The NRRC Reliable Information Access (RIA) Workshop

Donna Harman
National Institute of Standards and Technology
Gaithersburg, Maryland, 20899
donna.harman@nist.gov

Chris Buckley
Sabir Research, Inc.
Gaithersburg, Maryland, 20878
chrisb@sabir.com

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: relevance feedback

General Terms

Experimentation, Measurement, Performance

Keywords

relevance feedback, failure analysis

1. INTRODUCTION

Current statistical approaches to IR have shown themselves to be effective and reliable in both research and commercial settings. However, experimental environments such as TREC show that retrieval results vary widely according to both topic (question asked) and system [2]. This is true for both the basic IR systems and for any of the more advanced implementations using, for example, query expansion. Some retrieval approaches work well on one topic but poorly on a second, while other approaches may work poorly on the first topic, but succeed on the second. If it could be determined in advance which approach would work well, then a guided approach could strongly improve performance. Unfortunately, despite many efforts no one knows how to choose good approaches on a per topic basis [1, 3].

The major problem in understanding retrieval variability is that the variability is due to a number of factors. There are topic factors due to the topic (question) statement itself and to the relationship of the topic to the document collection as a whole, and then there are system dependent factors including the approach algorithm and implementation details. In general, any researcher working with only one system finds it very difficult to separate out the topic variability factors from the system variability.

In the summer of 2003 NIST organized a 6-week workshop as part of the ARDA NRRC Summer Workshop series.[1]. The goal of this workshop (RIA) was to understand the contributions of both system variability factors and topic variability factors to overall retrieval variability. The workshop brought together seven different top research IR systems and set them to common tasks. Comparative analysis of these different systems enabled system variability factors to be isolated in a way that had never before been possible.

[1]This research was funded by the Advanced Research and Development Activity in Information Technology (ARDA), a U.S. Government entity which sponsors and promotes research of import to the Intelligence Community which include but is not limited to the CIA, DIA, NSA, NIMA and NRO

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. PARTICIPANTS

There were 28 people from 12 organizations.

- Carnegie Mellon University (Jamie Callan and students) using the Lemur system, a freely available very flexible research statistical IR engine.
- City University, London (Andy MacFarlane) used the Okapi system, a well-known probabilistic retrieval system.
- Clairvoyance (David Evans, David Hull, Jesse Montgomery) furnished 2 versions of the commercial, NLP x vector-space CLARIT system and AWB, an interactive tool useful for analyzing retrieval results.
- MITRE (Warren Greiff) contributed statistical analysis of data.
- NIST (Donna Harman, Ian Soboroff and Ellen Voorhees) organized the workshop.
- Sabir Research (Chris Buckley) designed most of the infrastructure to support the workshop, was the day-to-day leader during the six weeks and furnished the SMART IR system, Version 14.2.
- University of Massachusetts at Amherst (Andres Corrada-Emmanuel) also used the Lemur system, but with a different language modeling approach.
- The University of New York, Albany (Tomek Strzalkowski, Sharon Small, and students) used a SMART/HITIQA hybrid system based on NLP principles. A subcontractor (Paul Kantor) helped with statistical analysis of the data.
- University of Waterloo (Charlie Clark, Gord Cormack, and students) brought the MultiText retrieval system, made available two versions of Question Answering MultiText that had been used for TREC QA, and supplied WUI, a flexible browser based user interface for examining retrieved documents.

3. OVERVIEW OF WORKSHOP

The main data used for experimentation and failure analysis was the TREC ad hoc data from TRECs 6, 7, and 8 (topics 300-450). The original TREC relevance judgments were used, along with the associated document sets, but new runs were made for the workshop.

There were two main tracks to the RIA investigation of system and topic variability, one "bottom-up" and one "top-down". The bottom up track was a massive comparative failure analysis. Each of six systems contributed one representative run. Then for each of 45 designated topics, a detailed manual analysis of each run with its retrieved documents was done. The analysis goal was to discover why systems fail on each topic. Are failures due to system dependent problems such as query expansion weaknesses or system algorithm problems, or are the problems more inherent to the topic? For each topic, what would be needed to improve performance for each system? How can this be predicted by the system?

In the top-down track, the seven systems performed a large number of variations of a query expansion task. The majority of these experiments involved "blind" relevance feedback, where the systems used results from the set of top documents retrieved rather than using actual relevance judgments. In some sets of experiments, the systems changed their own tuning parameter settings. In other experiments, each system used as the source of expansion terms those from each of the other systems, or the actual expansion terms determined by other systems. The overall goal of the analysis was to isolate the system effect and discover why each system is succeeding in its query expansion efforts on each topic.

- bf_base – 4 runs per group using no feedback, standard feedback for each system, and feedback using the top 20 documents with 20 and 100 terms for expansion.
- bf_numterms – 37 runs per group using the top 20 documents but varying the number of terms for expansion from 0 to 100 (at given intervals).
- bf_numdocs – 36 runs per group expanding with 20 terms taken from varying numbers of top documents (1 to 100 at given intervals).
- bf_swap_doc – 8 runs, each using different sets of the top 20 documents (one set from each group), with terms then selected by each system.
- bf_swap_doc_term – same as above except each system picks the top 5 terms to exchange.
- bf_numdocs_relonly – 36 runs, similar to bf_numdocs except only the relevant documents in the top document set were used for expansion.
- bf_pass_numterms – similar to bf_numterms except that the top 20 passages (usually paragraphs) were used instead of the top 20 documents for expansion.
- bf_swap_doc_cluster, bf_swap_doc_hitiqa, bf_swap_doc_fuse – 3 small experiments varying the source of the documents to be swapped. The first uses FullClarit clusters, the second used HITIQA selections, and the third uses fusion of the top documents retrieved by the various systems.

4. SOME PRELIMINARY RESULTS

4.1 Failure Analysis

Prelimary results from the bottom-up failure analysis indicate that the root cause of poor performance on any one topic is likely to be the same for all systems. Except for six topics (out of 45), all systems fail for the same reasons, although to different extents. Using some of the tools from the workshop, it appears that the systems are retrieving different documents from each other in general, but all systems were missing the same aspect in the top documents.

The other major conclusion is that if a system can realize the problem associated with a given topic, then for well over half the topics studied, current technology should be able to improve results significantly. This suggests it may be more important for research to discover what current techniques should be applied to which topics, rather than to come up with new techniques.

4.2 Selected Blind Feedback Experiments

To give some idea of the type of issues being uncovered by the blind feedback experiments, here are some very preliminary results from several experiments. The workshop report `http://nrrc.mitre.org/NRRC/publications.htm` gives more details on the experiments and there are other posters in this SIGIR that present further work in these areas.

- bf_numterms – choosing the best number of query terms to add based upon the results can improve results as much as 30% over using a fixed number of terms for all queries.
- bf_swap_doc – Systems were very sensitive to the initial set of documents, with scores varying from as little as 10% to 50% depending on which set of initial documents were used for query expansion. Another surprising feature is how often systems prefer to use documents from other systems rather than their own documents.
- bf_swap_doc_term – The various systems chose quite different term lists even though they were dealing with the same document sources; only 15% to 25% of terms overlapped in general,

5. REFERENCES

[1] C. Buckley and J. Walz. The TREC-8 Query Track. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 65–76, 2000.

[2] D. Harman. What we have learned, and not learned, from TREC. In *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, pages 2–21, April 2000.

[3] S.Cronen-Townsend, Y.Zhou, and W.B.Croft. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.