Inf Retrieval (2009) 12:615–641
DOI 10.1007/s10791-009-9101-4

Overview of the Reliable Information Access Workshop

Donna Harman · Chris Buckley

Published online: 18 July 2009

Abstract The Reliable Information Access (RIA) Workshop was held in the summer of 2003, with a goal of improved understanding of information retrieval systems, in particular with regard to the variability of retrieval performance across topics. The workshop ran massive cross-system failure analysis on 45 of the TREC topics and also performed cross-system experiments on pseudo-relevance feedback. This paper presents an overview of that workshop, along with some preliminary conclusions from these experiments. Even if this workshop was held 6 years ago, the issues of improving system performance across all topics is still critical to the field and this paper, along with the others in this issue, are the first widely published full papers for the workshop.

Keywords Information retrieval · Relevance feedback · Failure analysis

1 Introduction

The field of information retrieval has always closely modeled the application of a person seeking information. As librarians (or Google watchers) well know, there is not only a wide variety in the types of information that users seek, but a huge variation in how those users express their needs in a query. This variation is natural, and therefore successful evaluations of information retrieval systems must mirror this in test collections by having large numbers of test questions, hopefully from a "natural" source. The early Cranfield collection came with 225 test questions; the current TREC collections also have large numbers of test questions (called "topics" in TREC).

Despite the wide variety in the topics used in TREC, the graph in Fig. 1 shows that the average retrieval effectiveness approximately doubled in the first 7 years of TREC. This

D. Harman (✉)
National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
e-mail: donna.harman@nist.gov

C. Buckley
Sabir Research Inc., Gaithersburg, MD 20878, USA
e-mail: chrisb@sabir.com

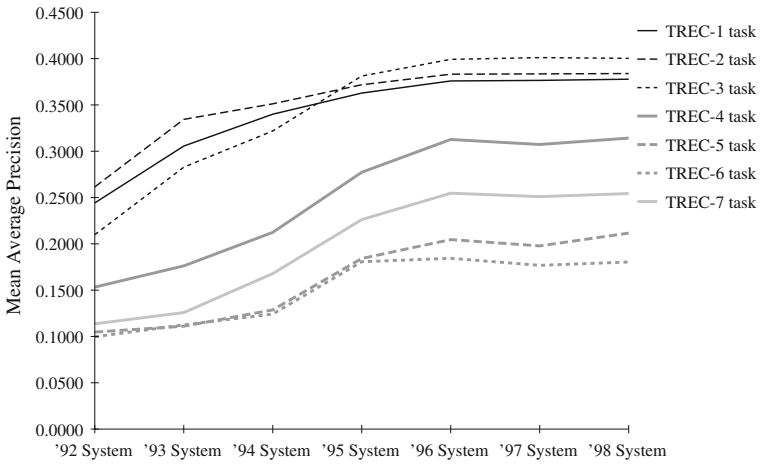

Fig. 1 Retrieval effectiveness improvement for Cornell's SMART system, TRE C-1–TREC-7

means, for example, that retrieval engines that could retrieve three good documents within the top 10 documents in 1992 were now likely to retrieve six good documents in the top 10 documents retrieved for the same search. The figure plots retrieval effectiveness for one well-known early retrieval system, the SMART system of Cornell University. The SMART system was consistently one of the more effective systems in TREC, but other systems were comparable with it, so the graph is representative of the increase in effectiveness for the field as a whole.

Figure 1 also shows a flattening of the improvements by TREC-7. Note that in general this flattening appeared for all of the systems and there was considerable discussion as to the cause of this performance ceiling. One issue is simply that researchers put more effort into the new tasks being run in the later TRECs, such as cross-language retrieval or web searching. But there was agreement that a major factor in this flattening or ceiling effect is the extremely large variation in performance across topics. This variation has been a problem since the beginning of research in information retrieval in that techniques that work well for one topic do not work well for others, leaving no improvement in performance on average. In the early TRECs, new techniques such as better weighting and pseudo-relevance feedback improved performance on most topics, therefore improving the averages. However, at some point, there were no new ideas that seemed to improve performance for the majority of topics—hence the flat curves.

Topic variation is reflected in many ways such as:

1. a wide variation across topics in the average precision score for the best performing system,
2. a wide variation in performance across topics for a given system (or system variant),
3. a wide variation in performance across topics of the effectiveness of particular devices such as relevance feedback,
4. a wide variation between two system variants with respect to the rank of the same retrieved document.

Figure 2 clearly illustrates the first two of these variation problems. First, the performance of the best system for each of the 50 topics varies from almost perfect performance to an average precision of barely 0.1. Past experiments (Voorhees and Harman 1997) have

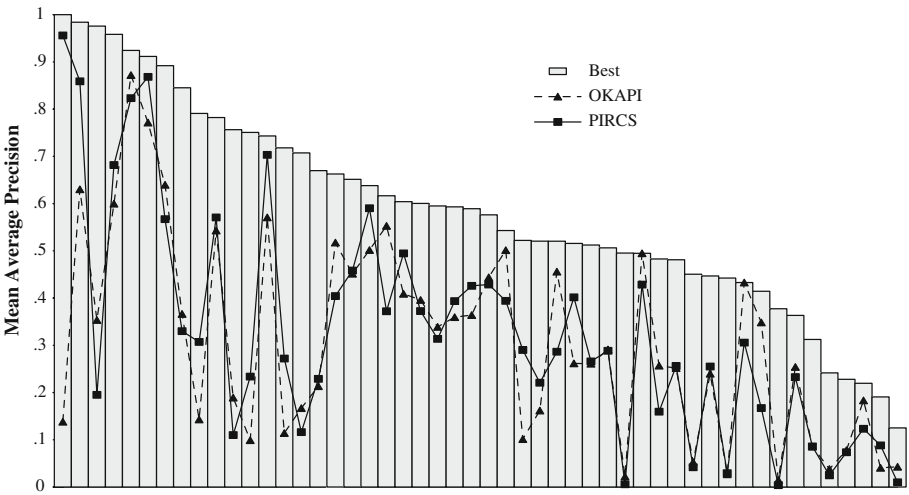

Fig. 2 Performance variations across topics, TREC-8

shown that this performance variation is not correlated with the number of relevant documents for a given topic, but is some function of the interaction between the topic, the document set being searched, and the retrieval system. When specific systems are examined, a second source of variation can be seen in Fig. 2. The results for the OKAPI system in TREC-8 show a wide variation in performance scores across the different topics, and this variation is not correlated with the performance of the best system, other than it is bounded by those results. Additionally, examination of a different system, such as the PIRCS system, shows the same types of variations, but with performance different than both the best system and the OKAPI system.

Table 1 illustrates the third example of topic variation. The table shows the number of topics that had the best performance using different topic input lengths (full topic, description only and title only) for three different systems. Further examination of the data reveals that topics that work best at a particular length for one group did not necessarily work best at that length for the other groups.

Because retrieval approaches can work well on one topic but poorly on another, determination in advance of which approach would work well for a given topic would allow tailoring of the systems to each topic. Unfortunately, despite many efforts (Cronen-Townsend et al. 2002; Yom-Tov et al. 2005), no one knows how to choose good approaches on a per topic basis. The major problem in understanding retrieval variability is that it is caused by a number of factors. There are topic factors due to the topic statement itself and to the relationship of the topic to the document collection as a whole. There are system dependent factors including the specific algorithms and implementation details. In general a researcher is working with only one system and thus finds it very difficult to separate out the topic variability factors from the system variability.

Table 1 Number of TREC-7 topics performing best by topic part

	Full	Desc.	Title
OKAPI	28	13	9
PIRCS	27	10	13
SMART	22	17	11

The goal of the Reliable Information Access (RIA) Workshop was to understand the contributions of both system variability factors and topic variability factors to overall retrieval variability. Comparative analysis of the different systems was to enable system variability factors to be isolated in a way that never before had been possible. The workshop was sponsored by ARDA in their summer workshop series.

2 Workshop description

Because of the complexity of the problem, it was critical that the workshop be highly focused; additionally experiments needed to concentrate on techniques that are common to all the systems. Note that almost all information retrieval systems use term occurrence statistics in some manner as a core of their systems, with the common technique of matching the words in the input questions against words in the documents. In general this implies that improvements must come from either re-weighting the importance of existing word matches, or from adding new words to the query that can be used for matching. Thus query expansion has been a central focus of statistical information retrieval throughout its research history, and is the only technique that has been consistently shown to improve performance *on average*. However while query expansion works well on average, there are several different mechanisms that could cause this improvement. Systems are in effect tuned to emphasize some choice(s) of these mechanisms, such as different term weighting methods, different query expansion methods, etc.

In a pre-workshop meeting in March of 2003, it was decided to focus the workshop investigation on one type of query expansion, that of pseudo-relevance feedback (also called "blind" feedback). This expansion works on the assumption that the initial top-ranked documents are relevant and uses these documents in the feedback process. The documents can then be mined for expansion terms or for re-weighting of existing terms or both. Between March and June, the various systems were installed at MITRE (the location of the workshop) and discussion continued on the details of what would be done during the 6 weeks. It should be noted that an additional part of this workshop was to investigate the relationship of improved retrieval as input to a question-answering system. This part of the workshop is not further covered here; see (Collins-Thompson et al. 2004) for more on this.

The final organization of the RIA workshop featured two approaches to the investigation of system and topic variability—a massive comparative failure analysis and a series of tightly controlled experiments examining variants of pseudo-relevance feedback.

For the massive comparative failure analysis, each system contributed one representative run. Then, for 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. Were failures due to system dependent problems such as query expansion weaknesses or system algorithm problems, or were the problems more inherent to the topic? For each topic, what would be needed to improve performance for each system? How could this be predicted by the system?

For the controlled set of experiments, the systems performed a large number of variations in the pseudo-relevance feedback technique. 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 documents from each of the other systems, or used 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 in its query expansion efforts on each topic.

For each of these two approaches the workshop participants collected enormous amounts of data. Only a small portion of the analysis of the data could be completed during the workshop. The preliminary analysis that has been done has already produced a number of surprising results. The entire collection of data has been released to the community (http://ir.nist.gov/ria), and hopefully will enable useful research for years to come.

By its very nature, the RIA workshop required participation from a large number of groups and experts. Bringing together seven of the top research systems in one location with both high-level theoretical expertise and also practical system expertise was difficult, especially given the 6-week duration of the workshop. There were two groups of participants; the senior experts who generally were present for 1–2 weeks of the workshop spread out over several trips, and the graduate students who for the most part were at the workshop for the full 6 weeks. Altogether, there were 28 people from 12 organizations that participated. The seven systems represented at RIA were CMU (from Carnegie Mellon University); City (from City University, London); CLJ and FullCL (from Clairvoyance Corporation); Sabir (from Sabir Research); UMass (from University of Massachusetts at Amherst); Albany (from University of New York at Albany); and Waterloo (from University of Waterloo), with the workshop being coordinated by NIST and held at MITRE Corporation.

The Appendix gives the organizations, people, and software that contributed to the workshop, along with detailed descriptions of each system as written by the participants. Note that this was an open workshop environment where everybody was constantly contributing ideas and efforts. As well as working with their own research systems, most graduate student participants were also in charge of several of the daily failure analysis sessions and one or two of the system experiments.

This paper starts with a short summary of the failure analysis part of the workshop, followed by summaries of each of the controlled experiments. Section 5 is a summary of the data that was collected and that is available on the website. Section 6 gives some very preliminary results from initial efforts at the workshop to develop automatic ways of categorizing topics; this is included mainly as a prompt for further experimentation by others. The paper concludes with a retrospective summary of lessons learned in terms of how to organize and run such a workshop, and also a set of suggested experiments to continue this work.

3 Massive comparative failure analysis

The failure analysis investigation was an attempt to discover why current research information retrieval systems fail and to propose concrete areas of concentrated research to improve effectiveness. What follows is a short summary; readers are referred to the paper in this issue (Buckley in press) for details and results. During the March pre-workshop meeting it was decided that all groups would submit a *standard* retrieval run that in some sense was representative of their group's approach to IR. There were no restrictions on what could be in the run as long as it was completely automatic. These runs became the basis of the failure analysis.

This failure analysis was a major activity of the workshop; with 90 min to 2 h per day allocated for the individual and group analysis. After a few false starts, a standard procedure was adopted, using a wide variety of tools. The major tool was the Waterloo User Interface, which allowed a user to view documents that either were relevant, but not retrieved in a top set, or that were non-relevant, but were retrieved in the top set. Given the

large time requirements for failure analysis (from 11 to 40 person-hours per topic), it was obvious that not all 150 topics could be examined (only 45 topics were actually finished). It was decided to focus on topics where the systems in general scored below the overall MAP average and where there was a large variance among system scores.

The first conclusion of the failure analysis was that the root cause of poor performance on any one topic was likely to be the same for all systems. Whereas the systems were retrieving different documents in general, all systems were missing the same aspect in the top documents. The other major conclusion was that 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 than to come up with new techniques. Again, for full details behind these conclusions, see the paper in this issue (Buckley in press).

4 Controlled retrieval experiments

4.1 Design of experiments

The retrieval experiments in the RIA workshop were a large investigation into how different systems vary while performing a single query expansion task, that of pseudo-relevance feedback. Pseudo-relevance feedback was chosen as the target task for several reasons. First, it is known to have a high degree of topic variance; within any one system it works very well on some topics but hurts performance on other topics. Most systems find a mild average benefit to the use of pseudo-relevance feedback. Secondly, most systems have used it at some point in their research; thus the implementation effort required for experimentation was minimized. And, finally, it has a number of important parameter settings that systems in practice set to different values, and that can be changed easily.

In a typical pseudo-relevance feedback task, systems automatically expand the original query by adding terms that occur in documents (or passages) that the system determined were closely related to the query. On each topic, a system

1. Performs an initial retrieval with terms from the text of the original topic,
2. Without any user looking at them (thus "pseudo-relevance"), the system assumes that the top X documents were responsive to the topic and would be useful for expansion,
3. The system chooses N terms from the top X documents and adds them to the original query terms,
4. All terms are reweighted,
5. The new expanded query is re-run against the entire document collection, and a ranking of the top documents is produced,
6. In a live system, these documents would then be given the user. In the experimental setting, the ranking is evaluated based on the ranks of known relevant documents.

For these retrieval experiments, variations of each of the possible parameter choices were studied. These included the number of documents to draw expansion terms from (X), the number of expansion terms to add (N), the choice of the expansion documents, and the choice of the expansion terms. There is an inherent system performance of each system due to their weighting, indexing, and matching algorithms. The major goal of the analysis was to see if the variability due to topics could be separated from that inherent system-dependent variability. Different expansion approaches work well on different topics. If it is possible to isolate the topic-dependent effect, then the factors that are discovered can

determine the success of an expansion approach and each system can adjust its approach and parameters based upon those topic dependent factors.

Somewhat more formally, evaluation scores can be explained in terms of the topic, the inherent system, and the run (system parameter settings).

$$p(t, s, r) \sim et + es + er + esr + etr + est + estr$$

where $p(t, s, r)$ is the score; t, the topic; s, the system; r, the run; et, the topic effect; es, the system effect; er, the run effect; esr, the effect of the interaction between system and run; etr, the effect of the interaction between topic and run; est, the effect of the interaction between system and topic; and $estr$, the interaction of all three parameters, which is ignored here.

In the basic sets of experiments, there were altogether 150 topics, seven systems, and about 100 different runs for a total of 105,000 data points. One goal of the experiments was to look at *etr*, the interaction of the topic and run. This could be used to classify topics according to what sort of approach and parameters should be used. Ideally, this classification could be matched to a classification based on topic information alone. In that case, there would be an effective decision procedure for how to choose the approach and parameters on a per topic basis.

Another major goal of these retrieval experiments was simply to increase the understanding of what is happening with query expansion and pseudo-relevance feedback. Most research groups have experimented extensively with pseudo-relevance feedback at some point or another, but because pseudo-relevance feedback is so topic and system dependent, it has been very hard to analyze why it works or doesn't work on particular topics. Most groups have been content to just optimize for maximum average performance.

When query expansion improves performance, it tends to be because one or more of the following is added:

1. better weighting to original query terms
2. synonyms
3. one or two good related words
4. a large number of related words that establish that some aspect of the topic is present (context)
5. specific examples of general original query terms

It is very likely that each of the five effects is of primary importance to some set of topics but not to other sets. Until it is known how important each of these effects is, the systems cannot adjust to improve expansion performance. The goal here was to understand for a system what worked for individual topics as compared to all other approaches that this system or other systems tried. Given the problems caused by topic variability, it is much easier to compare against other system results than to attempt to judge whether an approach succeeded or failed on some absolute basis.

4.2 Brief descriptions of each experiment

Each of the retrieval experiments done during the workshop is briefly described below. There was very little time for analysis of the experiments during the workshop, but included in each section is a summary of what has been written in later publications. Readers should refer to these publications for more information.

Each experiment listed below includes a brief description, the experimental goal, the leader and the participating systems, the basic methodology, a summary of the results (and

reference to other publications on these results), and some suggestions for further analysis. Note that these suggestions were made at the time of the experiments and therefore represent excellent leads into further research.

TREC data (http://trec.nist.gov) was used in the workshop, with most of the work being done with the 150 topics created for the ad hoc tracks in TRECs 6, 7, and 8 (topics 301–450), against the TREC disks 4 and 5 (without the Congressional Record sub-collection which was used only in TREC 6). This topic set is usually considered the "best" one for experimentation, both because the topic generation methodology used in TREC was stable by this point and because it is the only set with 150 topics against the same data. Note that additional runs were made as part of the database collection for other sets of topics (see Sect. 5). In general the description part of the topic was used for experimentation, with each system using their "normal" stopword and stemming techniques.

4.2.1 bf_base

- Description: Basic investigation of pseudo-relevance feedback
- Goal: Establish whether pseudo-relevance feedback works for the participating systems
- Leader: Andres Corrada-Emmanuel
- Participants: All 8 systems (2 from Clairvoyance)
- Methodology: Perform 4 runs per group:
  1. No feedback at all; initial retrieval (bf.0.0)
  2. Standard pseudo-relevance feedback run of system with whatever parameters the system normally uses (bf)
  3. Set the number of documents used for feedback to 20, and the number of expansion terms to 20 (bf.20.20)
  4. Set the number of documents used for feedback to 20, and the number of expansion terms to 100 (bf.20.100)
- Results and Comments:
  All groups got reasonable average performance increases of between 10 and 20% using expansion (see Table 2). Some groups got mildly better performance expanding by a lot of terms as opposed to a few; other groups got mildly worse scores.
  The parameters used for the standard bf run, where each system could choose its own parameters, varied widely as can be seen in Table 3. Systems such as CLJ, which tended to add very specific terms, used comparatively few documents and terms, while systems such as UMass, which added more general terms, used more documents and added more terms. CMU added a different number of terms for each topic, averaging an additional 412 terms per topic.
- Future Analysis: none suggested at the workshop

4.2.2 bf_numdocs

- Description: Vary the number of documents from which added terms are extracted in a pseudo-relevance feedback expansion
- Goal: Along with bf_numterms, one of the two major experiments in pseudo-relevance feedback parameterization

Table 2 MAP scores for bf_base runs

	bf.0.0	bf	bf.20.20	bf.20.100
Albany	0.126	0.154	0.139	0.154
City	0.186	0.216	0.213	0.193
CLJ	0.185	0.210	0.209	0.192
CMU	0.201	0.225	0.217	0.218
FullCL	0.169	0.188	0.196	0.196
Sabir	0.204	0.226	0.226	0.225
UMass	0.196	0.235	0.220	0.234
Waterloo	0.198	0.228	0.215	0.211

Table 3 Parameter choices for standard bf run

	Number of documents	Number of added terms
Albany	20	100
City	10	20
CLJ	6	30
CMU	10	Hundreds
FullCL	6	30
Sabir	20	60
UMass	30	100
Waterloo	25	25

- Leader: Jesse Montgomery
- Participants: All 8 systems
- Methodology: Perform 36 pseudo-relevance feedback runs, expanding by 20 terms taken from a variable number of top documents. Start by considering 1 top document, then 2, 3,...,20, 25, 30,...,100
- Results: The short paper presented at SIGIR2004 (Montgomery and Evans 2004) discussed the following major results:
  - Each system had an optimal number of documents to be used for feedback, i.e., a single peak occurred in mean average performance (MAP). However this optimal number differed across the systems (most happened between 10 and 20 documents used in feedback).
  - Some systems were more sensitive to using further documents. For example City and Sabir had more performance degradation as additional feedback documents were added, whereas UMass and CMU had little degradation.
  - There was no simple relationship that could be found between the optimal number of documents used for feedback and several obvious factors in the topics, such as the initial input query length and the number of relevant documents for the topic. Additionally there was no discernable pattern for any combination of these topic characteristics.

- Future Analysis:
  - Topics could be categorized by how often using more documents helped performance, with that categorization possibly correlated with categorization by how many terms helped performance.
  - It would be interesting to categorize topics by what percentage of the top documents should be relevant in order for feedback to help. The bf_numdocs_relonly experiment described later shows that if all documents used in feedback are relevant, then performance will increase as documents are added. Is there a percentage threshold above which adding more documents is expected to help?
  - As well as number of documents, are there particular documents that in general helped pseudo-relevance feedback across all systems? Are there documents that hurt pseudo-relevance feedback across systems even though they are relevant? Could these documents that either help or hurt be characterized?

4.2.3 bf_numdocs_relonly

- Description: Vary the number of potential documents from which added terms are extracted in a pseudo-relevance feedback expansion, but actually add only relevant documents
- Goal: This is a paired experiment with bf_numdocs. The goal was to determine how much the non-relevant top documents hurt the expanded query.
- Leaders: Rob Warren, Ting Liu, David Evans
- Participants: All 8 systems
- Methodology: Perform 36 pseudo-relevance feedback runs, expanding by 20 terms taken from a variable number of top documents. Start by considering 1 top document, then 2, 3,...,20, 25, 30,... ,100. For each run, delete all non-relevant documents from the top documents before query expansion. Thus, if the initial retrieval for a topic contains no relevant documents between ranks 11 and 20, then the 10 retrieval runs for sets 11 through 20 will be identical for that topic.
- Results: This is an upper-bound experiment. Among other things, it simulates having an actual user making relevance judgments from a set of top documents of size N, and using only those relevant documents for feedback. As would be expected, all systems have a slow, monotonic growth in MAP as the size of the candidate set of documents increases. The upper limit of MAP differs substantially among systems. For example, CMU had an upper limit MAP of 0.292, Waterloo had 0.316, and Sabir had 0.370. This gap is enormous; and should shed some light on differences between systems once it is fully understood.
- A short paper at SIGIR2004 (Warren 2004) discussed the following additional results.
  - Incremental benefits in performance seem to diminish after six relevant documents have been used for feedback.
  - Using a large number of relevant documents for feedback usually lowers system performance.
  - The use of some specific relevant documents clearly hurt performance for all systems when they are used as a source for query expansion terms.
- Future Analysis: It would be interesting to investigate if there is any way to automatically determine that a specific relevant document will hurt performance if it is used for expansion (or re-weighting).

4.2.4 bf_numterms

- Description: Vary the number of terms added to the original query by pseudo-relevance feedback expansion
- Goal: Along with bf_numdocs, one of the two major experiments investigating pseudo-relevance feedback parameters and variability
- Leader: Paul Ogilvie
- Participants: All 8 systems
- Methodology: Perform 37 pseudo-relevance feedback runs with expansion based on the top 20 documents. Start by adding 0 terms (just reweight original topic) then add 1 term, 2 terms,...,20 terms, 25 terms 30 terms,...,100 terms
- Results: Average behavior was different for each of the systems. This issue contains a more detailed analysis of this experiment (Ogilvie et al. in press).
  - All systems kept on improving on average as the number of terms increased from 0 to 15. As the number of terms continued to increase, some systems mildly improved further, other systems got worse. An oracle that chooses the best number of query terms to add based upon the results can improve results as much as 30%.
  - On a per topic basis, the systems with continuous improvement as number of terms increased tended to have a bi-modal distribution, i.e., either near 0 terms should be added or nearly 100 terms should be added.
  - Topics can be categorized by counting the number of added terms in the top 20 which actually improved performance as opposed to not adding that term. Strong improvements overall in expansion were strongly correlated with five or more helpful terms being added. Term expansion did not help strongly for any topic in which most systems agreed that only one to four terms should be added.
  - The above two points suggest that improvements across systems are coming from ensuring the context of the topic is represented in the documents, rather than in adding a small number of good synonyms, examples, or related terms. But this needs to be analyzed much more thoroughly.
- Future Analysis: This is the major experiment which needs to be understood on a per topic basis in order to understand pseudo-relevance feedback expansion. The topic categorization based on number of helpful terms needs to examined carefully, and compared against all the other topic categorizations.

4.2.5 bf_pass_numterms

- Description: Vary the number of expansion terms added to the original query in a pseudo-relevance feedback expansion. The initial retrieval was of passages rather than entire documents, thus there was considerably less text but presumably more focused areas to serve as the source of expansion terms.
- Goal: Understand how passage retrieval differs from document retrieval in the expansion process.
- Leader: Zhenmei Gu, Ming Luo
- Participants: 4 systems—City, CMU, Waterloo, FullClarit
- Methodology: The same methodology as the bf_numterms experiment, except each system returned a passage instead of the entire document. Each system had its own

definition of passage; the only enforced requirement was that a set of passages be non-overlapping.
- The FullClarit and Waterloo systems already expand queries by considering passages; Their runs were unchanged from the bf_numterms experiment. For the CMU system, a passage was defined as a text fragment of 100 words. The passages in the City system were of varied lengths up to a maximum of 10 sentences.
- Results: Both CMU and City got very mild average improvement (1–2%) over the corresponding bf_numterms runs when averaged over all 36 runs. One general observation was that the per topic performance with passages was more variable as the number of terms increased; possibly because rarer terms were being added from the passages as opposed to those that could have been added from the documents.
- A short paper presented at SIGIR2004 (Gu 2004) added the following observations:
  - A table showing that on a per topic basic, both CMU and City had improved performance (marginally) using passages for about 50% of the topics
  - CMU showed consistent (but marginal) improvement using passages when adding up to 100 terms; the City runs improved only when the number of feedback terms were small. One conjecture is that as City tried to draw more and more terms from twenty small passages, it could no longer find good terms.
  - Using passages for feedback tends to work better for topics in which the relevant documents have an average length that is shorter or longer than the mean average relevant document length for all topics.

4.2.6 bf_swap_doc

- Description: Each system used the top documents found by initial runs of other systems instead of using its own initial run.
- Goal: Determine how much the initial retrieval strategy of each system affects whether pseudo-relevance feedback works.
- Leader: Tom Lynam
- Participants: All 8 systems
- Methodology: All 8 groups prepared a list of their initial 60 retrieved documents in TREC results format. Each group then did 8 pseudo-relevance feedback runs, using a subset of each other's list of initial retrieved documents as the source of expansion terms, but using their own methods and default parameters to select documents and to choose and weight terms. At the end, each group had done a retrieval run based on (some of) Albany's top documents, a run based on City's documents, and so on for all 8 groups.
- Results and Comments: A separate paper in this issue (Clarke et al. in press) provides the details of this experiments, including results and analysis. Two major surprises were that some systems are much more sensitive to the initial set of documents than others, and that very often systems prefer to use documents from other systems rather than their own documents.
- Future Analysis: The effects of swapping documents is complex: there is a need to look much more closely at the characteristics of the topics for which swapping top documents made a large difference. It was not a question of just the number of relevant documents being considered.

4.2.7 bf_swap_doc_term

- Description: Each system used both top documents and expansion terms found by other systems instead of using their own documents and terms.
- Goal: Determine how much term selection algorithms of each system affect whether pseudo-relevance feedback works.
- Leader: Tom Lynam, Ting Liu
- Participants: 7 systems participated (CLJ did not)
- Methodology: This was a challenging experiment to perform (and explain). Please see the paper in this issue (Clarke et al. in press) for the detailed methodology and results.
- Future Analysis: There has been no topic analysis or categorization done for these runs. It would be interesting to examine those topics for which choice of terms does make a difference.

4.2.8 bf_swap_doc_cluster; bf_swap_doc_hitiqa

- Description: These were the first two of three small experiments in which the source of documents from which expansion terms were drawn was chosen using some outside criteria. The third experiment (bf_swap_doc_fuse) is reported in the swapdocs paper in this issue (Clarke et al. in press).
- Goal: Investigate the effect that criteria other than initial retrieval have on expansion performance.
- Leaders: Jesse Montgomery (bf_swap_doc_cluster), Sean Ryan (bf_swap_doc_hitiqa)
- Participants: 5 systems—Albany, City, FullClarit, Sabir, Waterloo
- Methodology (bf_swap_doc_cluster): This experiment was an upper bound experiment, clustering the retrieved set and choosing the cluster with the most relevant documents. The documents are from a FullClarit initial run where the top N documents are clustered by the FullClarit system, and the best cluster is chosen. There were two runs made with values for N being 50 and 100. For $N = 50$, the number of documents per topic ranged from 2 to 45. For $N = 100$, the number of documents per topic ranged from 2 to 73.
- Results (bf_swap_doc_cluster): The results are shown in Table 4. The most interesting point was that Waterloo was able to take advantage of the good clusters of documents, much more than other systems. The conjecture is that the Waterloo expansion by passages within each document was able to pick out a common good text piece that was responsible for both relevance and the clustering.
- Methodology (bf_swap_doc_hitiqa): For this experiment, the base initial set of documents was obtained by using the HITIQA NLP system to index and cluster a given initial set of documents. The HITIQA system matches passages against the query in a frame-based manner. The passages are then clustered with the documents being provided to other systems being the documents containing those clusters. Systems used all of the documents returned by HITIQA. The number of documents ranged from 3 to 72 per topic.
- Results (bf_swap_doc_hitiqa): The results are shown in Table 5. Overall, the results were below standard pseudo-relevance feedback runs. One factor affecting performance was that although HITIQA did a good job at finding good passages in long documents, this passage information was then thrown away and systems were only given the long documents themselves. With the exception of the passage-based

Table 4 MAP scores for bf_swap_doc_cluster

	$N = 50$	$N = 100$
Albany	0.204	0.221
City	0.236	0.236
FullCL	0.222	0.240
Sabir	0.224	0.249
Waterloo	0.255	0.271

Table 5 MAP scores for bf_swap_doc_hitiqa

	bf.hitiqa
Albany	0.166
City	0.197
FullCL	0.189
Sabir	0.179
Waterloo	0.220

Table 6 Summary of pseudo-relevance experiments

Experiment	Systems	Runs	Feedback	Documents	Terms	Comments
bf_base	8	4	no	0	0	
			Yes	Varied	Varied	System choice
			Yes	20	20	
			Yes	20	100	
bf_numdocs	8	36	Yes	[1,100]	20	
bf_numdocs_relonly	8	36	Yes	[1,100]	20	Use only relevant for feedback
bf_numterms	8	36	Yes	20	[1,100]	
bf_pass_numterms	4	36	Yes	20	[1,100]	Use passages
bf_swap_doc	8	64	Yes	Varied	Varied	System choice from 60 documents
bf_swap_doc_term	7	49	Yes	Varied	Varied	System choice from 60 documents
bf_swap_doc_cluster	5	10	Yes	Varied	Varied	FullClarit clusters
bf_swap_doc_hitiqa	5	5	Yes	Varied	Varied	HITIQA clusters

Waterloo system, the long documents proved less useful. There was not enough time to repeat the experiment in a passage environment.

Table 6 shows a short summary of all of the controlled experiments.

5 Run database

One of the major resources for future research produced by the workshop is the database of runs. This database is now stored on the NIST system (http://ir.nist.gov/ria) and a paper in this issue (Soboroff in press) describes the web site in detail.

Each group produced well over a hundred evaluated retrieval runs on the standard collection of 150 topics used in TRECs 6, 7, and 8, as described in the previous section. Then the major experiments were all rerun (replicated) for each group on the TREC 5 ad

hoc task, about 95 runs. In addition, 2 key experiments (bf_numdocs and bf_numterms, about 73 runs) for each group were replicated on each of the TREC ad hoc tasks from TREC 1, 2, 3, 4. Finally, one run was made for each group on the merged document collection formed from the news articles in TRECs 1–8, using all available topics (1–450). Altogether, there are 4,088 run results in the database, taking up over 22 gigabytes of disk space (Zhenmei Gu and Luo Ming were responsible for the run replications).

The replicated runs have not yet been examined in any detail; that lies in the future. The main purpose of the replicated runs was to validate the experimental analysis done on the results from the standard collection to verify that the experimental conclusions are not dependent on the particular topics and documents of the standard collection, but hold true on other collections as well. In addition to the validation purpose, the replicated runs are themselves useful for research as described below.

The primary difficulty in studying topic and collection variability has been the fact that evaluated retrieval runs from a single version of a system on large numbers of topics have not been available. The 50 topics in a typical TREC experiment run on a single collection have not been sufficient. The results from the 400 topics run here will provide the first good test bed to look at topic variability of TREC style topics. This is still not enough to represent the entire universe of topics, especially given the rather stylized nature of TREC topics, but it is enough to investigate how topics group together, both in their characteristics and in their resulting search behavior.

The runs done for the merged document collection (TRECs 1–8 news articles) should be a useful resource for research in themselves, even though there are only 6 runs total. The standard pseudo-relevance feedback approach (bf) for each group was used to retrieve the top 5,000 documents for each of the 450 topics. Only partial relevance judgments are available for each topic; only the documents from the two (out of five total) volumes of the TREC disks used during the year the topic was introduced were ever judged. Research that can be done using these runs includes

- Does retrieval improve when documents from outside the target collection are used for pseudo-relevance feedback? The results from the bf run here can be restricted to a particular TREC ad hoc task, and then compared against the results of the bf run only on that task.
- Does the ranking of systems for ad hoc retrieval on the same document collection agree with the rankings of systems for Question Answering? The document set used here is exactly the document set used for TREC 9 and 10 Question Answering. For several groups there are both the ranking results of IR topics 1–405 and the ranking results of QA questions 201–1393.
- Can a valid evaluation methodology be devised for comparing runs when there is only very partial relevance judgments? This is an increasingly important topic as new, much larger test collections with much more incomplete relevance information are built.

6 Preliminary experiments in topic categorization

One of the major goals for the workshop was to understand how topics differ from each other, and how this affects system performance. An initial approach to this, unfortunately not even started until the final week of the workshop, was to automatically assign topics to categories based upon performance scores and other features. What follows are some initial experiments and some very preliminary results that are only meant to suggest further work.

For these experiments, each of the topics were "scored" based on various features, such as those below. Note that some of these scores are system-dependent and therefore there will be a topic score for each system.

1. Non-relevance-dependent features:
   - a syntactic analysis of topic text using idf
   - a comparison of the document rankings from different systems or approaches
   - a comparison of the document rankings before and after feedback within a given system
   - the Clarity measure, developed at UMass (Cronen-Townsend et al. 2002), which uses the topic and ranking obtained from a language model system to predict how easy a topic is
   - readability and clusterability were also used but not discussed in this overview
2. Relevance-dependent features:
   - the mean average precision (MAP) of the topic for a given system
   - how much pseudo-relevance feedback improved the MAP for a given system
   - how often individual added terms improved the MAP for a given system

6.1 Experimental method

For the purposes of this initial investigation, the interest was in the extremes of scores for each feature. Was the behavior of the topic different for those topics which were given a high score for the feature, as opposed to those topics given a low score? Given the feature score for each topic, the 150 topics were divided into three categories:

- Positive: The top 30 topics according to the feature score
- Negative: The bottom 30 topics according to the feature score
- Neutral: The remaining (90) topics

Some of the more natural measures, such as MAP scores, were system dependent as well as topic dependent. This could have been handled by averaging the measure across systems, but outliers and system blunders can strongly affect the average. Instead, the system dependence was handled by a voting mechanism in a two step process.

1. Step One: For each system, divide the topics into the three above categories.
   - PositiveScore: The topic has a score greater than the top X% (typically 20–30%) of the observations across all topics.
   - NegativeScore: The topic has a score less than the bottom X% (typically 20–30%) of the observations across all topics.
   - NeutralScore: The remaining topics
2. Step Two: Vote on the above categorization among the systems (normally there were 7 or 8 systems).
   - Positive: Y% ($Y > 50\%$, typically 70%) of the systems called the topic PositiveScore in Step One.
   - Negative: Y% ($Y > 50\%$, typically 70%) of the systems called the topic NegativeScore in Step One.

- Neutral: Y% (Y > 50%, typically 70%) of the systems called the topic NeutralScore in Step One.
- Mixed: None of the above (no agreement between systems on this topic)

The parameters X and Y were chosen by hand on a per feature basis to give roughly 30 topics in each of the PositiveVote and NegativeVote categories.

6.2 Categorization experimental results

There were a total of 20 categorization experiments done, with 14 investigated in some detail, including one based upon the manual topic failure analysis. All of these experiments and the data are available from the web site.

Much more work needs to be done, but several interesting results have already been discovered. The following result discussions look at the intersection of two categorizations and concentrate on correlation between the Positive (or PositiveVote) categories defined by two different feature scores.

6.2.1 Similar document rankings among all systems versus pseudo-relevance feedback MAP

The document rankings for each topic for the 8 standard runs were compared against each other by using the "anchormap" measure. This (newly defined) measure is a general, asymmetric, pairwise ranking comparison measure that emphasizes the top elements in the two rankings. Anchormap computes the similarity of a pair of system retrieval rankings in the following manner. The top X (here 30) documents of Ranking A are used as the only relevant documents to calculate a MAP score for Ranking B. If those top documents of A are near the top of B, then anchormap will be high and the rankings are considered similar. Anchormap is a general measure, but was originally a measure to specifically look at how the top X documents used for feedback in the initial run of a pseudo-relevance feedback experiment are dispersed throughout the ranking for the feedback run.

In this particular categorization of topics, anchormap was used in its general form, computed over the 56 pairs of feedback runs for the 8 systems, and averaged for each topic. The topics were then sorted by this average anchormap score, and divided into Positive, Neutral, and Negative sets, as described before. The topic categories produced by anchormap were compared against the categories produced by the top MAP scores. The Pearson correlation between the topics in the Positive groups was an extremely high 0.557, i.e., the topics for which the systems found the same top documents were indeed the topics that the systems got the best scores on. Out of the 30 topics with the most similar rankings, 19 of them were in the top 26 highest scoring topics and 0 topics were in bottom 24 scoring topics. Conversely, of the 30 topics with least similar rankings, 0 were among the top MAP scores and 9 were in the bottom 24 scores. This allows the prediction that if different systems or approaches get similar top documents, then the topic can be considered easy and standard techniques should work well.

6.2.2 Similar rankings among all systems versus pseudo-relevance feedback improvement

This categorization comparison was the same as before except instead of comparing anchormap similarities against the top scoring topics, they were compared against the topics for which pseudo-relevance feedback improved the most. Here the correlation

among Positive categories was a very high 0.327. This would indicate that if systems or approaches get similar documents, then pseudo-relevance feedback is likely to help.

An interesting investigation would be to use the anchormap similarity and like approaches to detect and correct the problem of a system missing aspects of a topic. For instance, instead of anchoring the map score in the top documents of a base run and an expansion run, anchor it in only the top documents that have some threshold similarity to a topic aspect. The absolute value of the map score of a base run counting only the documents with high similarity to a topic aspect will indicate whether the aspect is being retrieved, and the anchormap similarity, given those documents with the aspect, of the base run and expanded run will indicate whether the expansion is moving toward or away from an aspect.

6.2.3 Similar rankings between base run and feedback run versus pseudo-relevance feedback MAP

To explore the pseudo-relevance feedback improvement more, instead of comparing the similarity among the rankings of 8 different systems, compare the ranking similarity between the initial run and the pseudo-relevance feedback run of the same system. Topics were categorized by the voting procedure described previously which chooses topics for which most systems agree have the same sort of ranking similarity. The correlation among positive groups was again a very high 0.371. This would imply that the topic results are likely to be successful if the top documents of an initial search using pseudo-relevance feedback remain the near the top of the expanded search ranking.

This seems to make sense, since the top documents of the initial search were used for expansion terms and weighting in the expanded search. If different documents were retrieved then it's very possible that the new search got off-topic by over-emphasizing one aspect of the top initial documents.

6.2.4 Similar rankings between base run and feedback run versus pseudo-relevance feedback improvement

This comparison was the same as above except directly comparing whether pseudo-relevance feedback improves performance. The Positive groups had a high correlation of 0.287, again suggesting that pseudo-relevance feedback should be used when the initial top documents remain stable in their rankings.

6.2.5 Clarity versus pseudo-relevance feedback MAP

The Clarity measure was used on the CMU base run to categorize topics and this was then compared against MAP scores. The correlation among Positive groups was 0.167. Since Clarity can predict hardness of a topic, this strongly suggests that the anchormap approaches, with a much higher correlation, should also be able to predict hardness. That remains for future work.

Note that it may be fairer to compare Clarity against MAP score of baseline systems instead of pseudo-relevance feedback systems. Doing so gives a correlation of 0.177, a mild improvement but in the same ballpark.

6.2.6 Clarity versus pseudo-relevance feedback improvement

It has never been claimed that Clarity can predict pseudo-relevance feedback improvement without modification of the Clarity measure. Indeed, the RIA investigations showed a correlation among Positive categories of only 0.038. The correlation between the Positive Clarity category and the Negative improvement category was .098, substantially higher.

6.2.7 Topic rare term versus pseudo-relevance feedback MAP

If the topic contained a comparatively rare term, then it was more likely to be easy. The score for each topic here was the maximum idf of any of its original topic terms, with the topic scores then being sorted and divided into the normal Positive, Negative, and Neutral categories. The correlation between Positive categories was 0.229.

6.2.8 Topic rare term versus pseudo-relevance feedback improvement

If the topic contained a rare term, as measured by the maximum idf of all original topic terms, then it was not particularly likely that pseudo-relevance feedback will help. The correlation between Positive categories was 0.038, or roughly neutral. What was quite interesting was that the correlation between the Positive idf category and the Negative improvement category was 0.294 (like Clarity, higher than between Positive categories). For a very substantial number of topics with rare terms, pseudo-relevance feedback hurts.

6.3 Preliminary categorization conclusions

Overall, the results of the initial categorization efforts surpassed expectations. There were high correlations between a number of categories, including several described above that should be able to be transformed into a predictive process that gives insight as to what sort of retrieval approaches are likely to be successful on a particular topic.

As yet, there are no real results comparing the categories determined by the manual failure analysis with the categories described above. There were too few topics in each failure analysis category to use the same procedure. A different approach needs to be developed.

7 Summary of research results and suggested future work

There are many detailed results and suggested further work given in the previous sections; these will not be repeated here. However, there are several broad areas that should be emphasized. These are drawn from the work above, and from the half-day review discussions that each 2-week workshop session ended with.

1. Current research IR systems are failing for the same reason on individual topics. They are retrieving different individual documents, but have the same general classes of failure documents (whether non-relevant retrieved or relevant not retrieved).
2. Current system failures are dominated by presence or absence of topic aspects in the retrieved documents. The relationship between aspects, needed for factoid Question Answering, is not an important failure mode yet. This suggests that IR systems must

do a better job of simply recognizing aspects of a topic, or of recognizing that the retrieved documents do not include an aspect of the topic.

3. The data is now available for understanding why pseudo-relevance feedback improves results. The five possibilities listed in this paper's introduction can be looked at. Preliminary work here indicates that when pseudo-relevance feedback works well across systems, it works because large numbers of terms (five or more) are helpful, possibly ensuring the context of a retrieved document is correct.
4. Automatic (non-relevance-based) categorization of topics is needed as topics have to be treated differently in the retrieval process. Some categories have been introduced that need to investigated further, and others need to be added. Additionally a methodology for looking at whether those categories can be useful has been shown.
5. Categorizing topics by measuring the similarity of retrieval rankings of different approaches is both possible and informative. The anchormap similarity between rankings of several different approaches both predicts the hardness of the topic and identifies topics for which feedback should work. Topics that have retrieved sets that are comparatively stable using different approaches are more likely to be successful and more likely to improve using pseudo-relevance feedback. Other anchormap-like similarities of retrieval rankings should also be investigated. For example, comparing a full topic ranking against a ranking based on only one aspect of the topic will give a measure of the importance of that aspect to the retrieved set.
6. There is now massive data across several collections to support statistically differentiating the effect of the topic and the system upon results. Incorporating this with the automatic categorization of topics, and with the manual categorization due to failure analysis, should give insights as to how different approaches can be used for each topic.
7. At a lower level of analysis, the massive data should support finding the expansion source documents and expansion terms that most aid retrieval. The next question is determining the properties of these terms and documents that can be used to select the best candidate terms and documents in the future.

8 Conclusion and retrospective thoughts

The RIA workshop presented a very special opportunity to the IR community to *start* work on understanding how and why systems vary in performance across questions (topics). Once there is a better understanding of this, then there will be more robust IR systems, which will in turn lead to better QA systems. The initial work has been done, what remains is further analysis of the results by the entire IR community.

The workshop was both a major effort and a major success, although there was never enough time to do everything. One of the major successes was simply the act of bringing so many systems and graduate students together to work on a common task. The enthusiasm and the daily interaction of the seven groups led not only to better understanding of the various systems but to increased awareness of many different IR issues. The logistics of focusing on both a failure analysis and a common set of experiments turned out to be a good use of the 6 weeks. The early decision to create a large data set for later analysis, and the successful organization and creation of that data, turned out to be critical in the management of such a large group of work, and in providing an excellent record of what was done, allowing for future analysis.

There were two issues that created problems for the workshop, both involving the lack of time, and hopefully these can be considered "lessons learned". First, the logistics of setting up such a workshop are huge; even though the systems were set up at MITRE before the workshop, there were always things that needed changing. This involved not only system changes to run the experiments, but the building of failure analysis modules and the organization and creation of the results data. Some of this could have been done beforehand, IF the needs had been known. The second issue was the surprise as to the difficulty of the topic variation problem. It had been expected that the early experiments would lead to some hypotheses that could then be tested and would lead to more concrete conclusions. This did not happen and became part of the reason that there was so little time for analysis or categorization experiments.

A short workshop was held at SIGIR 2004 to discuss recommendations for the future. The following list is the outcome of that workshop. Note that the list is relatively unedited in that these are various ideas as opposed to an ordered list.

- What could be done differently next time
  1. monitor consistency of failure analysis, including having solid definitions of what is wanted
  2. modify the systems beforehand to autorecord data for failure analysis
  3. develop hypotheses and test them, either in new failure analysis or separately
  4. as a new set of experiments, look at the weighting issues separately
     (a) using query terms only
     (b) using query terms plus expansion terms
     (c) using pseudo-relevance feedback to check the weights on query terms, i.e., if terms are not in the top documents, then modify weights
     (d) work on a topic by topic model
- additional work with the number of documents experiments
  1. use discounting of the presence of terms in later retrieved documents (this would require system work)
  2. use of the Clarity measure or the new anchormap measure for prediction of how many documents to use
  3. tailor the initial runs for high precision
  4. look into the issue of good versus bad documents to use for feedback
  5. check into the correlation of performance of feedback with the density of relevant documents in the top 20
  6. find a way to pick the "best" cluster of documents automatically since this gives the biggest boast to performance
  7. analyze the clusters of documents to see what types of aspects appear in them
- additional work with the number of terms to add experiment
  1. manually classify whether the new terms are "key" terms or do they provide new context or aspects
  2. plot performance after each term is added
  3. classify the terms by "extraction" type (person, place, etc.)
  4. investigate whether all the systems get improvement from the same terms

- additional swapping experiments
  1. randomly swap documents
  2. do some type of fusion of terms for feedback

Acknowledgements This research was funded in part 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 includes but is not limited to the CIA, DIA, NSA, NIMA and NRO.

Appendix

- Carnegie Mellon University
  - Participants: Jamie Callan, Paul Ogilvie, Yi Zhang, Luo Si, Kevyn Collins-Thompson
  - CMU used the Lemur system, a freely available (http://www.lemurproject.org/), very flexible research statistical IR engine (Zhai and Lafferty 2001), and worked with a KL-divergence based language modeling approach with massive expansion. In this approach, queries and documents are modeled as unigram language models, or probability distributions over a vocabulary. Documents are ranked so that the documents whose probability distributions diverge the least from the query are higher in the list. The query expansion used was the divergence minimization approach (Zhai and Lafferty 2001). The divergence minimization approach estimates a language model that minimizes the divergence between a new query expansion language model and the feedback documents while using the collection language model as a controlling factor.
- City University, London (City)
  - Participant: Andy MacFarlane (working remotely)
  - City's contribution to the RIA workshop used the Robertson/Sparck Jones Probabilistic model (Robertson and Sparck Jones 1976) implemented in the OKAPI System. In this model indexed terms are weighted independently on the basis of their estimated (or probable) relevance. The BM25 weighting function was used for all experiments (Robertson et al. 1995). The Term Selection Value (Robertson 1990) was used for term selection in the pseudo relevance feedback experiments. The full range of the OKAPI BSS was used to support the experiments including passage processing and term extraction.
- Clairvoyance Corporation
  - Participants: David Evans, David Hull, Jesse Montgomery
  - Clairvoyance Corporation used two systems: CLARIT, a commercial information management toolkit written in C++ (Evans and Lefferts 1994, 1995; Milic-Frayling et al. 1998) and CLJ (CLARIT Java), a recently developed IR research toolkit built on top of CLARIT. The CLARIT system provides both indexing and retrieval functionality, as well as a wide range of other information management tools, including text classification and filtering, extraction, and summarization. CLJ consists of a set of retrieval functions (only) that run on top of a CLARIT index. CLJ was included in the experiments primarily because it was a more suitable environment to make the rapid modifications required for the RIA workshop.

In practice, it was found that the CLARIT toolkit was also flexible enough to complete the tasks required for the workshop within the time constraints. In addition, Clairvoyance contributed its Analyst Workbench, a graphical interface for text mining highly suitable for detailed failure analysis on individual topics and exploratory data analysis.

There are two special distinguishing characteristics of the CLARIT indexing process. CLARIT uses NLP for tokenization, storing individual words, noun phrases, and sub-phrases as index terms. Terms can be filtered by part-of-speech categories at indexing time; all the major content-bearing categories were used in the RIA experiments. CLARIT indexes on paragraph-sized "sub-docs" (passages) instead of full documents, typically varying in size between 8 and 20 sentences and averaging about 12 sentences in length. Document score/rank in retrieval is determined by the highest scoring sub-doc in a document. Passages of different (often smaller) sizes can be re-created on the fly for query expansion.

- Sabir Research
  - Participant: Chris Buckley
  - Sabir used Version 14.2 of SMART, a flexible IR research engine based on the vector space model as developed by Gerard Salton (Williamson et al. 1969, 1971; Buckley 1985). Documents and topics are broken into their component words and phrases, and are then statistically weighted for importance and matched. The version and parameters choices used in RIA were kept as simple as possible to allow full understanding of the effects as algorithms changed within RIA. In fact, the settings and algorithms were those used for the Cornell TREC 4 base run 9 years ago (Buckley et al. 1996). The only slightly non-standard setting used in RIA was the use of SMART statistical phrases.

    Sabir also supplied a version of SMART performing retrospective, upper bound runs. Chris Buckley supplied and modified trec_eval, a program to evaluate run results in TREC results format (http://trec.nist.gov/trec_eval/), designed most of the infrastructure to support the workshop, and served as the day-to-day leader of the workshop.

- University of Massachusetts at Amherst
  - Participant: Andres Corrada
  - The Center for Intelligent Information Retrieval at UMASS (CIIR) also used Lemur, but with different language modeling approaches. The system used for the "standard" run in the IR portion of the Workshop used the query-likelihood algorithm (Ponte and Croft 1998) using unigram scoring and Dirichlet smoothing of document probabilities.

    For the runs investigating the behaviour of feedback algorithms, UMASS used a hybrid of query-likelihood and Relevance Models (Lavrenko and Croft 2001) designed to fit the feedback strategies utilized by the other systems at the workshop. First a query-likelihood initial ranking of documents was performed, which was then used to build a Relevance Model for a query. The Relevance Model, which can be thought of as an expanded query, was then combined with the initial query to create a hybrid query. The weights used to combine the two queries were designed to never give the original query less than half the probability mass and approach the initial query as the number of feedback terms went to zero. That is, a run with one feedback term allowed, had most of its mass assigned to the initial query and gave

some small probability mass to the top feedback term.
During the QA portion of the workshop, a dynamic passaging system that calculated the query-likelihood of fixed-byte-size passages within documents was used. This is a system that is designed to identify answer passages and cannot extract "exact" answers as defined in the current TREC QA main task. For the initial passage extraction run that was used as input to the QA systems, 250-byte passages were ranked.

- University of New York at Albany
  - Participants: Tomek Strzalkowski, Sharon Small, Sean Ryan, Ting Liu, Paul Kantor (from Rutgers University)
  - Albany used a SMART/HITIQA hybrid system. The SMART that was used was an old version (Version 11.0, 1991) of the SMART retrieval engine, and was modified during the workshop in order to participate in the document swapping retrieval experiments. HITIQA is an analytic question answering system (Small et al. 2004; Strzalkowski et al. 2008) developed under the ARDA AQUAINT program. It uses the SMART system to fetch an initial set of documents from a database.
    The HITIQA QA capabilities were utilized during the last session of the workshop to test the effects of the different retrieval approaches on the effectiveness of question answering. HITIQA is an interactive open-domain question answering technology designed to accept complex analytic questions in natural language. Many of the TREC topics used in RIA experiments could be considered as synopses of reasonably complex analytic questions. The interactive features were not used during RIA.
    Typically, top 50 documents retrieved from the database are passed for answer search within HITIQA. In addition to using SMART output, HITIQA also accepted external document sets provided by all the other retrieval systems participating in RIA. HITIQA answer search includes segmenting the documents into passages, and then clustering these passages into a small number of tight topics. Representative passages from each topical cluster are subsequently mapped onto templates (called frames) which identify key topical relations and their attributes. Frame-level comparison with the input question determines the degree of fit for each frame. Frames with more than 1 attribute mismatch with the question are not considered part of the answer. In the interactive mode of HITIQA, conflict frames are negotiated with the user through a clarification dialogue, which may result in changes to the answer space. For example, frames may be added to the answer if the user decides to accept or override their matching conflicts with the question. Since the clarification dialogue was not used in RIA experiments, the initial answer space produced by HITIQA was also the final answer.
    The effectiveness of the question answering process was measured by the number of frames comprising the answer obtained from a given set of retrieved documents. The QA process was at its most effective when the size of the answer space was maximum. Separate statistics were collected for exact-match frames (zero-conflicts) and for one-conflict near miss frames, as well as for the combined set.
- University of Waterloo
  - Participants: Charlie Clarke, Gord Cormack, Tom Lynam, Egidio Terra
  - The MultiText Project, University of Waterloo, adapted passage-retrieval and term-extraction methods from their QA system to the task of pseudo-relevance-feedback

query expansion. The MultiText passage-retrieval algorithm locates "hotspots" within the corpus where many query terms cluster in close proximity (Clarke et al. 2001). After stopword elimination and stemming, the terms from the description field are used by the algorithm to locate the top ranked hotspots. Feedback terms are extracted from these hotspots, a score is computed for each extracted term, and the highest scoring feedback terms are added to the original query set. This expanded query is then executed using the MultiText implementation of OKAPI BM25 to return the top 1000 documents. Details may be found in the MultiText TREC 2003 paper, where the technique was used for their Robust track runs (Yeung et al. 2004).

Waterloo also made available two versions of Question Answering MultiText that had been used for TREC QA. They also supplied WUI, a flexible browser based user interface for examining retrieved documents, which became critical to the failure analysis part of RIA.

- NIST
  - Participants: Donna Harman, Ian Soboroff, Ellen Voorhees
  - NIST was the organizer of the workshop, and also contributed the Beadplot software used in failure analysis (http://www.nlpir.nist.gov/projects/beadplot)
- MITRE and others
  - Participants: Warren Greiff (MITRE), Paul Kantor (Rutgers University), Robert Warren, Zhenmei Gu, Luo Ming, Jeff Terrace
  - Warren Greiff and Paul Kantor from Rutgers University contributed statistical analysis of data during the workshop.
  - summer students: Robert Warren, Zhenmei Gu, Luo Ming, Jeff Terrace
  - Robert Warren was extremely active in developing the infrastructure to support the workshop. Of particular note, he and Jeff Terrace constructed an elaborate web-based system that allowed dynamic browser access to the entire database of research results, notes, data, and reports. Zhenmei Gu was drafted to be the local representative for City, making changes to the software as necessary and performing the needed experimental runs. Zhenmei and Luo Ming were responsible for replicating experimental runs of all systems on additional experimental databases.

References

Buckley, C. (1985). *Implementation of the SMART information retrieval system*. Technical Report 85-686. Computer Science Department, Cornell University, Ithaca, NY, May.

Buckley, C. (in press). Why current IR engines fail. *Information Retrieval*. doi:10.1007/s10791-009-9103-2.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART: TREC-4. In D. K. Harman (Ed.), *The Fourth Text REtrieval Conference (TREC-4)* , Gaithersburg, MD, pp. 25–48.

Clarke, C. L. A., Cormack, G. V., & Lynam, T. R. (2001). Exploiting redundancy in question answering. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* New Orleans, LA, pp. 358–365.

Clarke, C. L. A., Cormack, G. V., Lynam, T. R., Buckley, C., & Harman, D. (in press). Swapping documents and terms. *Information Retrieval*. doi:10.1007/s10791-009-9105-0.

Collins-Thompson, K., Callan, J., Terra, E., & Clarke, C. L. A. (2004). The effect of document retrieval quality on factoid question answering performance. In K. Järvelin, J. Allan, P. Bruza, & M. Sanderson (Eds.), *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* Sheffield, UK, pp. 574–575.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, & K. Järvelin (Eds.), *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* Tampere, Finland, pp. 299–306.

Evans, D. A., & Lefferts, R. G. (1994). Design and evaluation of the CLARIT-TREC-2 system. In D. K. Harman (Ed.), *The Second Text REtrieval Conference (TREC-2)* Gaithersburg, MD, pp. 137–150.

Evans, D. A., & Lefferts, R. G. (1995). CLARIT-TREC experiments. *Information Processing and Management, 31*(3), 385–395.

Gu, Z. (2004). Comparison of using passages and documents for blind relevance feedback. In K. Järvelin, J. Allan, P. Bruza, & M. Sanderson (Eds.), *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* Sheffield, UK, pp. 482–483.

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In W. B. Croft, D. J. Harper, D. H. Kraft, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* New Orleans, LA, pp. 120–127.

Milic-Frayling N., Zhai C., Tong X., Jansen, P., & Evans, D. A. (1998). Experiments in query optimization: The CLARIT system TREC-6 report. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6)* (pp 415–454). Gaithersburg, MD.

Montgomery, J., & Evans, D. A. (2004). Effect of varying number of documents in blind feedback. In K. Järvelin, J. Allan, P. Bruza, & M. Sanderson (Eds.), *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* Sheffield, UK, pp. 476–477.

Ogilvie, P., Voorhees, E., & Callan, J. (in press). On the number of terms used in automatic query expansion. *Information Retrieval*. doi:10.1007/s10791-009-9104-1.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval* Melbourne, Australia, pp. 275–281.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation, 46*, 359–364.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). OKAPI at TREC-3. In D. K. Harman (Ed.), *The Third Text REtrieval Conference (TREC-3)* Gaithersburg, MD, pp. 109–126.

Small, S., Strzalkowski, T., Liu, T., Shimizu, N., & Yamrom, B. (2004). A data driven approach to interactive question answering. In M. T. Maybury (Ed.), *New directions in questions answering* (pp. 129–140). AAAI/MIT Press.

Soboroff, I. (in press). A guide to the RIA workshop data archive. *Information Retrieval*. doi:10.1007/s10791-009-9102-3.

Strzalkowski, T., Small, S., Hardy, H., Kantor, P., Min, W., Ryan, S., et al. (2008). Question answering as dialogue with data. In T. Strzalkowski & S. Harabagiu (Eds.), *Advances in open-domain question answering* (pp. 149–188). Springer.

Voorhees, E. M., & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference (TREC-5). In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5)* Gaithersburg, MD, pp. 1–28.

Warren, R. H. (2004). A review of relevance feedback experiments at the 2003 Reliable Information Access (RIA) Workshop. In K. Järvelin, J. Allan, P. Bruza, & M. Sanderson (Eds.), *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* Sheffield, UK, pp. 570–571.

Williamson, D., Williamson, R., & Leskm, M. E. (1969). The Cornell implementation of the SMART system. In G. Salton (Ed.), *ISR Report 16 to the NSF*, pp. 1–62.

Williamson, D., Williamson, R., & Lesk, M. E. (1971). The Cornell implementation of the SMART system. In G. Salton (Ed.), *The SMART retrieval system* (pp. 12–51). Prentice-Hall.

Yeung, D. L., Clarke, C. L. A., Cormack, G. V., Lynam, T. R., & Terra, E. L. (2004). Task-specific query expansion. In E. M. Voorhees (Ed.), *The Twelfth Text REtrieval Conference (TREC-12)*, Gaithersburg, MD.

Yom-Tov, E., Fine, S., Carmel, D., & Darlow, A. (2005). Learning to estimate query difficulty. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, & N. Ziviani (Eds.), *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* Salvador, Brazil, pp. 512–519.

Zhai, C., & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Tenth International Conference on Information and Knowledge Management (CIKM 2001)* Atlanta, GA, pp. 1–2.