Error Analysis of Difficult TREC Topics

Xiao Hu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
xiaohu@uiuc.edu

Sindhura Bandhakavi
Department of Computer Science
University of Illinois at Urbana-Champaign
bandhakavi@uiuc.edu

Chengxiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
czhai@cs.uiuc.edu

ABSTRACT
Given the experimental nature of information retrieval, progress critically depends on analyzing the errors made by existing retrieval approaches and understanding their limitations. Our research explores various hypothesized reasons for hard topics in TREC-8 ad hoc task, and shows that the bad performance is partially due to the existence of highly distracting sub-collections that can dominate the overall performance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Error analysis and performance evaluation

General Terms
Measurement, Performance

Keywords
Error analysis, hard topics, sub-collection, measures, dominance, distraction

1. PROBLEM DESCRIPTION
Experimental error analysis is very useful for understanding the limitations of existing IR approaches, and further, for putting forward improved methods or reasonable hypotheses. So far, a majority of retrieval evaluation has been done in the context of TREC [1]. However, evaluation is usually based on the average performance over multiple topics, rather than on the performance of an individual topic, and error analysis is rarely performed. Papers comparing the results of different TREC conferences [2] aimed at illustrating long-term TREC trends, rather than analyzing any particular experimental result or performance.

In this paper, we investigate the unexplored domain of retrieval error analysis. In particular, we focus on some difficult TREC topics on which most participating retrieval systems perform poorly. Analysis of such topics can be expected to help us understand the limitations of existing retrieval methods. We attempt to answer three research questions: (1) Are there topics hard for all the systems? (2) If a topic is hard for all the systems, is it hard for some document set while easy for others? (3) Are there any statistical characteristics of the document set or query that can explain the poor performance?

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58813-646-3/03/0007.

2. DATA SELECTION
The official experimental results and testing data of the ad-hoc task in TREC-8 [1] formed the starting point of our analysis task. The TREC-8 test collection is composed of four different document datasets (sub-collections), namely, Foreign Broadcast Information Service (FBIS), Financial Times 1991-1994 (FT), Federal Register 1994 (FR94), and Los Angeles Times (LA). We chose title only automatic runs for our purpose, because they represent the real world better than the long ones. We used the official TREC-8 ad hoc results of all participating groups, and eliminated non-optimal runs from the same group, which gave us 14 different runs. We measure the hardness of a topic by the *best* average precision that *any* system has achieved on the topic. From the 50 topics provided, we identify 10 hardest topics, on which no system has an average precision above 0.17. We also choose 5 easy topics all with an average precision above 0.9 for comparison.

3. SUB-COLLECTION ANALYSIS
Our first hypothesis is that a topic hard for the whole TREC-8 collection may turn out to be easy for some sub-collections. To test this hypothesis, we process raw result files and compute the average precision for each sub-collection for each topic. Furthermore, in order to quantify the influence of sub-collections on the overall performance, we propose several measures defined on the break-even cutoff of the whole set results. Breakeven cutoff is the point where the document rank equals the total number of relevant documents for the topic. Intuitively, it is the ideal number of documents that the whole set should retrieve. The three sub-collection measures we propose are defined as follows:

1. "Dominance" indicates how much more a sub-collection contributes to the result than it is ideally supposed to. If we regard the number of relevant documents in a sub-collection as its "quota", then "dominance" is defined as the number of documents retrieved up to the break-even point from that sub collection divided by the quota.

2. "Distraction" measures how much a sub-collection contributes towards non-relevant documents, and is given by the number of non-relevant retrieved documents from that sub-collection divided by the number of retrieved documents from that sub-collection, all computed up to the break-even point.

3. "Average Rank" of the documents retrieved at the break-even point indicates how the retrieved documents from each sub-collection rank in the result. This measure along with the former two measures gives a more complete picture of the influence of each sub-collection.

The values of these measures for seven representative topics are showed in Table 1. The topics selected had relatively more relevant documents than others in each sub-collection, thus promising more reliable results. While average precision per sub-collection may qualify its "easiness", high dominance and distraction values may serve to explain the poor overall performance in spite of the existence of sub-collections with good individual performances. Topic 439 is such an example with poor overall performance caused by the two distracting sub-collections, FBIS and LA, even though FR94 performs much better than the whole set. In summary, our study of the sub-collections shows that one of the causes of poor performance is the existence of one or more dominating and distracting sub-collections.

Table 1: Measures across Sub-collections

Topic	Measure	FBIS	FT	FR94	LA	Whole set
437	# Rel Docs	5	68	0	0	73
	Avg. Precision	0.062	0.005	NA	NA	0.049
	Dominance	0.2	1.059	0/0	0/0	
	Distraction	1	0.889	0/0	0/0	0.89
	Avg. Rank	71	36.5	0	0	
442	# Rel Docs	17	4	0	73	94
	Avg. Precision	0.063	0	NA	0.087	0.052
	Dominance	2.118	4	0/0	0.575	
	Distraction	0.944	1	0/0	0.857	0.915
	Avg. Rank	45	51.5	0	45.8	
439	# Rel Docs	48	73	89	11	221
	Avg. Precision	0.03	0.042	0.358	0.026	0.074
	Dominance	2.021	0.603	0.506	3.182	
	Distraction	0.938	0.841	0.444	0.943	0.819
	Avg. Rank	107.6	102.6	125.2	106.2	
421	# Rel Docs	32	31	10	10	83
	Avg. Precision	0.244	0.101	0.025	0.107	0.092
	Dominance	0.407	1.032	2.1	1.7	
	Distraction	0.538	0.875	1	0.941	0.867
	Avg. Rank	35.3	40.53	44.8	41.5	
449	# Rel Docs	10	44	5	8	67
	Avg. Precision	0.204	0.161	0.221	0.01	0.122
	Dominance	1.8	0.591	0.8	2.375	
	Distraction	0.778	0.731	0.75	1	0.821
	Avg. Rank	33.3	29.3	36.5	40.4	
413	# Rel Docs	24	43	1	1	69
	Avg. Precision	0.269	0.111	0	0	0.143
	Dominance	1.625	0.581	4	1	
	Distraction	0.744	0.76	1	1	0.768
	Avg. Rank	31.7	38.9	34	66	
410	# Rel Docs	19	42	0	4	65
	Avg. Precision	0.895	0.926	NA	1	0.912
	Dominance	0.789	1.095	0/0	1	
	Distraction	0.133	0.174	0/0	0.25	0.169
	Avg. Rank	31.3	32.3	0	31	

4. TF-IDF ANALYSIS

Since most retrieval methods rely on TF-IDF weighting, another hypothesis is that the difficult and easy topics differ in TF-IDF statistics. To test this hypothesis, we calculate the average IDF value of query terms as well as the "TF ratio", which is the ratio of average TF in relevant documents to that in the whole document collection. We expected to see higher average IDF values and TF ratios for the easy topics when compared to the difficult ones. However, results show that neither the IDF value nor the TF ratios have any significant difference in the cases of easy and hard topics. (The actual values are omitted due to space limitation.) This means that the difficulty of some topics is not necessarily due to the limitation of the TF-IDF style weighting, and it needs to be explained by further analysis.

5. CONCLUSIONS AND FUTURE WORK

Based on our observations we can conclude that:

- There exist hard topics that are hard for all the systems we studied.
- There exist hard topics that are hard across all the document sub-collections.
- There are some topics for which one document subset performs much better than others (e.g., 437, 439 and 421). So better retrieval performance at least partially relies on selection of the right document set, which is consistent with the findings in the work of distributed IR [3].
- From "Dominance" and "Avg. Rank", it can be seen that some sub-collections play a dominant role by contributing more retrieved documents than they should (with high "dominance") and by ranking them high (with low "Avg. Rank"). The difficulty of such topics is partially due to the dominance of the poorly performing sub-collections over the well performing ones.
- Different topics tend to have different distracting sub-collections (e.g., FT to 437; LA to 439, FR94 to 413), but the LA sub-collection appears to be distractive far more frequently than other sub-collections, as is evident from its values of "Distraction" and "Avg. Rank".
- Some sub-collections are "easier" (with higher average precision among the four sub collections) for one topic, but "distractive"(with high "distraction" value and low rank) for another (e.g., FR94 for 439 and 413). There is no sub-collection generally easy for all the topics.
- Comparing the TF ratio and average IDF between hard and easy topics, we have not found any obvious correlation between the hardness of the topics and the TF ratios or average IDF values.

As future work, we plan to examine more TREC data and the actual content of documents to further study what exactly makes a topic hard. Further analysis is also needed to understand why a particular sub-collection such as LA tends to be distractive, and such analysis can be expected to provide insight into how to deal with such a distracting collection in a retrieval model.

6. REFERENCES

[1] E. Voorhees and D. Harman, Overview of the Eighth Text REtrieval Conference (TREC-8), In Proceedings of TREC-8, pages 1-24.

[2] K. Spärck Jones, Summary Performance Comparisons TREC-2 Through TREC-8, In Proceedings of TREC-8, Pages B-1.

[3] A. L. Powell, J.C. French, J. Callan, M. Connell, and C.L. Viles, The impact of database selection on distributed searching, In Proceedings of SIGIR 2000, pages 232-239.