cardinalities between
roles in different choreographies.
</desc>

<narr> Narrative:
In the context of semantic web, the relationships between entities can have
different cardinalities and roles. Relevant expert will have an explicit knowl-
edge of such choreographies. Experts in Semantic web are not relevant with-
out explicit knowledge in choreographies.
</narr>
</top>
```

Figure 4: A sample expert search task topic.

then judged the pools through the CWI system. We received judgments for 49 of the 55 topics. The names of the groups who contributed their considerable time and effort to this task are listed in Table 3.

A sample topic is shown in Figure 4. Note that this topic resembles a TREC ad hoc topic, except that the user is looking for people rather than documents. The topic statements were composed by the contributor, and only lightly edited to correct the spelling of key words and any ambiguous grammar.

Systems produced a ranked list of expert candidates for each topic. In addition, for each candidate, systems returned a (possibly zero-length) ranked list of documents supporting the designation of that person as an expert in the topic. The purpose of requiring support documents is twofold. First, in an actual application, it is important for the system be able to illustrate why a person is being recommended as an expert. Second, the groups making the relevance judgments could make use of the support documents in deciding whether a person was an expert, rather than doing their own research or relying on background knowledge.

The pools for expert search included the top 20 ranked people for each topic, along with the top 10 support documents for each of those people, from the two highest-priority runs per group. This created very large pools with 6,217 expert-document pairs per topic on average. Ideally,

all support documents are assessed before making a judgment on the candidate's expertise. However, considering the size of the pools, we decided to distinguish between *judged* and *partially judged* expert search topics. In a partially judged topic, the assessment of the candidate's expertise has not been done on the basis of judging all support documents, but using a handful of (positive or negative) support documents only (i.e., some of the pooled support documents are skipped). Unfortunately, making partial judgments did not reduce the workload very much - on average, assessors who judged expertise using partial judgments still made an assessment for more than one out of six support documents in the pool. We explore some possible ideas for reducing the judging load below in the discussion of results.

The relevance scales for expert search are somewhat unusual, to allow for the possibility of indeterminate expertise and support documents which in fact did not support a judgment of expertise either way. The scales used in the expert search relevance judgments were

- Candidate experts:

  0: candidate is not an expert.

  1: unknown.

  2: candidate is an expert.

- Support documents:

  0: negative support (document indicates person is not an expert).

  1: no support either way.

  2: positive support (document indicates person is an expert).

Note that the threshold for correctness for both people and documents is 2, rather than the usual value of 1.

3.2 Results

The evaluation results measure the quality of the ranked list of people using traditional retrieval measures including MAP and precision at fixed ranks. Two sets of measures were provided. The first measures the ranked expert list without regards to the support documents; if a correct expert is returned, the system is credited with returning that expert even if no supporting documents were retrieved. These results are shown in Table 4. Manual runs, where a person was involved at some point in the query process, are shown in italics.

The evaluation scores in Table 5 only gave credit for retrieving a relevant expert if a supporting document was retrieved as well. Credit was awarded if a supporting document appeared anywhere in the list of (maximum) 20 support documents for that person. If no supporting document was retrieved, the person was considered not relevant.

Figure 5 plots each run's no-support-required MAP score against its supported-experts MAP score. The tau correlation of the two rankings is only 0.76. Three runs from ICT did not return any support documents at all, and as such they are found along the x-axis in Figure 5; when we remove those runs from both rankings, the tau improves to 0.82. This is still low enough to indicate noticeable differences in the two rankings. The graph seems to show groups of runs with very closely-scoring expert rankings that differ in their supported-expert ranking. We need to look more closely at those runs that are returning unsupported experts to understand what is happening here more fully.

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
kmiZhu1	**0.6431**	**0.6242**	**0.6391**	**0.8245**	**0.7347**	**0.6031**	**0.9609**
SJTU04	0.5947	0.5783	0.5913	0.7673	0.7041	0.5694	0.9358
SRCBEX5	0.5639	0.5599	0.5642	0.7224	0.6551	0.5469	0.9043
PRISEXB	0.5564	0.5808	0.5614	0.7592	0.6653	0.5459	0.8486
IBM06MA	0.5235	0.5192	0.5180	0.7673	0.6449	0.4857	0.9286
UMaTDFb	0.5016	0.5108	0.5049	0.7265	0.6388	0.5000	0.8571
THUPDDSNEMS	0.4954	0.4978	0.4916	0.6694	0.5939	0.5071	0.8265
ICTCSXRUN01	0.4949	0.4977	0.4858	0.6898	0.5837	0.4908	0.8194
FDUSO	0.4814	0.4989	0.4936	0.7020	0.6306	0.5153	0.8612
UvAprofiling	0.4664	0.4957	0.4707	0.6612	0.5878	0.4959	0.8510
DUTEX2	0.3779	0.4175	0.4077	0.6245	0.5184	0.4184	0.8094
qutmoreterms	0.3673	0.4043	0.3907	0.6327	0.5388	0.4367	0.7683
UMDemailTLNR	0.3503	0.3775	0.3552	0.5388	0.5041	0.4245	0.7064
UIUCe2	0.3364	0.3580	0.3388	0.5388	0.4816	0.3959	0.7187
ex3512	0.3158	0.3425	0.3299	0.5347	0.4612	0.3898	0.7912
uwXSOUT	0.3132	0.3780	0.3364	0.5796	0.5143	0.4112	0.7140
uogX06csnQE	0.3024	0.3433	0.3292	0.5306	0.4429	0.3531	0.7831
PITTPHFREQ	0.2770	0.3513	0.3166	0.5510	0.5041	0.3857	0.7366
sophiarun1	0.2248	0.2864	0.2565	0.4980	0.4306	0.3286	0.6307
w1r1s1	0.2154	0.2818	0.2523	0.5184	0.4245	0.3265	0.6368
l3s2	0.1313	0.1480	0.1401	0.5714	0.2918	0.1459	0.8010
quotes	0.1308	0.1778	0.1844	0.3184	0.2653	0.2224	0.5095
SPlog	0.1126	0.1555	0.1671	0.2531	0.2204	0.1878	0.4709

Table 4: Expert ranking scores without taking support documents into account. The best run in each group according to MAP is shown. Runs in italics are manual runs.

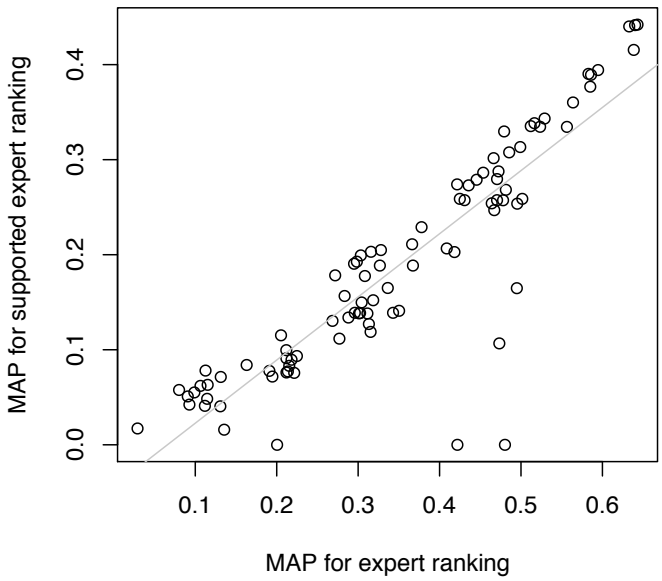

Figure 5: Scatterplot of MAP scores when support is or is not required when considering whether a retrieved person is relevant.

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
kmiZhu1	**0.4421**	**0.4835**	**0.4986**	**0.6612**	**0.5633**	**0.4459**	**0.8369**
SJTU04	0.3943	0.4304	0.4581	0.5714	0.5204	0.4143	0.8132
SRCBEX5	0.3602	0.4092	0.4299	0.5551	0.4735	0.3969	0.7350
IBM06MA	0.3346	0.3829	0.4135	0.5878	0.4878	0.3602	0.7339
PRISEXB	0.3345	0.4203	0.4228	0.5429	0.4571	0.3847	0.6695
UvAprofiling	0.3016	0.3637	0.3743	0.4980	0.4265	0.3582	0.7177
FDUSF	0.2796	0.3148	0.3356	0.4653	0.4041	0.3204	0.6767
UMaTiDm	0.2740	0.3205	0.3350	0.4980	0.4102	0.3204	0.6344
THUPDDFBS	0.2573	0.3035	0.3155	0.4082	0.3673	0.3020	0.6117
DUTEX2	0.2290	0.2918	0.3028	0.4898	0.3898	0.3031	0.6703
qutbaseline	0.2110	0.2561	0.2527	0.4082	0.3531	0.2694	0.6115
ex3512	0.2031	0.2466	0.2724	0.3959	0.3286	0.2786	0.6481
UIUCe2	0.1650	0.2271	0.2582	0.3143	0.2898	0.2347	0.5874
ICTCSXRUN01	0.1648	0.2338	0.2497	0.2857	0.2347	0.2143	0.4245
UMDemailTLNR	0.1410	0.2015	0.1997	0.3388	0.2980	0.2357	0.5561
uwXSHUBS	0.1389	0.2028	0.1938	0.3551	0.2878	0.2449	0.5185
uogX06csnQE	0.1387	0.2046	0.2180	0.3061	0.2551	0.2071	0.5430
PITTPHFREQ	0.1117	0.1843	0.1744	0.3143	0.2857	0.2031	0.5085
allbasic	0.0996	0.1479	0.1409	0.3020	0.2429	0.1786	0.5233
sophiarun1	0.0934	0.1415	0.1322	0.3184	0.2449	0.1582	0.4646
SPlog	0.0781	0.1179	0.1470	0.2000	0.1694	0.1347	0.4265
l3s2	0.0714	0.0827	0.0820	0.3429	0.1755	0.0878	0.5840
body	0.0484	0.0809	0.1004	0.1224	0.1122	0.0918	0.2606

Table 5: Expert ranking with retrieval of a correct supporting document required. Runs in italics are manual runs.

Number of pooled		Experts	Tau correlation	
Experts	Documents	per Topic	No support	Support required
20	10	30.1	0.96	0.99
20	5	27.9	0.96	0.98
20	1	21.6	0.93	0.94
10	10	22.5	0.95	0.97
10	5	20.5	0.94	0.96
10	1	16.0	0.90	0.92

Table 6: Comparison of system rankings using pools of 20 or 10 experts and 10, 5, or 1 support documents per pooled expert, using the tau correlation to the official ranking.

3.3 Reducing pool size

As indicated above, the inclusion of support documents for experts caused the pools to be very large. Using these pools, we can examine if equivalent evaluation results could be obtained with smaller pools. The judged pools included the top 20 retrieved experts and the top 10 retrieved support documents for each candidate expert. In the process of this experiment, we discovered a bug in the support document pooling. The outcome of this bug was that if an expert was in the pool, the top 20 support documents were pooled even from a run that did not retrieve that expert in its top 10. This increased the size of the pools by a factor of 1.5 on average, and it seems likely that most of those documents were not relevant, simply because they came from less effective runs. After correcting this error and re-creating the relevance judgments based only on what should have been pooled, we found that nearly all the relevant experts found in the official pools were still present in the corrected version. The tau correlation to the official results was 1.0 for unsupported MAP and 0.99 for supported MAP. Thus, we have not changed the official reported results based on the original relevance judgments.

Starting from these corrected pools, we further reduced them by taking only the top 1, 5, or 10 supporting documents, and similarly by taking the top 10 experts only and the corresponding top 1, 5, or 10 supporting documents. Since the expert judgments were presumably informed by the supporting documents, we could not just apply the original expert judgment in the reduced pools. Instead, we used the following heuristic: if a document supporting expertise was retained in the reduced pool, we judged the candidate an expert. Similarly, if the reduced pool contained a document judged as indicating that the candidate was not an expert, we judged the candidate to not be an expert. If both supporting and detracting documents were in the reduced pool, we retained the original assessor's judgment of expertise. If no supporting or detracting documents were in the reduced pool, the candidate's expertise was labeled unknown.

Table 6 compares the system rankings based on the relevance judgments from these reduced pools to the official rankings reported above. For both supported and unsupported MAP, all reduced relevance judgment sets provide system rankings that are nearly identical to the official ranking.

An important concern when using small pools is that runs that did not contribute to their creation may be unfairly scored, because these runs are more likely to have retrieved candidates and support documents that are not in the pool. To gauge this effect, within each set of reduced pools we held out each group's runs in turn and measured them using the relevance judgments that would have been produced if their group had not contributed. Again, this works as in the reduced pools themselves; candidates that are only found by the held-out group are left unjudged, and holding out a group's unique support documents can change a judgment of candidate expertise.

We consider both changes in score, as well as how a group would have been ranked differently

Number of pooled		No support			Support required		
Experts	Documents	max	min	rank	max	min	rank
20	10	+0.0048	-0.0639	+0/-15	+0.0044	-0.0641	+5/-15
20	5	+0.0026	-0.0061	+0/-16	+0.0046	-0.0751	+5/-15
20	1	+0.0043	-0.1269	+0/-22	+0.0069	-0.1084	+6/-26
10	10	+0.0138	-0.0443	+6/-12	+0.0156	-0.0550	+8/-14
10	5	+0.0173	-0.0581	+5/-15	+0.0123	-0.0579	+7/-14
10	1	+0.0394	-0.1116	+5/-21	+0.0258	-0.1060	+11/-26

Table 7: Changes in MAP score and rankings when groups' runs are left out of the pools. "max" and "min" are the maximum and minimum MAP score difference. "rank" gives the largest movements up and down in the ranking when a group's runs are held out.

had it not contributed to the pool. Table 7 shows these results. "max" and "min" show the maximum and minimum change in MAP score among held-out groups. "rank" shows the largest moves up and down in the ranking. For example, when the pool is reduced to contain only 10 candidates per run and a single support document per candidate (10-1), one run drops 26 places in the supported experts ranking of 91 runs when all runs from that group are held out of that reduced pool. Note that this large change indicates that most runs scored very closely together; a change of -0.1060 in MAP covers more than a quarter of the ranking.

These results seem to indicate that the pools can be significantly reduced and still adequately measure the pooled runs, but that some caution should be exercised to ensure that the judgments are reusable by groups that did not participate. Reducing down to a single support document has a very large effect, greater than pooling fewer experts. Ten candidates with five support documents each is probably reasonable.

4 Conclusion

The second year of the enterprise track was very successful. We built a second set of topics for searching for discussions in mailing lists. We have also built a test collection for expert search. Taken together, the enterprise track collections are the first of their kind. While we still need to study their stability and reusability, we hope they will be a valuable resource for researchers.

An important lesson we have learned is that it can be difficult to situate information needs within an organization when you are not actually part of that organization. The topics largely give the impression of someone on the outside looking in, perhaps representative more of a new member of an organization rather than a veteran. When we began the track, we were concerned that the technical nature of the organization would be the chief obstacle to topic development. Now with topics created both by TREC participants and by NIST assessors, we appreciate that the greater challenge is to think of the information needs that people inside the organization have.

To that end, the collection will change in TREC 2007. The collection will be a snapshot of CSIRO, the Australian Commonwealth Scientific and Industrial Research Oganization. More importantly, the topics will be developed by employees at CSIRO. This will result in a topic set that reflects the range of information needs found within the organization.

5 Approaches

The following are descriptions of the approach taken by different groups. These paragraphs were contributed by participants and are intended to be a road map to their papers in the TREC proceedings. Below each group name is a list of their runs submitted to each task.

5.1 Beijing University of Posts and Telecommunications

Expert: PRISEXB, PRISEXR, PRISEXRM, PRISEXRMT

Candidates are ranked by their relevant description files. Each description file is constructed with the words co-occurred with a candidate, i.e., in the same window of text, in a document. Support documents are also ranked according their corresponding description files. Special data structures like headword and email are also considered to improve performance.

5.2 Case Western Reserve University

Expert: allbasic, basic, w1r1s1

This was Case Western Reserve University's first participation in TREC. We participated in the expert search task of the enterprise track. Our motivation for participation was our work developing an expert search capability for a prototype vertical digital library, MEMS World Online (`memsworldonline.case.edu`). For the expert search task we mostly relied on the email list portion of the W3C corpus. The emails are likely to be the most accurate indicator of an individuals expertise. Additionally, we give higher weight to response emails, which are also likely to be good indicators of expertise. We also used an additional weighting factor which is related the expertise of the individual's closely related colleagues in the social network extracted from the corpus. This is based on the intuition that the experts of the same topic are likely to work closely together. Finally, we used WordNet for synonyms in one run, though we did not expect much from this because of the technical nature of the task topics. We did not do any significant file preprocessing and only used automatic queries.

5.3 Chinese Academy of Sciences – ICT

Discussion: IIISRUN

Expert: ICTCSXRUN01 – 05

In this year, our team's research and experiments mainly focus on the mail list corpus and the link relationship amongst the candidates expert and other users. The W3C corpus includes a large archive of the W3C's mail lists. These lists are email forums for people who want to share information about W3C's research and projects. We can treat these forums as social networks. In our experiments, we find some interesting features of the community structures of these networks: In most of the mail lists, the candidate experts are not well connected. The social network in these mail lists can be divided into some communities which includes a few candidate experts and a lot of other users. The candidate experts are mostly in the center of their communities. And also, we use some link analysis approaches to rank the candidates in the social networks. In our experiments, we choose the PageRank algorithm and a revised HITS algorithm as link analysis methods. These approaches gives satisfying results in our experiments.

5.4 City University

Expert: ex3512, ex5512, ex5518, ex7512

A naive string matching algorithm is used to extract the full name and email addresses of identified experts, using a fixed window size (of 2000 characters), in order to build a profile for

those experts. We then index these profiles using Okapi, and used BM25 to rank the experts to generate our results.

5.5 DaLian University of Technology

Discussion: DUTDS1 – 4
Expert: DUTEX1 – 4

For email discussion search, we first preprocessed the cleaned W3C collections based on which an index was built by Indri (or Lemur). Then we handled the query topic in the same way of cleaning the documents, i.e. stripping the special character and stopping word. Ultimately, relevant documents were retrieved by Indri (or Lemur).

For expert search, we first created a correlative document pool for each candidate from the cleaned W3C collections and then gained the expert list and the support document with the pool. In the stage of correlative document pool generation, firstly, we collected the identities of each candidate, including his name, email, phone, nick, personal main page and so on. There were two stages in this process, automatic and manual. In the automatic we made several rules for identity extraction combining the technique of named identity recognition, then adjust and recruit the result in the manual stage.After candidate identity extraction was finished, an index was built based on the cleaned W3C collections and utilized the candidate identities to query. We singled out a number of words around the candidate identity to form the correlative document pool.

In the stage of expert list and supporting document generation, an index was built based on the correlative pool firstly. We attempt to compose the query in several ways for each topic and introduced the query to the Indri. The expert list was gained through the retrieved Indri score.Different from lasts year, every retrieved expert should be provided with his supporting documents which can explain why the candidate is an expert in this subject. Accordingly, we dealt with the correlative document pool. We took the (document ID-candidate ID) as the supporting document ID, in this way the correlative document pool of a candidate was divided into some supporting documents. Then we added the candidate identities to the original query and utilized Indri to gain the supporting documents of the expert.

5.6 L3S, University of Hannover

Expert: l3s1 – 4

We performed experiments on Expert Search in scope of Enterprise Track 2006. We based our technique solely on W3C mailing lists. The main assumption was that the author of an email is an expert on the subject addressed by the email. We tested 4 different heuristics with different threshold on the document score as well as the expert score. Using set of data-driven thresholds on similarity values, we cut off different number of experts per each query. One finding of our experiments was that complexity of the information need does not correlate with the number of relevant experts returned by the system. It was an interesting result, since normally the more specific your question, the less experts you expect. This result should be investigated more carefully, since definition of the task specificity is somewhat vague. It would be interesting to agree on one common scheme for task specificity definition in the expert search community.=20
We also scheduled more experiments with additional dataset, which we are creating in our group. This dataset will include real world documents, publications and wiki pages. The difference with W3C collections is that it could be enhanced with specific expert search interface and will allow tracking user logs while searching experts in it.

5.7 Lowlands Team

Expert: MAPCrelTret, MAPTrelCret, SP, SPlog

The lowlands team worked on the expert search task. We experimented with directly comparing two sets of document rankings: one for topics one for candidates. For each candidate we produce a ranked list of the 1000 most relevant documents based on a name+email address query. For each topic we produce a separate ranked list of the top 1500 most relevant documents. The intuition is that candidates for whom the document ranking has a high correlation with the ranking based on a given topic are likely to be experts for that topic. Experiments with various ways of producing the candidate based rankings and various ways of computing the correlation, showed that with a good document ranking for the candidates, good results can be obtained independent of the correlation method used.

5.8 Open University

Expert: kmiZHU1, kmiZhu2, kmiZhu4, kmiZhu5

Our group have used a two-stage language modeling approach consisting of a document relevance model and a window-based co-occurrence model in expert search. Document relevance measures the relevance of a document to a topic, and the co-occurrence model measures the relevance of an expert to a topic. Boolean query, span query, BM25, and TF/IDF are used for document relevance. There are mainly three innovative points in our group's approach. First, document authority in terms of their PageRanks is taken into account in the document relevance model, and the assumption is that more authoritative documents are linked or referenced more often by the others. Second, document internal structure is considered in the co-occurrence model. The occurrence of an expert's name in different parts of a document has influence on judging his/her relevance to a topic. We used templates of documents to segment these documents and consider structures of various documents, e.g., technical report, emails, and research papers. Third, we used incremental window sizes in the co-occurrence model. In selecting window sizes, small windows often lead to more accurate associations between experts but may miss some of them, while large windows often cover more associations to compensate small windows but may introduce noise. We gave higher weights to small window based than large window based relevance and aggregate their relevance together. Window sizes can reflect from phrase level, sentence level up to document level associations. In addition to the three points, partial match of queries, query construction from description and narrative of topics, and query construction by domain experts were also studied.

5.9 Queen Mary University of London

Expert: body, listbq, quotes, www

For Enterprise TREC, our group tried a strategy which integrates information retrieval with database management techniques. We use a probabilistic framework that allows us to evaluate expert finding strategies expressed in probabilistic variants of SQL and Datalog. Documents in the ETREC collection are parsed into a relational representation, to aid the integration of IR and DBMS. For the identification of experts, we assumed that some parts of emails in the collection are better at discriminating experts than others. We used different runs to check this claim, using only quotations, only bodies, or the whole email text for expert finding, and compared the performance of these different strategies.

5.10 Queensland University of Technology

Expert: qutbaseline, qutlmv2, qutmoreterms

We have participated in the expert search using the Terrier search engine for topic based retrieval, and then post-processed the top 100 documents to identify the experts. The concept of an expert was identified through the frequency with which the expert appears in the top 100 documents (emails, news, standards or drafts). The heuristic is pretty straight forward – one would expect a higher frequency for an expert in publication, citation, email discussion, etc. Furthermore, the persons appearing in the W3C standards or drafts as editors or authors should be experts. We did not have an opportunity to refine the selection to take account of indicative context. We based our expert selection on frequency alone without any attention to context or other details. The performance of the system was quite reasonable considering its simplicity. The system outperformed the median score when measured over all topics, but was not quite competitive enough relative to the best topic scores although it got close for several topics.

5.11 Ricoh Software Research Ctr.

Discussion: srcbds1 – 5
Expert: SRCBEX1 – 5

We participated in expert search and discussion search of Enterprise Track in TREC 2006. In the discussion search, we take advantage of the redundant pattern of emails to parse them according to their data structure. The collected pieces of information are subsequently stored in XML format and include the subject part, author part, sent time part, content part, quoted message part, greeting part and ad part. As the words in different parts are known to have different semantic weight, we use the so-called Field-Based weighting method to find relevant documents. We not only consider content relationships between the query and the target document but also non-content features such as time-line, mail thread, author, category and quoted chain. Tests showed that these non-content features are effective in improving the precision of discussion search. Our expert search consists of four features. Firstly, we make two kinds of data clean - webpage clean and candidate clean to adopt a profile-based document search. Core information is extracted from the W3C corpus such as the title, bolds, abstract, etc. Candidates are then matched with each web page and a profile is created for each candidate. Secondly, we use two variation weighting models, variation BM25 weighting model and DFR_BM25 weighting model. Query-based document length, not profile length, is used as document length in these weighting models to eliminate multiple topic noise. Query-based document length is the summation of lengths of extracted web pages that are relevant to the query. Thirdly, we use variation phrase weighting model to decease semantic confusion. Fourthly, field based two stage search method is used to make refined search. We demonstrate, on the basis of experiments, how these four approaches can effectively improve expert search.

5.12 Shanghai Jiao Tong University

Expert: SJTU1 – 4

In this research, we propose a new evidence-oriented framework to expert search. Here, the evidence is defined as a quadruple like (Query, Expert, Relation, Document). Each quadruple denotes that a "Query" and an "Expert", with a certain "Relation" between them, are found in a specific "Document". Within this framework, the task of Expert Search can be accomplished in three steps, namely, 1) evidence extraction: various kinds of co-occurrences between the expert and the query are extracted; 2) evidence quality evaluation: many novel factors like matching quality and context quality, are proposed as evidence quality evaluation; and 3) evidence merging: we proposed and compared two novel methods for evidence merging. The experimental results show that the new exploited evidences are quite useful and the evaluation of evidence quality improves the expert search significantly.The results also show that with cluster based merging, the result becomes even better.

5.13 Tsinghua University

Discussion: THUDSSUBPFSM, THUDSSUBPFSS, THUDSTHDM, THUDSTHDPFSM, THUDSTHDPFSS

Expert: THUPDDEML, THUPDDFBS, THUPDDL, THUPDDS, THUPDDSNEMS

Our expert finding system derives from that of last year, which first reorganize original documents to form PDDs, and then search and rank experts from these PDDs by employing retrieval model based BM2500 algorithm and bi-gram weighting. Our work this year focuses mainly on refinement of PDD documents and result reranking. We take advantage of email documents by producing Email-PDDs, appending Email subjects to original PDDs to form new PDDs, and combining search results of new PDDs and Email-PDDs. Regarding the result reranking stage, we have examined whether certain query-independent features – such as person activity and expert degree – help to find experts more accurately. Another new reranking approach we probed is to make use of social network, which is synthesized based on co-occurrences in web pages or email communications.

In Discussion Search task, several approaches have been probed. First, we discard useless and meaningless information in the email corpora to diminish the noise that affects the retrieval results. Then we examine the effectiveness of different field features in email such as quoted text and subjects of the email, some field features are emphasized by enforced as PFS (Primary Field Space) in our retrieval model. Finally we combined the adjacent serial emails to email threads and calculate the similarities of the single email and its threads respectively then integrate them together. Queries were constructed from the "query" field and "description" field. And all the experiments are base on our search engine TMiner.

5.14 University of Amsterdam

Discussion: UAmsBase, UAmsPOSBase, UAmsPOStQE, UAmsThreadQE

Expert: UvAbase, UvAPOS, UvAprofiling, UvAprofPOS

Following upon our last year's TREC Enterprise participation, we employ a standard language modeling setting for both tasks. Our aim for the discussion search task was to experiment with various query expansion techniques. Our first method employs blind relevance feedback, but instead of using the top ranked documents, we also include the contents of the accompanying threads. Our second method enriches the query by adding noun phrases from the description and narrative fields. We also experimented with combining the outcomes of the different approaches. Results indicate that adding terms from the description and narrative fields helps in most cases but not all. Thread-based query expansion did not deliver the desired results, due to topic drift. As to the expert search task, our baseline method calculates the probability of a candidate being an expert given the query topic. This probability is estimated by iterating over all documents that are associated with the given person. Moreover, we introduce the topical profile of an individual, which reflects the person's competency over a set of knowledge areas. The expert search topics were used as knowledge areas, and the topical profile of each W3C candidate was calculated. A rank-based combination of expert finding and profiling methods resulted in remarkable improvements over the baseline.

5.15 University of Glasgow

Expert: uogX06csnP, uogX06csnQE, uogX06csnQEF, uogX06ecm

In our participation in the Enterprise Track, we aim to develop our novel voting model for expert search. Our newly-proposed approach models expert search as a voting process. In our model, a candidate's expertise is represented by a profile, which is a set of documents associated to the candidate. Then, using the ranked list of retrieved documents for the expert search query,

we propose that the ranking of candidates can be modeled as a voting process, from the retrieved documents to the profiles of candidates. The votes for each candidate are then appropriately aggregated to form a ranking of candidates, taking into account the number of voting documents for that candidate, and the topicality of the voting documents. Our voting model is extensible and general, and is not collection or topics dependent.

This year in TREC, we test two new approaches for appropriately aggregating the votes for candidates. Moreover, we integrate a new component into the model that takes into account the candidate's profile length. Finally, we test a selection of approaches to increase the accuracy of the voting documents.

5.16 University of Illinois at Urbana-Champaign

Expert: UIUCe1, UIUCe2, UIUCeFB1, UIUCeFB2

We submitted four automatic runs, all using the title field of a topic and the whole corpus. Our goal is test the effectiveness of a new language model for expert retrieval. The new language model is based on the model 2 proposed in (Balog et al., 2006) with the following three extensions: (1) We model the document-candidate association using a mixture model that allows for putting different weights on matching the email and matching the name of a candidate. Thus we have a complete unigram language model for this task. (2) We use the count of email matches in the supporting documents for a candidate to define a prior on candidates such that a candidate whose supporting documents have many email matches would be favored. (3) We perform topic expansion and generalize the language model from computing the likelihood to computing the KL-divergence.

5.17 University of Maryland

Expert: UMDemailTLNR, UMDemailTTL, UMDthrdTTLDS, UMDthrdTTL, UMDthrdTTLNR

We have adopted a simple unsupervised approach that focuses only on mailing lists as the source of evidence of candidate expertise. The system first retrieves a set of emails or threads that are relevant to the topic and scores the candidates based on references in the headers and mentions in the text to their names and email addresses in the retrieved set. The credit given by each reference or mention is weighted according to (1) the retrieval similarity (to the topic) score of the email where the reference appears, and (2) in which field (headers, new text, quoted, etc.) in that email it appears.

5.18 University of Massachusetts

Discussion: UMaTDMixThr, UMaTiMixHdr, UMaTiMixThr, UMaTiSmoThr

Expert: UMaTDFb, UMaTiDm, UMaTNDm, UMaTNFb

This year the University of Massachusetts took part in both tasks of the Enterprise track. For the DS task we compare two methods for incorporating thread evidence into the language models of email messages. To group emails by thread we used the *all-in-reply-to* list provided by William Webber, concatenating the text of related messages.

One approach for incorporating thread context is to estimate a language model of the thread and interpolate it with the smoothed language models of other email components (header and mainbody). We use Dirichlet smoothing and automatically set the α parameter equal to the average component length. An alternative way to take advantage of thread information is to use it as a background model for smoothing email components. The idea is that threads would provide a more reasonable fallback distribution than a word distribution for general English. Our experimental results show that smoothing with a thread-based fallback model is more effective

than smoothing with a general collection model. However, constructing a mixture of language models from header, main body and thread text is more effective.

Our approach to the ES task represents candidate experts as mixtures of language models from associated documents and then ranks candidates according to query likelihood. Since the candidate representations are probability distributions over terms, we can build richer models by interpolating models estimated from different subcollections or different types of documents, or different entity definitions; in short, retrieval settings representing different descriptions (aspects) of a person entity. For example, we use two subcollections (www and lists), and two definitions (full name and last name). This model also preserves the information inherent in individual documents, such as structure and term proximity. Therefore we can use document retrieval techniques to capture higher-level language features. We use pseudo-relevance feedback and phrase expansion.

5.19 University of Ulster and St. Petersburg State University

Expert: sophiarun1 – 3

The SOPHIA group used the Contextual Document Clustering algorithm to cluster the W3C document corpus (documents from www and lists catalogs) into hundreds of thematically homogeneous clusters. Given a topic, the most relevant clusters were used to select experts for that topic. The expert relevancy score was calculated based on the number of mails sent by the expert from within the relevant clusters and similarities between these mails and the topic.

5.20 University of Waterloo

Discussion: uwTbaseline, uwTDbaseline, uwTDsubj, uwTsubj

Expert: uwXSHUBS, uwXSOUT, uwXSPMI

For the discussion search task, we hypothesized that the author's of an email tend to give their subjective opinion about the topic in discussion. In this year's discussion search track, we tested this hypothesis by re-ranking the email lists based on the presence of certain subjective adjectives in the proximity of the query words.

Experts, people who are knowledgeable about a given topic, tend to associate themselves with the topic over certain period. For expert search, in one approach, we estimated the association with the topic by studying the patterns in the mailing lists. We used graph-based ranking algorithm like HITS algorithm and PageRank to rank the candidates. In other approach, we estimated the expertise using statistical measures like mutual information etc, b/n the candidate and the topic.

References

Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–50, Seattle, WA, August 2006.

Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC 2005 enterprise track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005. URL `http://trec.nist.gov/pubs/trec14/t14_proceedings.html`.

Yejun Wu, Douglas Oard, and Ian Soboroff. An exploratory study of the w3c mailing list test collection for retrieval of emails with pro/con arguments. In *Proceedings of the Third*

Conference on Email and Anti-Spam, Mountain View, CA, July 2006. URL `http://www.ceas.cc/2006/26.pdf`.

Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, November 2006.