L3S Research Center at TREC 2006 Enterprise Track

Sergey Chernov, Gianluca Demartini, and Julien Gaugaz *

L3S Research Center
University of Hanover, Hanover, Germany
{chernov, demartini, gaugaz}@l3s.de

Abstract. The L3S Research Center submitted four runs at Enterprise Track for the first time in 2006, all of them are based solely on the W3C mailing lists. The first run serves as a fully automatically produced baseline. The second run uses a threshold on the document scores to limit the number of documents used for expert ranking. The third uses in addition a threshold on the experts scores in order to decide how many experts to retrieve. Our last run exploits the manually assigned topic specificity values, which predicts a number of relevant expert for each query. The results show that the simple threshold techniques outperform the baseline, while the current definition of query specificity does not improve the result quality.

1 Introduction

We performed experiments within an Expert Search task in the scope of the Enterprise Track 2006. We based our four techniques solely on the 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 different thresholds on the document score as well as the expert score. Using a 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 the specific information needs do not assume fewer relevant experts. It was an unexpected result, since normally the more specific your question, the less experts you expect to find. 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 topic specificity definition in the expert search community. 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 the W3C collections is that it could be enhanced with a specific expert search interface and can allow tracking user logs while searching experts with it.

* Authors are listed in alphabetical order

2 Document Collection Management

We based our techniques solely on the W3C mailing lists. The main assumption was that the author of an email is an expert on the subject addressed by the email she wrote. To manage the W3C mailing lists first we created a XML valid file[1] containing the structure of the mailing list collection.

After this step we parsed the file with an XML parser and created a Lucene (an open source information retrieval library[2]) Index with the fields described in the Table 1, in order to retrieve relevant emails.

Email Field	**Description**
Body	Full text of the email
CC	Email addresses in CC
DocNo	ID of the supporting document for TREC
From	Email address of the sender
ID	Unique ID of the email
InReplyTo	ID of the email to which this one is an answer
Name	Name of the sender
Received	Date of mail receiving
Sent	Date of mail sending
Subject	Subject of the email
To	Email address of the receiver

Table 1: The Email Fields Indexed with Lucene.

The mapping between the candidates and the email authors considered only exact match of the email address.

3 Runs Description

After building the inverted index of the mailing list, it is possible to retrieve emails relevant to the topics proposed in Enterprise Track 2006 for the Expert Search Task. We used an internal Lucene TFxIDF ranking function, to get the retrieval status values (RSV) for each email. The produced scoring was used to estimate the expert scores. The authors of the majority of the relevant emails (with at least one query term) were considered experts on the topic. The model was tuned on the results from Expert Search task 2005. In the following sections we describe each run in details.

3.1 Baseline (l3s1)

The topic structure in Enterprise Track 2006 is different from the topics of 2005. In year 2006 for each topic a Title, a Description, and a Narrative is available in

[1] available at http://www.l3s.de/∼demartini/w3c/w3c-lists-supercleaned.html
[2] http://lucene.apache.org/

contrast to Title part only of topics of 2005. To have at least one fully automatic and data-independent method we decided to use only the Title part of the query in our first run. It makes the run perfectly comparable with the runs from other participants and from the past year.

We first retrieved all the emails relevant to the query (composed by the keywords in the Title of the topic) and then we ranked the authors according to the number of relevant emails they have written. In this simple scenario the *expert score* (ES) is given by the number of emails they have written on the topic. After ranking the authors this first algorithm retrieves the top 5 experts. We assume that in real-world task it is unrealistic to browse as many experts as we normally browse documents. While it is common fact that majority of users do not look after the first 10–30 results, we expect that number of top-k expert should lie in the interval of 5–10 experts. This run is considered to be the baseline.

3.2 Using Document Score Threshold (l3s2)

The second algorithm uses the whole topic description provided by TREC and builds a weighted "OR" query of all the parts (Title, Description, and Narrative) of the original topic. The final query is obtained with an additional boosting of term weights as follows:

$$Query = \quad 3.0 * Title \quad \text{OR} \quad 2.0 * Description \quad \text{OR} \quad 1.0 * Narrative \qquad (1)$$

The problem is that for such long queries we have in average the 80% of the collection documents relevant to each query. In this case a good ranking of the retrieved documents should put on the top of the list the most relevant ones. To resolve the "too-many-results" problem we decided to fix a *document threshold* to limit the number of document considered to assess the expertise of the candidates. We assume that with low RSV we need more documents to decide about the expertise. We fixed a value of RSV to be filled by the top-k retrieved documents.

We learned the parameters from the topics of 2005 test collection (which uses the same document collection as of 2006) where the relevance judgments were available. We compared different possible thresholds to see which one performed better in terms of a Mean Average Precision (see Figure 1). The best results were achieved when we consider the top 240 documents on average. Instead of using the fixed number of documents, we calculated that the total sum of RSV for the first 240 documents is equal to 76.5 (on average). The performance of other thresholds is shown on the Figure 1. To smooth differences between popular and rare queries, we used this RSV threshold rather than fixed number of documents. So for every query we took into account only the top documents until their sum of RSV reaches the value of 76.5.

After ranking the documents and limiting the set of relevant documents, the expert search were computed as the sum of RSVs of their emails and, as in the run l3s1, only the top 5 experts were retrieved.

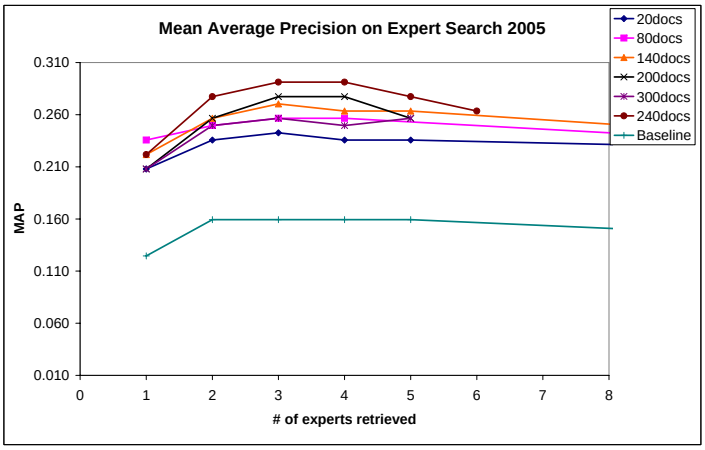

Fig. 1: Performance of different document thresholds on Enterprise Track 2005

3.3 Using Expert Score Threshold (l3s3)

The third algorithm we proposed is based on the second run with an additional improvement. The process of selecting the documents to decide about the expertise of the candidates is the same as in the run l3s2. The enhancement is done using an *Expert Score Threshold* (EST) to avoid the retrieving of a fixed number of expert for each query. Our assumption is that there are different types of topics and for some of them there will be more experts and of some others less experts depending on the different characteristics of the topic itself. The relationship between the type of topic and the number of experts on it will be considered more in details for the run l3s4.

In the run l3s3, instead of retrieving the top N (with N fixed) candidates after ranking them, we decided to compute the ES as the RSV sum over all emails in the relevant set written by the expert and to put a threshold in order to retrieve only the experts with scores above a fixed threshold.

We decided to retrieve 5 experts on average, but for some topics we retrieved less than 5 and for some other topics we retrieved more than 5. Using the 2005 test collection we found that the average expert score at rank 5 is 1.2, so we fixed the EST at this value, and all experts with the ES above 1.2 were retrieved, but minimum one expert.

3.4 Using the Topic Specificity (l3s4)

In the last run we assume that the number of available experts depends on the *specificity* of the topic. For example, there could be a lot of experts for the topic "Web Service Architecture", while only a few for "DOM traversal and range". We manually judged the specificity of each topic as 1 (very broad), 2 (usual topic) to 3 (very narrow). The topic specificity was defined as "If you input the

query into a search system, do you expect to find big, moderate or small number of experts?". This definition is both collection and user dependent. On the one hand, only a part of all the possible expertises is present in W3C collection, on the other hand, your expectation about the topic specificity is influenced by your background, for example, some people see a "DOM traversal and range" problem as a very broad area.

In our experiments we considered three assessors, their pairwise disagreement lies in the interval [0.26;0.38] which is very substantial. It indicates that the notion of topic specificity should be defined in a more consistent manner, which does not allow such a vague interpretation. The average real numbers among three available judgments were used. The correlation between user-defined topic specificity and number of relevant experts in the collection for topics from years 2005 and 2006 is shown on the Fig. 2 and Fig. 3.

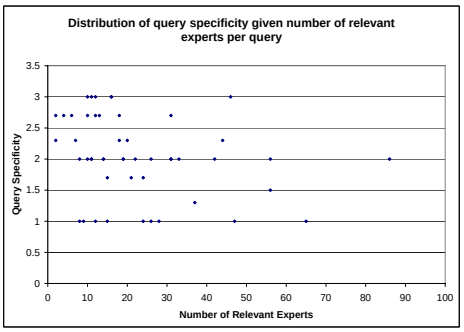

Fig. 2: Distribution of the Topic Specificity in Enterprise Track 2005 Queries.

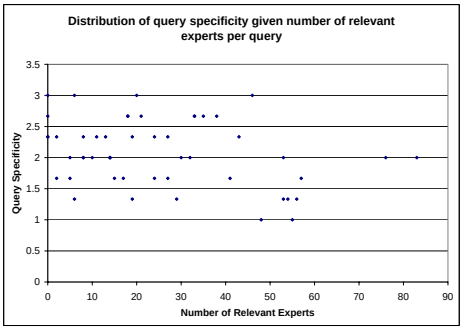

Fig. 3: Distribution of the Topic Specificity in Enterprise Track 2006 Queries.

The actual assignment of the specificity values is presented in the Table 3 (see at the end of the report). From the plots we do not observe any correlation between specificity values and number of experts. This can be either the real situation or just a side effect of our not perfect definition of topic specificity. Currently, it depends heavily on user expectations, while we would like to establish more robust measure for prediction of the expected number of experts. In our run l3s4 we multiplied EST with the coefficients 0.5, 1.0, and 1.5 (for specificity 3, 2, and 1 accordingly) to modify the number of the experts retrieved.

4 Discussion of the results

In this section we compare the results obtained with our four runs showing which method performs better than the other.

run	MAP	bpref	P5
l3s1	0.029	0.042	0.167
l3s2	**0.131**	**0.140**	**0.571**
l3s3	0.106	0.112	0.416
l3s4	0.115	0.122	0.444

Table 2: Average of some measures across the 2006 topics

One possible way to compare our 4 runs is using the 2006 results only. We can see which run performed better comparing the average value of the evaluation measures across the 2006 topics. The results are presented in Table 2 where it is possible to see that the run l3s2 had the best performance and that the run l3s1 (the baseline) was the worst one. The results indicate two main facts. First, the variable number of experts was not helpful in the run l3s3. Second, our definition of topic specificity does not work well for the Expert Search task. While it is still interesting to explore broad range of specific and broad information needs, better predictor is needed for the number of expert to return.

5 Conclusions and Future Work

We consider the Expert Search task as one of the most important directions of a future information retrieval research. In this report we described the L3S Research Center runs at the Enterprise Track 2006. We conducted experiments on the W3C mailing lists and tested several thresholds on the document and on the expert scores. While the thresholds on the document scores proved to be helpful, the threshold on expert scores did not improve the retrieval performance. The manually assigned topic specificity values, for prediction of the number of relevant experts for each topic, did not work for the current setup. We believe, that it can be an effect of a vague definition of the topic specificity and leave the development of better notion of specificity as a future work.

TN	Title	NRE	TS	TN	Title	NRE	TS
1	Semantic Web Coordination	11	2.0	52	ontology engineering	168	2.0
2	Research and Development Interest	8	1.0	53	W3C translation policy	180	1.7
3	Cascading Style Sheets (CSS)	26	1.0	54	xml digital signature	163	1.7
4	Web Services Addressing	44	2.3	55	Semantic Web Rule Language	234	1.3
5	Hypertext Coordination	20	2.3	56	Rich Web Client	88	1.7
6	Mobile Web Initiative	16	3.0	57	OWL Lite Specification	178	2.0
	Workshop Program Committee			58	text XML query language	172	2.3
7	WCAG reviewers	2	2.3	60	SOAP security considerations	165	2.0
8	P3P Specification	24	1.7	61	VoiceXML Browser Implementation	148	2.0
9	XML Query	47	1.0	62	Mereology	177	2.0
10	XML Schema	28	1.0	63		172	2.3
11	Voice Browser	86	2.0	64	MathML specification	154	2.3
12	Web Services Description	33	2.0	65	RSS	163	1.7
13	Web Content Accessibility Guidelines	42	2.0	66	parsing MathML	167	2.7
14	Rules Workshop program committee	18	2.3	67	Privacy on the Web	201	1.3
15	XSL/FO Task Force	12	3.0	68	semantic search	160	1.3
16	Semantic Web Best	56	2.0	69	CSS3	187	2.7
	Practices and Deployment			70	Evaluation and Report Language	188	2.0
17	Education and Outreach	26	2.0	71	css floating elements	226	2.3
18	Compound Document Formats	31	2.7	72	PNG specification	183	1.7
19	Device Independence	19	2.0	73	DOM traversal and range	184	3.0
20	Math Interest	15	1.7	74		144	2.3
21	Internationalization Tag Set (ITS)	13	2.7	75	W3Photo	152	2.7
22	W3C's Tenth Anniversary	11	3.0	76	URI Fragment identifier	0	2.3
	Birthday Celebration attendees			77	XML schema test collection	115	2.0
23	ERCIM employees	8	2.0	78	rdf ontology	177	1.0
24	Protocols & Format	12	1.0	79	Semantic Web	135	1.0
25	Patent and Standards Interest	22	2.0	80	User Agent Accessibility Guidelines	148	2.0
26	Chairs	65	1.0	81	Description logics in the Semantic Web	163	1.7
27	Authoring Tool Guidelines	7	2.3	82	APPEL A P3P Preference	0	2.7
28	Multimodal	56	1.5		Exchange Language		
29	Internationalization Core	11	2.0	83	Semantic interpretation for	234	2.3
30	HTML	15	1.0		speech recognition		
31	XSL	24	1.0	84	W3C validation services	233	1.3
32	RDF Data Access	31	2.0	85	Timed text specifications	133	2.7
33	Advisory Committee	391	2.0	86	RDF graph serialization	211	2.7
34	SVG	37	1.3	87	XML Processing Model	202	2.0
35	Social Meaning of RDF and	16	3.0	88	patent policy	213	1.7
	URIs Task Force			89	XML interchange	143	2.3
36	Technical Plenary Attendees	4	2.7	90	CSS test suite	152	2.7
37	XForms	21	1.7	91	SVG Accessibility	216	3.0
38	Mobile Web Best Practice	31	2.0	92	Notation 3	202	2.0
39	Technical Architecture	9	1.0	93	machine translation	160	1.3
40	XML Coordination	14	2.0	94	Voice Browser	178	2.0
41	Tech Plenary Program Committee	10	3.0	95	XSL Transformations	162	1.3
42	SYMM	18	2.7	96	orphaned annotations	175	3.0
43	URI Coordination	6	2.7	97	Device Independence Principles	111	2.3
44	Advisory Board	10	2.0	98	RDF Semantics Datatype Interpretations	160	2.0
45	Evaluation & Repair Tools	19	2.0	99	P3P for my website	0	3.0
46	XML Binary Characterization	46	3.0	100	XML Encryption standard	0	2.3
47	XML Core	14	2.0	101	Implementation of EPAL	182	2.7
48	Internationalization Guidelines,	12	2.7	102	User Agents testing	167	2.3
	Education & Outreach (GEO)			103	Annotea server protocol	153	2.3
49	AC Meeting attendees	2	2.7	104	Web Service Architecture	0	1.0
50	MWI Device Description	10	2.7	105	Authoring tool web accessibility		1.7

Table 3: Topic Specificity Values (TN — Topic Number, TS — Topic Specificity, NRE — Number of Relevant Experts).