Examining Assessor Attributes at HARD 2005

Mark Baillie
CIS Department
University of Strathclyde
Glasgow, UK
mb@cis.strath.ac.uk

Ian Ruthven
CIS Department
University of Strathclyde
Glasgow, UK
ir@cis.strath.ac.uk

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Relevance feedback.

General Terms: Performance, Experimentation

Keywords: Query expansion, clarification forms

1. THE HARD TRACK

The TREC HARD (High accuracy Retrieval from Documents) track was motivated to investigate techniques for personalised retrieval of documents. Through the use of a limited dialogue with the TREC assessors, the track facilitated the gathering and exploitation of information about the assessors' personal search context (e.g. knowledge of search topic) which could be used to improve document retrieval. In this paper we describe experiments, run within the context of the 2005 HARD track, which indicate that assessor attributes such as familiarity, interest and confidence when searching a topic can help determine when the utilisation of automatic query expansion improves retrieval over the original document ranking.

2. COMPARING STRATEGIES

The HARD track protocol is as follows. First a set of 50 topics were distributed to the participating groups who each perform a baseline retrieval run on the AQUAINT document corpus. Each group is then allowed to ask the topic assessor to complete a clarification form for each topic they will assess. Each form was designed to gather personal, contextual information about the assessor's relationship to the topic which can be used for personalising the search. Contextual knowledge could consist of information on the assessor's knowledge of the topic or their confidence in judging retrieved documents. Once this data had been gathered it was used to re-rank the baseline run, typically by performing some form of query modification. The new retrieval runs are returned to TREC, with the results from by both the baseline and modified retrievals evaluated by the assessor who completed the clarification form.

Our interest was to compare techniques that might perform well under different searcher contexts. Specifically we looked at assessor knowledge of the search topic, assessor interest in the topic and assessor confidence in judging documents about the topic. We asked the assessor about these aspects in the clarification form using a 3-point scale (high/average/low for each attribute) [1, 2]. Our aim was to analyse the relative effectiveness of retrieval strategies across these attributes, with assessors separated by their responses, against a baseline run; Okapi BM25 using the short title of each TREC topic. The strategies we investigated were:

Query expansion using representative terms: Terms from the top N ranked documents (topic language model) were scored by how representative they were to a topic using the Kullback-Leibler distance between the topic and collection language model [3]. Representative terms are those that are very general or common to a topic. We then selected the top Q representative terms to expand the original query, performing a new retrieval. We hypothesised that assessors with low topical knowledge, for example, would benefit from this technique as it would lead to the retrieval of more general, introductory documents about a topic.

Query expansion using discriminatory terms: Discriminatory terms are those that are specialised or infrequent with respect to a topic. We selected the top Q discriminative terms for a topic and added these to the query to perform a new retrieval. We hypothesised that assessors with high knowledge of a topic would benefit from this technique as it would lead to the retrieval of less general but more detailed documents on a topic.

Query expansion using emotive terms: Documents that are more interesting to read may be ones that could lead the assessor to read in more detail (and hence find relevant information) or are more suitable for assessors who are less interested in the topic being search. To test this assumption, we carried out a query expansion using emotive terms. Emotive terms are those that carry an emotional impact, e.g. dramatic, significant, amazing. We extracted 280 of these from a thesaurus and, for each topic, selected those terms that contributed most to the topic language model, expanding the original query to provide a new retrieval [1]. The new retrieval prioritises those documents that contain emotive text.

Retrieval by readability: Assessors with low interest, knowledge, or confidence in assessment may find it easier to assess documents that are easier to read. To investigate this we used a combined readability score measure which combines the document Retrieval Status Value with the Flesch readability score, which assigns high values to documents that are more readable [4].

Pseudo-relevance feedback: Our final retrieval strategy was a query modification strategy based on Pseudo-relevance feedback. The original query was then expanded

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

Group		# group	BM25	Disc.		Rep.		Read.		Motiv.		Pseudo.	
Famil.	Low	11	0.220	3.9%	6/11	0.5%	4/11	-0.5%	5/11	-0.7%	5/11	3.0%	9/11
	Ave.	34	0.249	4.0%	23/34	1.6%	15/34	-0.2%	17/34	-0.4%	13/34	3.8%	21/34
	High	4	0.214	10.7%	4/4	4.4%	4/4	0.4%	3/4	0.4%	2/4	12.6%	3/4
Interest	Low	7	0.180	8.0%	5/7	1.2%	3/7	-0.3%	3/7	0.0%	2/7	4.6%	6/7
	Ave.	12	0.310	7.6%	9/12	2.5%	5/12	-1.1%	4/12	-1.4%	5/12	8.0%	9/12
	High	30	0.230	2.1%	19/30	1.5%	16/30	0.3%	19/30	-0.2%	13/30	3.1%	19/30
Conf.	Low	3	0.381	14.2%	3/3	7.8%	2/3	-1.4%	1/3	-2.8%	0/3	5.9%	3/3
	Ave.	24	0.178	0.3%	13/24	0.1%	8/24	-0.1%	13/24	0.0%	12/24	1.6%	15/24
	High	22	0.290	6.8%	17/22	1.7%	13/22	-1.1%	10/22	-1.1%	10/22	6.6%	15/22
Assess.	a	8	0.266	-0.1%	5/8	3.2%	5/8	-0.3%	5/8	-0.7%	5/8	3.8%	5/8
	b	8	0.216	2.8%	5/8	-0.6%	3/8	0.4%	5/8	-0.4%	3/8	2.7%	5/8
	c	8	0.255	8.2%	6/8	3.0%	3/8	-0.2%	3/8	0.0%	4/8	10.0%	8/8
	d	8	0.185	-0.2%	4/8	1.4%	2/8	-0.4%	3/8	-0.2%	2/8	0.3%	3/8
	e	8	0.239	4.9%	5/8	-1.1%	4/8	-0.3%	4/8	-0.5%	3/8	2.1%	6/8
	f	10	0.260	9.6%	9/10	3.7%	6/10	0.0%	6/10	-0.8%	3/10	7.0%	7/10

Table 1: The % improvement in R-precision over the baseline for each technique across all groups, across the different assessor groups and assessors. We also provide the number of times a technique was successful on the right-hand side of the column. i.e. improvement over the baseline BM25.

by the top Q terms that contribute most to the top N documents. This was a benchmark strategy to compare against our novel approaches.

Our *hypotheses* - compressed into a general hypotheses here for space reasons - was that individual techniques would work better for different assessor attributes (knowledge, interest, confidence). We also investigate the relative effect of assessor attribute and topic to retrieval success. Our investigation examines which strategies performed best for each assessor group e.g. assessors with low topical knowledge.

2.1 Results

We examined various combinations of N and Q ranging from $N = 1, \ldots, 25$ and $Q = 1, \ldots, 40$. We present the percentage improvement (or deterioration) in R-precision over the OKAPI BM25 baseline for the *optimal* parameter setting only for each technique (Table 1). We also present the number of topics where there was an improvement over the baseline across the various aspects of assessor attributes. Overall, expanding the query with discriminative terms (*Disc.*) and also Pseudo-relevance feedback (*Pseudo.*) performed best for all 50 topics, while utilising document readability and query expansion using emotive terms were worst.

When examining the performance of each technique across the different levels of assessor familiarity (*Famil.*) we found that expanding the query using *Disc.* improved over the baseline by 10% for assessors with *high* topic familiarity. This was in agreement with our original assumption. Using pseudo-relevance feedback did provide better improvement, 12.6%, although for one topic there was no gain at all. However, expanding the query with representative terms (*Rep.*), did not benefit those assessors with *low* topic familiarity. Examining the performance of each technique across the different levels of assessor interest in a topic we found that using *Disc.* improved document ranking for those assessors with *average* to *low* topic interest. Also, when examining the performance of each technique across the different levels of assessor confidence in assessing a topic we again discovered that *Disc.* improved document ranking for those assessors with *low* confidence. Finally, when examining the performance of each technique across the various assessors separately, it was highlighted that assessors 'c' and 'f' were more receptive to *Disc.* than others such as 'a' and 'd'.

3. CONCLUSIONS

There was evidence to suggest that assessor attributes such as familiarity, interest and confidence could be utilised as part of a decision mechanism to determine when performing automatic query expansion for improving document ranking. Although sample sizes are relatively small, the results indicate that particular assessor attributes may be more conducive to successful query expansion, especially when using discriminative terms. Further investigation is required to confirm these assumptions. Of particular interest is to determine whether expanding the query with discriminative terms does indeed retrieve more suitable documents or if the improvement in retrieval is a consequence of the actual document assessment process. For example, assessors 'c' and 'f' assessed more documents relevant, on average, than the remaining assessors, which may be a contributing factor to the improved performance [2]. Future work will investigate these issues in greater detail.

4. REFERENCES

[1] M. Baillie, D. Elsweiler, E. Nicol, I. Ruthven, S. Sweeney, M. Yakici, F. Crestani and M. Landoni. University of Strathclyde at TREC HARD. *TREC-2005*, 2006.

[2] I. Ruthven, M. Baillie and D. Elsweiler. The relative effects of knowledge, interest and confidence in assessing relevance. *To Appear: Journal of Documentation*, Emerald.

[3] D. J. Harper, G. Muresan, B. Liu, I. Koychev, D. Wettschereck and N. Wiratunga, The Robert Gordon University's HARD Track Experiments at TREC 2004. *TREC-2004*, 2005.

[4] N.J. Belkin, I. Chaleva, M. Cole, Y.-L. Li, L. Liu, Y.-H. Liu, G. Muresan, C. L. Smith, Y. Sun, X.-J. Yuan and X.-M. Zhang, Rutgers' HARD Track Experiences at TREC 2004. *TREC-2004*, 2005.