The TREC-8 Query Track

Chris Buckley and Janet Walz
Sabir Research, Inc
{chrisb,jwalz}@sabir.com

1 Introduction

The Query Track in TREC-8 is a bit different from all the other tracks. It is a co-operative effort among the participating groups to look at the issue of "query variability."

The evaluation averages presented in a typical system evaluation task, such as the TREC Ad-Hoc Task, conceal a tremendous variability of system performance across topics/queries. No system can possibly perform equally well on all topics: some information needs (expressed by topics) are harder than others. But what is quite surprising, especially to people just starting to look at IR, is the large variability in system performance across topics as compared to other systems. In a typical TREC task, no system is the best for all the topics in the task. It is extremely rare for any system to be above average for all the topics. Instead, the best system is normally above average for most of the topics, and best for maybe 5%-10% of the topics. It very often happens that quite below-average systems are also best for 5%-10% of the topics, but do poorly on the other topics. The Average Precision Histograms presented on the TREC evaluation result pages are an attempt to show what is happening at the individual topic level.

This large topic/query variability presents a great opportunity for improving system performance. If we can understand why some systems do well on some queries but poorly on others, then we can start introducing query dependent processing to improve results on those poor performance queries.

Unfortunately, we just don't have enough information from the results of a typical TREC task to really understand what is happening. The results on 50 to 150 queries are just not enough to draw any conclusions. The Query Track at TREC is an attempt to gather enough information from a large number of systems on a large number of queries to be able to start understanding query variability.

1.1 Query vs Topic

For the purposes of this track, a *topic* is considered an information need of a user. It includes a full statement of what information is wanted as well as information the user knows that pertains to the request. A *query* is what the user actually types to a retrieval system. It is much shorter than a topic, but is the only information from the user that the system has. Topic 51 (the first topic used in the Query Track) is given below. A query corresponding to Topic 51 might be something as simple as "Airbus subsidies".

TOPIC 51

<top>
<head> Tipster Topic Description
<num> Number: 051
<dom> Domain: International Economics
<title> Topic: Airbus Subsidies
<desc> Description:Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<smry> Summary:Document will discuss government assistance to Airbus Industrie, or mention atrade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<narr> Narrative:A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S.government, over federal subsidies to Airbus.
<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

<def> Definition(s): ...

1.2 Issues to Examine

There are a number of issues that we wish to examine in this and future Query Track experiments. They include

- Can we distinguish between easy and hard queries/topics?
  - Are queries hard or are topics hard?
  - Even if we can distinguish this from the results, can NLP analysis of a query distinguish this before-hand?
- What categories of queries can potentially yield performance differences?
- Where do query performance differences come from?
  - Examine system vs topic vs query.
- Can we easily create test collections with large numbers of queries with judgments?

If we can answer these questions, then we may make it possible to improve retrieval systems dramatically.

2 Query Track Test Collection Creation

The construction of the Query Track test collection consists of 2 sub-tasks. In the first sub-task, groups take each of topics 51-100 from TREC 1 and create one or more queries based on the topic. In the second sub-task, each group runs one or more versions of their system on all the queries from all the groups. The results are then evaluated and analysis can begin!

2.1 Query Creation Sub-Task

Groups create one or more versions of each of TREC topics 51-100 in categories

- Very short: 2-4 words based on the topic and possibly a few relevant documents from TREC disk 2.
- Sentence: 1-2 sentences using topic and relevant documents.
- Sentence-Feedback only: 1-2 sentences using only the relevant documents. The aim is to increase vocabulary variability.
- Weighted terms: lists of unstemmed terms with weights, possibly obtained through feedback on relevant documents from TREC disk 2.

The five participating groups produced 23 Query Sets. Each query set consisted of 50 queries corresponding to topics 51-100, for a total of 1150 queries. 15 Query Sets were produced by students and the rest by experts (retrieval system designers).

APL	INQ	Sab	Acs	Pir
Johns Hopkins	Umass	Sabir	Acsys	Queens
Expert	Students	Expert	Expert	Expert
2 weighted terms	5 short 5 sentence 5 feedback	3 short 1 feedback	1 short	1 short

Several versions of queries for topic 51 are given below. It was quite surprising how few duplicate queries there were, about 16%.

Sample of queries for Topic 51

•51 01 recent airbus issues
•51 02 Airbus subsidies dispute
•51 03 Airbus subsidy battle
•51 04 Airbus subsidies dispute
•51 05 U.S. Airbus subsidies
•51 06 What are the reactions of American companies to the trade dispute and how the dispute progresses?
•51 07 What are the issues being debated regarding complaints against Airbus Industrie?
•51 08 News related to the Airbus subsidy battle.
•51 09 U.S. and Europe dispute over Airbus subsidies
•51 10 Is European government risking trade conflicts over issue of Airbus subsidies?
•51 11 How is the Airbus business in the world ?
•51 12 why did the US put duties on airbus?

2.2 Retrieval Sub-Task

After the Query Sets were constructed, they were distributed to all the groups to run one or more retrieval runs on the TREC Disk 1 document collection (about 510,000 documents). The five groups performed 9 retrieval runs:

- APL : 1 run - words plus blind feedback
- INQ: 3 runs
  - only query terms
  - query terms plus structure
  - query terms plus structure plus blind feedback
- Sab: 3 runs
  - query terms plus adjacency phrases
  - query terms plus phrases plus 6 terms expansion from blind feedback
  - query terms plus phrases plus 27 terms expansion
- acs: 1 run - no expansion, base run
- pir: 1 run - blind feedback

The groups submitted the results (top 1000 documents retrieved for each query) to NIST for evaluation. There were a total of 203 runs; not all groups were able to run the 2 weighted term query sets. Thus the total was 9 runs * 21 NL queries plus 7 runs * 2 weighted terms queries.

The runs evaluated at NIST using trec_eval, concentrating on Mean Average Precision. The results of the initial evaluation were given to the five groups. This included

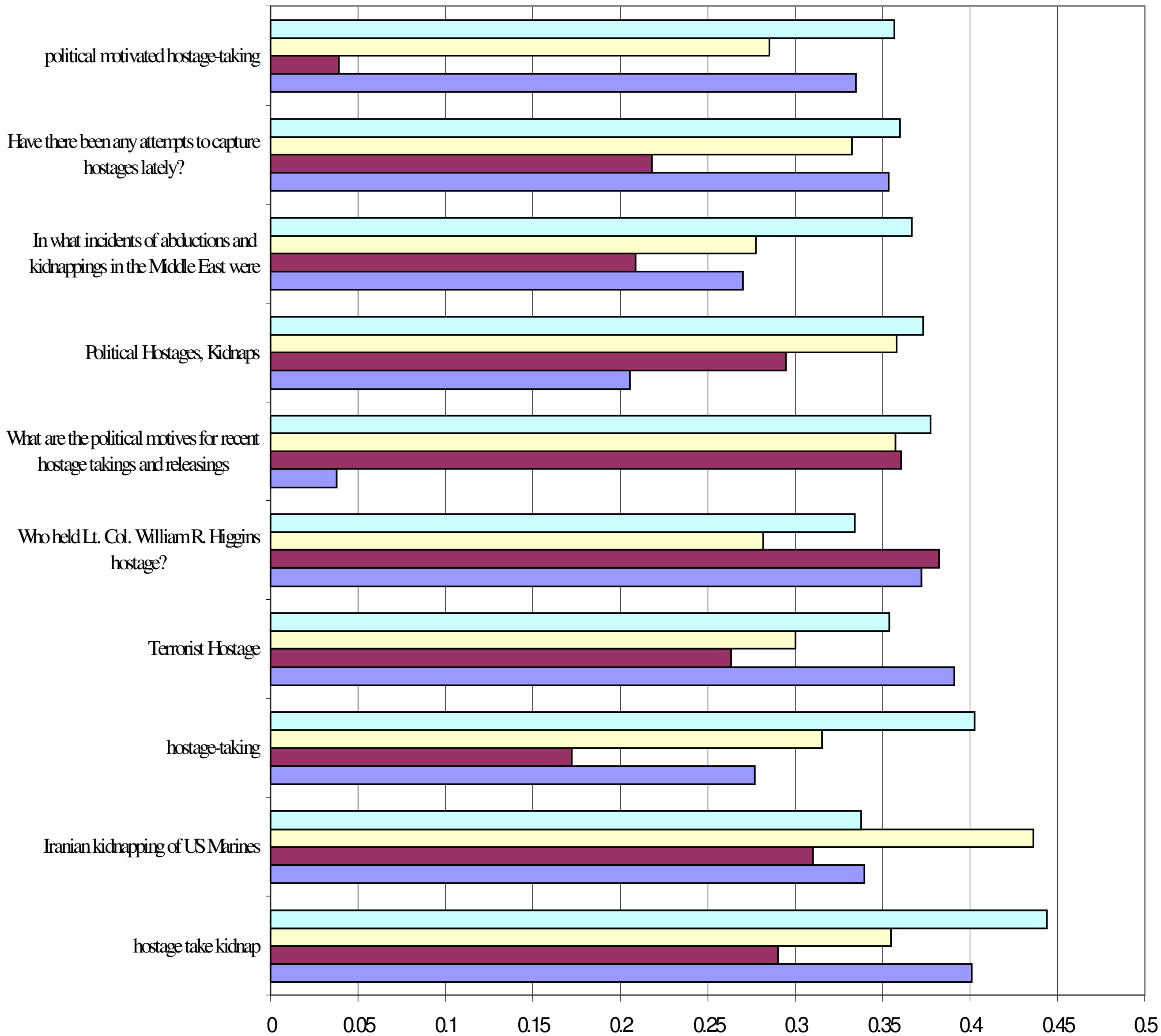
- Rankings of all documents (440 Mbytes in size)
- MAPs of all groups on all queries
- Various averages and standard deviations

3 Query Track Analysis

We present a very preliminary analysis of some aspects of the Query Track data. Other groups, notably the APL group of Johns Hopkins, have done more analysis. In addition, Walter Liggett of NIST has a paper in this proceedings.

3.1 Individual Query Analysis

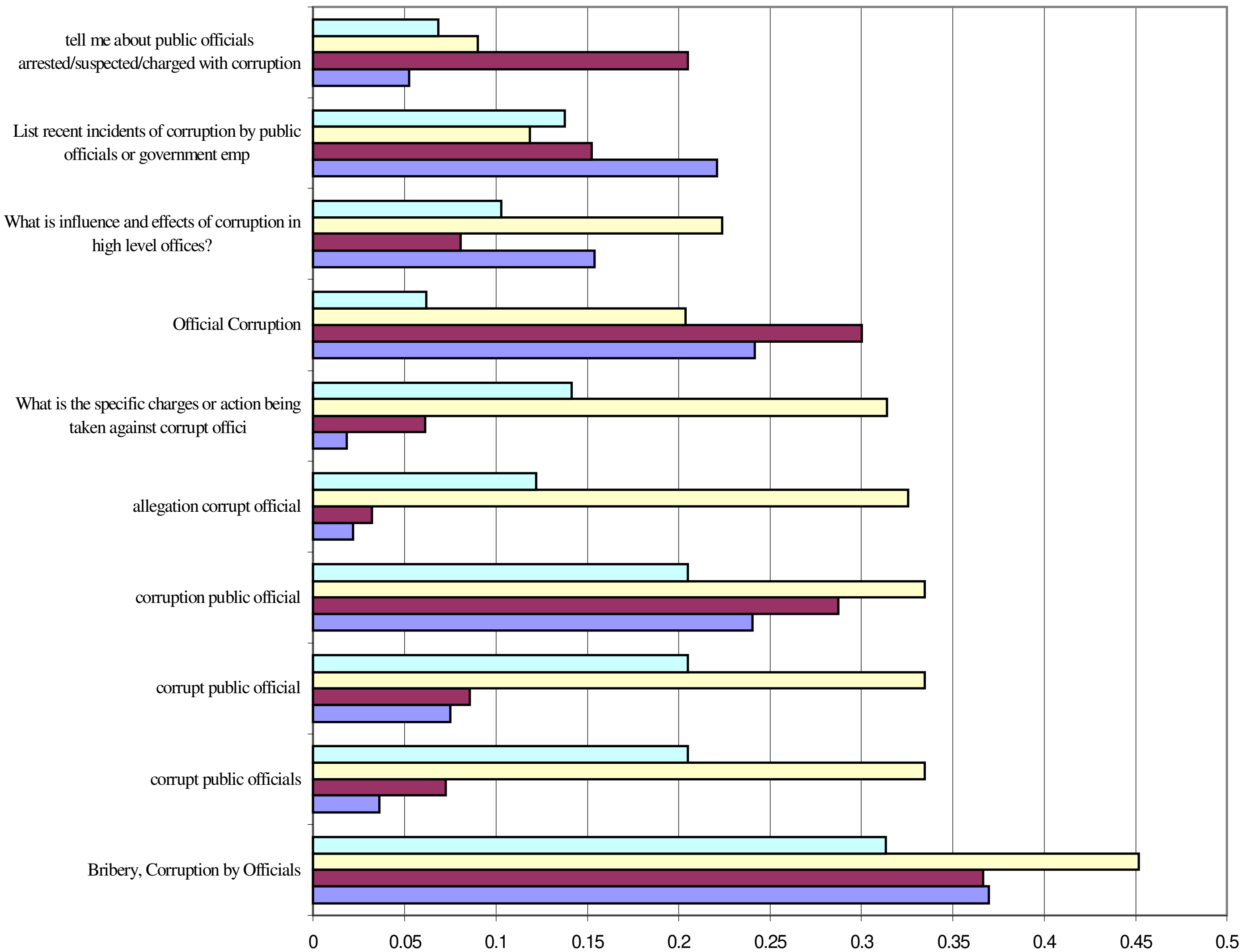
We look at the performance of 4 good runs on the top 10 queries per topic. The PIR, INQe, Sabe, and APL runs are the best runs of their respective groups, all using their own version of query expansion based on blind feedback. We want to examine how performance varies due to both system differences and query differences. Here, we look at how the 4 systems do on 4 topics, looking qualitatively at outliers, and doing an analysis of variants on each query.



ANOVA

rce of Varia	SS	df	MS	F	P-value	F crit
Rows	0.049356	9	0.005484	0.820275	0.602844	2.250133
Columns	0.07319	3	0.024397	3.649156	0.024951	2.960348
Error	0.180511	27	0.006686			
Total	0.303057	39				

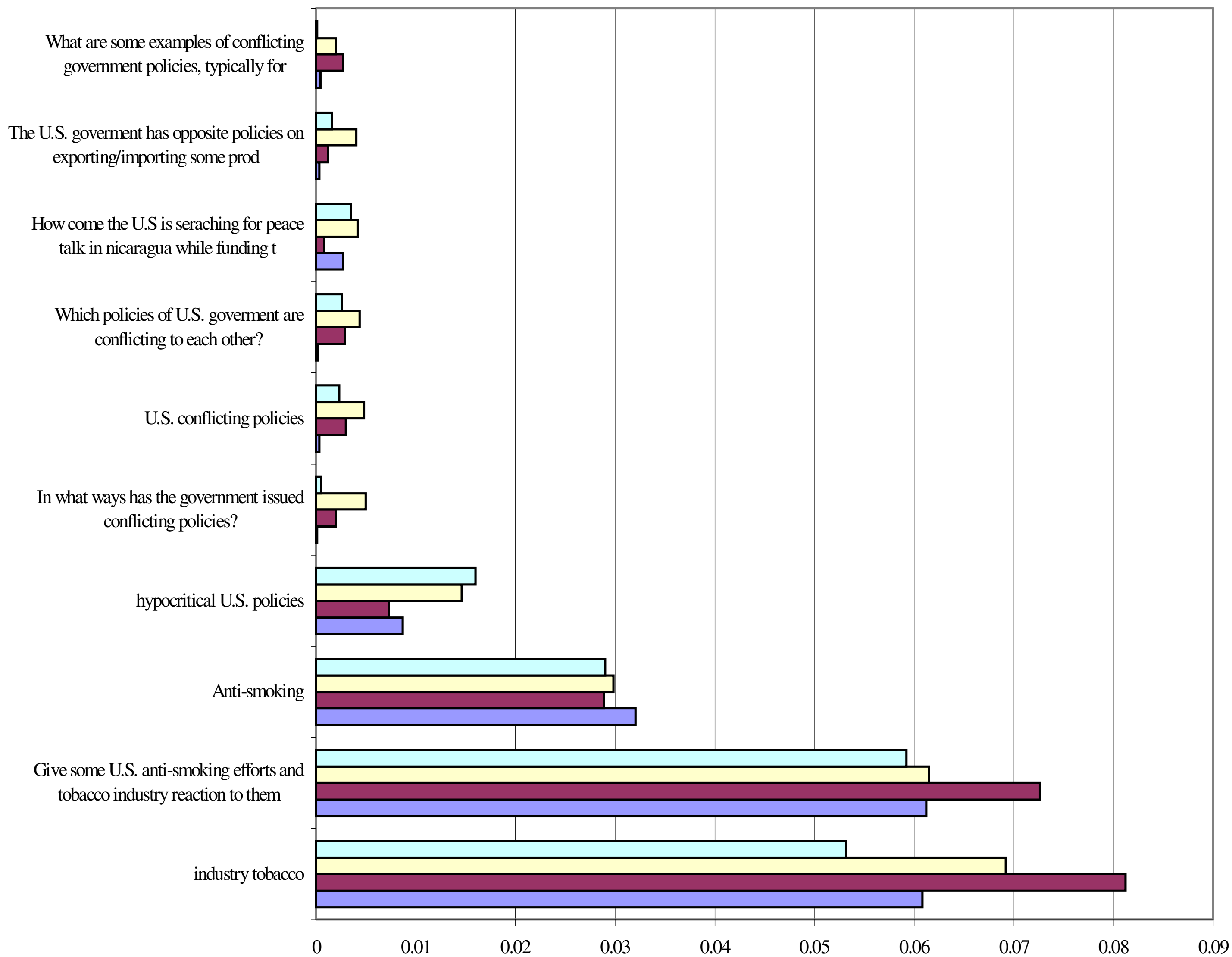
Topic 64



ANOVA

ce of Vari	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.2374	9	0.0264	3.8435	0.0031	2.2501
Columns	0.1078	3	0.0359	5.2333	0.0056	2.9603
Error	0.1853	27	0.0069			
Total	0.5305	39				

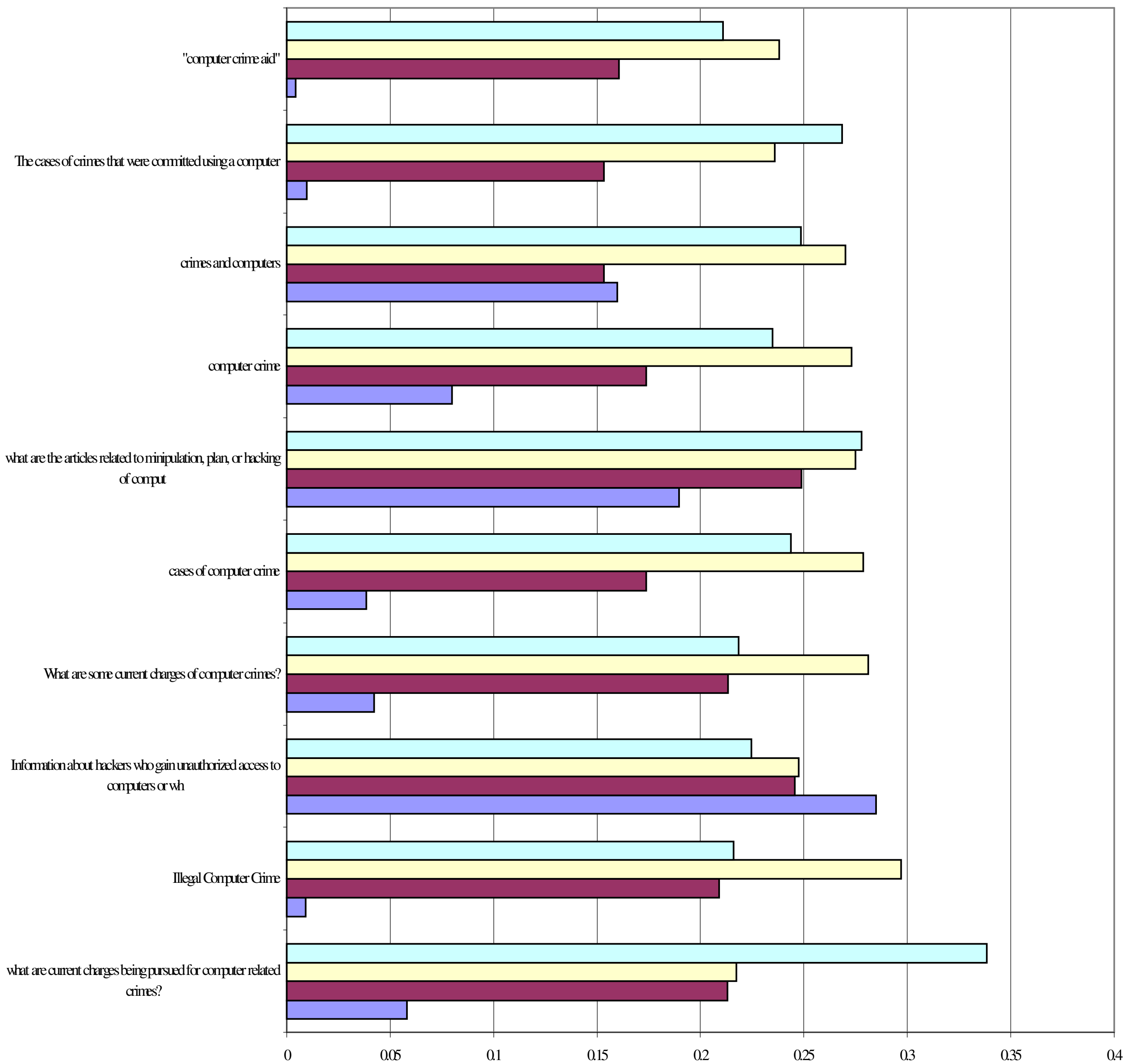
Topic 85



ANOVA

ce of Vari	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.0243	9	0.0027	134.5	5E-20	2.2501
Columns	0.0001	3	4E-05	1.8921	0.1548	2.9603
Error	0.0005	27	2E-05			
Total	0.025	39				

Topic 74



ANOVA

ce of Vari	SS	df	MS	F	P-value	F crit
Rows	0.0359	9	0.004	1.4476	0.2177	2.2501
Columns	0.1875	3	0.0625	22.681	2E-07	2.9603
Error	0.0744	27	0.0028			
Total	0.2978	39				

Topic 94

In Topic 64, all of the queries do well in general, but some of the systems do poorly for one or two queries. For example, the third system has problems with the hyphenated query "hostage-taking", handling it inappropriately here. This sort of analysis highlights the system 'blunders' well; showing clearly that a system has a problem with a particular query syntax.

Topic 85 is more interesting. It is another easy topic, but one where there is a large variation due to both systems and queries. Some systems are doing better than others at focusing in on the key words in the longer queries; the second system does better with the shorter queries while the fourth system likes the longer queries. All the systems do well with a good short query that is augmented by a specific concept like "bribery". Again, you can see the differences in the systems due to stemming and word order (phrases).

In Topic 74, the systems all behave the same (at a low level of performance), but the queries differ greatly. Performance improves as the queries shift from a general conceptual query, to a particular example. Obviously, this is a case where the topic itself is difficult.

Finally, in Topic 94 the systems are different, but the queries behave the same. The first three systems are all reasonably consistent across the queries, but the fourth system varies dramatically across queries.

In general, looking across all the topics, while using the 4 systems on the top 10 queries, we conclude that

- The queries provide a significant source of variance about half the topics.
- The top 4 systems are generally significantly different only due to "blundered runs"(e.g., stemming, hyphenation, spelling errors).

Looking at only the top 10 queries means we avoid the effect of "blundered queries". Most topics have one or two queries that are simply inappropriate for the topic. For example, query 51-06 in the earlier list of queries for topic 51 is such a blundered query; it talks about *the dispute* without ever mentioning that the dispute is *airbus subsidies*. However, restricting analysis to the top 10 queries also means we avoid hard, but good, variants of the topics.

If we do an analysis of variance for each topic working with the entire set of results (all queries and all systems), we find that queries and systems almost always provided significant sources of variation, with the variation due to query generally much higher than the variation due to system. But it is impossible draw any conclusions from this given the presence of blundered queries, and the fact that we had multiple versions of the same basic system for SMART and INQUERY engines that are designed to be at different levels of effectiveness.

3.2 Query Type Analysis

The 21 natural language queries can be broken apart based upon the original category of their formation.

	Number of queries in set	Average MAP
Short Queries	10	.227
Long Sentences	6	.209
Long Feedback Sent	5	.146
Long (overall)	11	.183

The short queries do noticeably better than the longer queries, contrary to what would normally be expected. Analysis done by Walter Liggett elsewhere concluded that the long queries are much more variable: often a long query is the best query version for a topic, but more often a long query is also the worst query version. However, it is hard to say whether this is really a length factor or just a query origination factor. Half of the short queries were done by experts and half by students, but only 1 out of the 11 long query sets were done by an expert. This question needs to be re-examined when this confounding factor can be removed.

RunSet	MAP
APL	.216
INQa	.167
INQp	.194
INQe	.229
Saba	.205
Sabm	.224
Sabe	.244
Acs	.147
Pir	.224

The table above gives the performance of the 9 system variations averaged across all the queries. Note the performance increase among the INQUERY and SMART (Sabir) systems as query structure and query expansion terms are added. The differences between the different versions of the same overall system are significant. The differences between the top 4 systems (APL, INQe, Sabe, Pir) are not significant. Note that the scores are much higher (ranging from .288 to .329) when averaged only over the top 10 queries per topic. These scores are much closer to the original TREC 1 scores, where systems had access to the entire long topic statement.

4 Conclusion

We've reaffirmed the tremendous variation that sometimes gets hidden underneath the averages of a typical IR experiment.

- Topics are extremely variable
- Queries dealing with the same topic are extremely variable. Even short queries were rarely duplicated (16%).
- Systems were only somewhat variable.

The lack of system variability could be due to the particular systems involved. They are all "bag-of-words" statistical systems, with the good systems all doing either implicit or explicit blind feedback query expansion. We need to repeat this experiment with more systems of different types.

We examined differences between using long or short queries. In this experiment, the short queries performed better. That could be because the particular systems being tested were not set up to take into account the relationships between query words that full sentences give you. On the other hand, students constructed almost all the long queries while experts constructed half of the short queries, so we could just be seeing a user experience effect. This experiment needs to be repeated.

We have started to analyze components of variance. However, there were a limited number of independent systems being tested. It is clear we need many more systems before we can reach conclusions here.

More systems would also be useful for learning to distinguish between a poor query, and a good query that is hard. The current operational definition of a good hard query is a query on which one system does well, but other systems do poorly. This implies enough information exists in the query, but that the state of the art is such that most systems cannot take advantage of the information. A collection of good hard queries might be especially useful for developing future systems.

The query collection as it exists is already a major resource for future experiments.

- One of the only query collections with spelling and other mistakes!
- Excellent test-bed for system tuning. Comparisons within a topic are valuable: what query syntax does a system not handle well?
- Provides a large number of queries (1150) with relevance judgments. This will be quite useful as systems start to do NLP analysis of queries.
- Provides repeatable, but non-identical, experiments in a controlled environment.

This last point may be especially valuable because it enables experiments of a type we have not been able to do before. If we view a particular retrieval task as responding to a given information need with a set of good documents (the relevant documents for that topic), we now have 23 different ways to accomplish that task (actually, a few less than 23 because of query blunders and duplication). We can start to study variability of approaches; are some approaches more stable than others? Eliminating topic variability from such studies is essential.

Analysis of the Query Track data has just begun; there is a wealth of data available. We encourage you all to play with the data and to add to it in future Query Tracks. Who knows what we will find in the future!