SIGIR 2004 Workshop: RIA and "Where can IR go from here?"

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

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[i]. 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[ii,iii].

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 called Reliable Information Access (RIA) as part of the ARDA NRRC (Northeast Regional Research Center) Summer Workshop series. The goal of this workshop 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.

The SIGIR 2004 workshop focused on discussing the implications of lessons learned from RIA; how they affect our understanding of what is currently happening in research systems, and what they suggest are areas of IR research that warrant immediate concentrated work. Some of the most important lessons have to do with how to study IR systems, and how to make use of multiple IR systems to sharpen our understanding of the principles behind the performance observed. This is a significant change in IR research over the "single lab model" which has been dominant up to now.

One of the major products of the summer RIA workshop is a massive (over 40 GBytes) data collection, including intensive manual failure analyses on 45 TREC topics showing why each system failed on those topics. The database also includes over 3,000 total runs done by these top research systems over 9 different TREC test collections. This database is now publicly available at http://ir.nist.gov/ria and will hopefully become a dynamic resource for the research community.

Because the main goal of the SIGIR workshop was presentation to a wider audience of what occurred in RIA, and lots of discussion with the audience, the talks were all given by RIA workshop participants. Four groups presented their work done at the initial RIA workshop, their work since that workshop, and a list of questions to the audience to provoke discussion of how others see the issues.

It should be noted that whereas there are no proceedings from this workshop, the main SIGIR proceedings contain 7 poster papers covering the work at the RIA workshop. References to these papers are included later in this article.

Presentation Summaries

Introduction to the 2003 RIA workshop, presented by Donna Harman (NIST), provided background for the SIGIR workshop audience. She explained the motivation for the workshop, and introduced the two major sets of experiments done in the workshop. The first set of experiments involved a massive failure analysis of 45 of the 150 old TREC topics to gain a deep understanding of where the various systems either jointly or separately failed.

For the second set of experiments, 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 was succeeding (or not) in its query expansion efforts on each topic.

There is a poster paper (*The NRRC Reliable Information Access (RIA) workshop*, by D. Harman and C. Buckley, pages 528-529) in the SIGIR proceedings that provides further detailed background information on the RIA workshop.

Review of failure analysis work, presented by Chris Buckley (Sabir Research), both summarized the extensive failure analysis work done at RIA and presented further thoughts on how to categorize failures. To do the failure analysis in RIA, 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. Preliminary results 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 was 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 that 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. A poster paper (*Why current IR engines fail*, by C. Buckley, pages 584-585) in the SIGIR proceedings provides further details on this.

Review of the feedback experiment varying number of documents, presented by Jesse Montgomery and David Evans (CLARIT), detailed the first set of pseudo-relevance feedback experiments in RIA, those that tested the effect of the number of documents used in pseudo-relevance feedback (the *bf_numdocs* and *bf_numdocs_relonly* experiments). For *bf_numdocs*, there were 36 runs made per group expanding with 20 terms taken from varying numbers of top documents (1 to 100 at given intervals). The runs explored the fact that different systems had their optimal results using widely differing numbers of documents for mining of the terms. An additional set of experiments examined this phenomenon when only the relevant documents in those top-ranked set were used, as opposed to the full set. The details of this talk are mostly covered by two poster papers in the SIGIR proceedings (*The effect of varying number of relevant documents in blind feedback*, by J. Montgomery, L. Si, J. Callan and D.A. Evans, pages 476-477, and *A review of relevance feedback experiments at the 2003 Reliable Information Access (RIA) workshop*, by R.H. Warren and T. Liu, pages 570-571).

Review of the feedback experiments involving number and source of terms, presented by Paul Ogilvie (Carnegie Mellon U.), provided details of the second set of pseudo-relevance experiments in RIA, those that tested the effect of the number of terms used and also the source of those terms (the *bf_numterms* and *bf_pass_numterms* experiments). For *bf_numterms,* 37 runs per group were made using the top 20 documents but varying the number of terms for expansion from 0 to 100 (at given intervals). The bf_pass_numterms experiment was similar except that the top 20 passages (usually paragraphs) were used instead of the top 20 documents for expansion. One of the interesting results of the bf_numterm experiment is that choosing the optimal number of query terms can improve results as much as 30 % over using a fixed number of terms for all queries. Some of the details of this talk (the passage experiment) are contained in the poster paper *Comparison of using passages and documents in blind relevance feedback in information retrieval*, by Z. Gu and M. Luo (pages 482-483). Other details will be published in a special issue of *Information Retrieval* on the RIA workshop.

Review of the feedback experiments involving the swapping of documents or terms across systems, presented by Tom Lynam (U. of Waterloo), reported on three sets of experiments involving the swapping of objects across systems. The *bf_swap_doc* examined the use of 8 runs, each using different sets of the top 20 documents, basically one set from each group, with terms then selected by each system. The idea here was to see if some systems were better at picking the initial documents. It was found that systems were very sensitive to the initial set of documents, with improvement in scores varying from as little as 10 % to 50 % depending on which set of initial documents were used for query expansion. Another surprising feature was how often systems documents were better able to use documents from other systems rather than the documents in their own top-ranked set. The *bf_swap_doc_term* experiment was the same as the document swap experiment except that here each system picked the top 5 terms to exchange. It was interesting to note that 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. More details on these experiments can be found in *A multi-system analysis of document and term selection for blind feedback*, by T. Lynam, C. Buckley, C.Clarke, and G. Cormack, to be published in the proceedings of the 2004 CIKM conference.

Tour of new web site, including open discussion by C.Buckley (SabIR), presented a full tour of the new web site (http://ir.nist.gov/ria) containing all the results from the RIA workshop. The home page lists three major areas: results of the failure analysis, results from the 16 experiments, and an alternative way of looking at those results on a per topic basis.

The failure analysis section provides detailed analysis of 45 TREC topics, one section for each topic. Within those sections, there are the basic system parameters used by the 8 systems, including their input query (and weights), their results (trec_eval), and some basic statistics about retrieval in the various TREC document sets. Then there are individual failure analysis reports for each system and finally a summary report across all systems.

The experimental section provides the motivation for each experiment, the experimental setup, and then sections for each system and for analysis of the results. For each system, there is a detailed parameter description for each run, a result list (top 1000 documents retrieved), the trec_eval results, and a list of the input query that was used, along with the weights.

Conclusions

Although there were only 12 participants in the workshop initially, others dropped in to join the discussion. The group used this opportunity to make plans for a special issue of *Information Retrieval* to contain details of the RIA workshop and results found using the data since then. However a major part of the afternoon was spent refocusing on the summer workshop and creating a list of future experiments. This included experiments that could be done with the existing RIA data, such as investigating the existence of “golden” documents (the ideal ones for feedback) or the existence of “poison pills” (documents that cause poor performance in feedback). There was also a lot of discussion of ways to continue the investigations started in RIA, such as alternative ways of doing failure analysis or alternative ways of running some of the experiments. This list is ongoing and will be published in the special issue and continued on the web site. It is sincerely hoped that the data on the web site will provide a wealth of information for others to use in research, and that the web site will become a dynamic instrument to store multi-site investigations.

It should be noted that there are two additional poster papers in the SIGIR proceedings that address issues from the RIA workshop that were not presented at the SIGIR workshop. Please see *The effect of document retrieval quality on factoid question answering performance,* by K. Collins-Thompson, J. Callan, E. Terra, and C.L.A. Clarke (pages 574-575) for information on experiments in question answering, and *Topic prediction based on comparative retrieval Rankings*, by C. Buckley (pages 506-507), for some initial experiments on predicting when to use blind feedback.

[i] Harman, D.(2000). What we have learned, and not learned, from TREC. In Proceedings of the 2nd annual colloquium on information retrieval research, 2--21.

[ii] Buckley, C. and Walz, J. (2000). The TREC-8 query track. In Proceedings of the eighth Text REtrieval Conference (TREC-8), 65--76.

[iii] Cronen-Townsend, S. Zhou, Y. and Croft, W.B. (2002). Predicting query performance. In Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, 299--306.