Inf Retrieval (2009) 12:652–665
DOI 10.1007/s10791-009-9103-2

RELIABLE INFORMATION ACCESS WORKSHOP

Why current IR engines fail

Chris Buckley

Published online: 1 August 2009

Abstract Observations from a unique investigation of failure analysis of Information Retrieval research engines held in 2003 are presented. The Reliable Information Access Workshop invited seven leading IR research groups to supply both their systems and their experts to an effort to analyze why their systems fail on some topics and whether the failures are due to system flaws, approach flaws, or the topic itself. There were surprising results from this cross-system failure analysis. One is that despite systems retrieving very different documents, the major cause of failure for any particular topic was almost always the same across all systems. Another is that relationships between aspects of a topic are not especially important for state-of-the-art systems; the systems are failing at a much more basic level where the top-retrieved documents are not reflecting some aspect at all. The investigatory framework and the lessons learned can serve as a model for needed future research in this area.

Keywords Information retrieval evaluation · Failure analysis · Failure categorization · Comparative system analysis

1 Introduction

Experimental environments such as TREC show that retrieval results vary widely according to both user topic and retrieval system. This is true for both the basic Information Retrieval (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 we could determine in advance which approach would work well, then a dual approach could strongly improve performance. Unfortunately, no one knows how to choose good approaches on a per topic basis.

The goal of the Reliable Information Access (RIA) Workshop during the summer of 2003 was to understand the contributions of both system variability factors and topic

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

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 enables system variability factors to be isolated in a way that never before has been possible.

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

As may be gathered from the list of questions, the track was much more of an investigation than a controlled experiment. There were some general expectations and hypotheses before the workshop, but no attempt was made to rigorously prove or disprove them. Instead the effort was to have experts in the field observe what is actually happening in practice with IR research systems, and suggest what can be done about it. These observations have yielded hypotheses for individual topics which may be tested in later experiments.

2 Topic failure analysis investigation

The RIA topic failure analysis investigation was an attempt to discover why current research IR systems fail and propose concrete areas of concentrated research to improve effectiveness performance. IR systems improved tremendously at the beginning of the 1990s as the TREC corpuses offered new large collections upon which to base statistical information retrieval. However, the improvement rate in the core IR retrieval algorithms has slowed down in the past few years.

One major part of the failure to improve is the perception that any individual group analyzing retrieval performance on a topic will find it very difficult to separate out the system-dependent failures from the topic-dependent failures. Some topics, or even aspects of topics, are hard for all systems. Other topics may be easier for some systems to do well on than other systems, but nobody currently understands why.

This investigation looked at IR system performance across a set of different research IR engines on a topic by topic basis. By comparing how each system succeeds or fails on any individual topic, strategies for successfully dealing with that topic can be devised. By considering enough individual topics, overall strategies for retrieval can be devised, along with identification of areas of research that need to be investigated before improvement can be expected.

Before the start of the workshop, the following observations had been made

- Good research systems all have about the same average performance as measured by Mean Average Precision (MAP), or any other standard TREC evaluation measure.
- Systems perform differently across topics.
- On any given topic, performance varies across systems.
- Systems retrieve quite different documents even if performance on a topic is about the same.

These facts suggested the following general hypotheses

1. Systems are failing on individual topics in different ways.
2. Systems are emphasizing different aspects of each topic.
3. Systems are looking at different relationships between aspects in a topic.
4. Systems will need to look at relationships between aspects to improve and these relationships can be categorized.

2.1 Failure analysis standard runs

The collection used for the standard runs were the 150 topics of TREC 6, TREC 7, and TREC 8, run on disks 4 and 5 of the TREC document distribution, not including the Congressional Record subcollection since that was used only for TREC 6. There were a total of 528,155 documents, averaging a bit over 3600 characters each.

Before the RIA workshop began, six of the seven participating groups each submitted a standard retrieval run that in some sense represented their group's approach to IR. There were no restrictions on what was done in the run as long as it was completely automatic. The resulting set of runs were widely differing:

- CMU: Lemur software implementing the KL-divergence based unigram language modeling approach Lemur (2009); Zhai and Lafferty (2001). Used blind feedback query expansion, expanding each query by several hundred terms (much more than other systems).
- UMASS: Lemur software implementing query-likelihood unigram language modeling approach with Dirichlet smoothing of document probabilities Lemur (2009); Ponte and Croft (1998). There was no query expansion or blind feedback done.
- Waterloo: MultiText system with blind feedback using passage retrieval and hotspot term-extraction, followed by an implementation of the Okapi BM25 algorithm for the final documents Clarke et al. (2001); Yeung et al. (2004).
- Sabir: SMART Version 14 vector space system with Lnu-ltu weighting. Used blind feedback and statistical phrases, implementing the base run from TREC 4 Buckley (1985); Buckley et al. (1996).
- Clairvoyance: CLARIT Java system based on full CLARIT indexing of sub-documents. Used blind feedback based on comparatively few documents adding precise (low frequency) terms Evans and Lefferts (1994, 1995); Milic-Frayling et al. (1998).
- City: Okapi probabilistic system using the Okapi BM25 algorithm. There was no query expansion or feedback, and the title field of the topics instead of the description field Robertson and Jones (May–June 1976); Robertson et al. (1995).

During week 1 of the workshop, it was discovered that the *standard* runs done before the workshop were not quite as standard as desired. Some groups had potentially indexed all fields of the documents while others had abided by the TREC restrictions of the appropriate years and not indexed the manually added fields (for example, the "SUBJECT" section of the LA TIMES). In addition, groups threw out different portions of some of the highly stylized TREC description topics ("*A relevant document must identify*…"). To make results more comparable, it was decided to standardize a set of patterns in topics that all systems would discard, and to not include the manual field of the LA TIMES documents. All systems reran their standard runs during week 1, and the resulting runs were then unchanged for the rest of the workshop.

2.2 Topic failure analysis process

Topic failure analysis was a major activity of the workshop, with 1 1/2 to 2 h per day allocated for the individual and group analysis. There were 28 people from 12 organizations participating in the six week workshop, including 15 senior IR researchers, 12 PhD students in fields relating to IR, and one entering college freshman providing Web and systems expertise. On any one day, there would be from 13 to 25 participants present at the workshop; all participants were expected to do the failure analysis. Thus the total failure analysis effort during the workshop was a bit in excess of 1000 person hours.

This was the first time this sort of group comparative failure analysis had ever been done, so substantial effort during the first week was spent developing the process of failure analysis; what we wanted to do given our goals and the available tools. The details of the process changed as the participants gained experience throughout the workshop but the overall process remained the same after the first week.

1. The topic (or pair of topics) for the day was determined, with a leader being assigned the topic, on a rotating basis among all participants.
2. Each participant was assigned one of the six standard runs to examine, either individually or as a team.
3. Each participant or team spent from 60 to 90 min investigating how their assigned system did on the assigned topic, examining how the system did absolutely, how it did compared to the other systems, and how performance could be improved for it. A template (see Fig. 1) was generally filled out to guide both the investigation and subsequent discussions.
4. All participants assigned to a topic discussed the topic for 20–30 min, in separate rooms if there were two topics. The failures of each system were discussed, along with any conclusions about the difficulty of the topic itself.
5. The topic leader summarized the results of the discussion in a short report (a template was developed for this by week 3 of the workshop). If there were two topics assigned for the day, each leader would give a short presentation on the results to the workshop as a whole.

Initially one topic per day was analyzed. People worked as teams on pairs of systems, with one person of the team being familiar with each system. The two systems were compared against each other as well as individually. People were rotated each day with the goal of each participant having a chance to work with somebody from each other system. This worked well though it could not be done fully; for example, there was nobody from City physically present. In this way, each participant learned the important details of the other systems from an expert, as well as learning how the various failure analysis tools could be used.

Given the available tools, it was just as easy to compare one system against all the other systems instead of just one other paired system, so that was done by week 2 of the workshop. By the end of week 2, people were assigned systems as individuals instead of teams, and two topics could be done in parallel. As new people arrived during the course of the workshop, they would be assigned to work as a team with an experienced participant for 2–3 days, and then were assigned to work individually.

There were several conflicting goals to consider when coming up with a template for failure analysis. The whole idea of failure analysis is to apply human intellect and experience to the question of what is happening with a given system's retrieval of documents. Having a detailed template that would be required to be filled out would defeat this

Fig. 1 Topic failure analysis template for individual system

- Behavior on top relevant documents *How many of the top documents for this system were relevant and could they be categorized and distinguished from others.*
- Behavior on top non-relevant documents *Why were the top non-relevant documents retrieved.*
- Behavior on unretrieved relevant documents *Why weren't these relevant documents retrieved within the top 1000.*
- Beadplot observations *How does the ranking (especially among the top 50 documents) of this system compare to all other systems.*
- Base Query observations *What did the system think were the important terms of the original query and were they good.*
- Expanded Query observations *If the system expanded the query (4 out of 6 systems did), what were the important terms of the expansion, and were they helpful.*
- Blunders of system *What obvious mistakes did the system make that it could have easily avoided. Examples might be bad stemming of words or bad handling of hyphenation*
- Other features of note *Anything else*
- What should system to do improve performance? *The individual's conclusion as to why the system did not retrieve well, and recommendations as to what would have made a better retrieval*
- What added information would help performance? How can system get that information?*Is there implicit information in the query, that a human would understand but the system didn't? Examples might be world knowledge (like Germany is part of Europe).*
- Assessing agreement (were there major issues? was relevance determined by "Desc"?) *The NIST assessor who originally judged relevance of documents might have a different idea of what was relevant than the text of the description indicates or than the workshop participant thinks should be relevant. It also may be unclear where and why the NIST assessor drew the line between marginally relevant and non-relevant documents*

purpose, especially given the very wide range of experience of both systems and tools among the workshop participants. For example, the Analyst Workbench (AWB) of Clairvoyance was an excellent tool for grouping documents and investigating what was happening with those groups. But it was difficult to learn well enough to be helpful, and was comparatively time consuming, and so it was only used by three or four participants. The template was designed to accommodate the reporting of results from various types of failure analysis investigations rather than prescribing the failure analysis was to be done. Figure 1 shows this template, along with explanatory text and comments (which were not part of the template or explicitly given to the participants) in italics.

The individual templates appeared to work well. They focused the less experienced participants on problems they could address while allowing the more experienced participants freedom to report what they found. A major weakness of the template was that people tended to keep track of the raw data of their observations as they went along, which was good, but did not make explicit their conclusions from their observations, which was not good. It might have been nice if people had added more to their conclusions after the group discussions of the topic, but in general they did not.

There was no template for the report by the topic leader on the topic as a whole initially. It was felt that the topics varied so much that a template would be overly constraining. However, the variability of the topic reports in the first few weeks prompted the

- *Failures common to several systems* If all or most of the systems fail on this topic in some common ways, describe them here. For example, common stemming failure, or a critical word that no one found during expansion.
- *Notable failures unique to one system* Idiosyncratic failures, notable ones only please.
- *Winning strategies* If one or two systems performed spectacularly on this topic, can you see why? For example, an unusual expansion term might enable one system to find a lot more relevant documents.
- *Classes of missed and false alarm documents* Are there identifiable clusters among missed documents (relevant, not retrieved)? What about false alarms (not relevant, retrieved)? Do most of the misses and/or false alarms fit into these classes, or are there a lot of special cases?
- *Notes on topic statement and relevance judgments* Everyone's got a gripe, but sometimes systematic idiosyncrasies in judging or critical words missing from a description are good to know.
- *Testable Hypotheses* The above observations may lead to a few hypotheses one how this topic's performance can be stabilized. List them here. Hypotheses should be testable, that is, it should be possible in principle to implement the solution in one or more of the systems and re-run the experiment.

Fig. 2 Topic failure analysis template for overall topic

development of a second template (see Fig. 2) by the end of week 3. For this template, the explanatory text was included in the template.

Again, there was the conflict between demanding enough detail in the template to provide good information common across topics and demanding so much detail that the template questions no longer fit the important parts of the topic analysis. Like the individual template, we opted for mostly general questions; though having a top-level very general template item would have been helpful.

There was debate in the workshop on several occasions as to whether the topics should be categorized as to major failure causes as they were being analyzed. It was decided that for future investigations, coming up with a set of failure categories would be useful and needed, but during this workshop we did not know enough about the types of expected failures until the end. The danger of prematurely defining categories is that people may be tempted to force a topic into a pre-defined category that it may not really fit. One of the results of this workshop is a list of failure categories.

2.3 Failure analysis topic choice

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. We had hoped we could do 50 topics and we actually finished 45 topics. Ideally, we should randomly choose topics that fit our desired criteria. In practice, we did not know enough about what criteria we wanted. Our first attempt failed badly. We ranked topics by variability of the ranked document results; topic 368 was the most variable. While retrieved documents and scores varied widely, all systems did fairly well on it: MAP varied from .38 to .76. However, during the failure analysis for this topic participants ended up spending their time judging how well each system matched the TREC assessor's line between marginally relevant documents and barely non-relevant documents; a task impossible to do well, and not

intellectually interesting. It was then decided to focus on topics where the systems in general scored below the overall MAP average. The overall criteria used were

- Average MAP of systems at or below the overall MAP average of about .21.
- Large variance between system scores.
- Variance in general not due to the City Okapi run (different vocabulary since it used topic title instead of description field).

In addition, we analyzed several topics (about 4 each) that

- Performed differently than others in some other RIA experiment.
- Had the basic form of a TREC Question Answering question.

Note that the analyzed topics cannot be said to be random selection of all the topics. We did not analyze most of the easy topics (we analyzed only two out of the top 56 topics, ranked by average MAP), since they had high MAP for most systems. We also did not analyze many of the topics that all systems did extremely poorly on (only four out of the bottom 23 average MAP), since there was generally very little variability between system scores on those topics. Since each topic took at least 11 person-hours to analyze, we concentrated on those topics for which there was evidence of some system-dependent effect and some evidence of system failure, and analyzed over half of those topics.

2.4 Tools for topic failure analysis

There were a wide variety of tools used in failure analysis. They include

- WUI: The Waterloo User Interface was the major tool used for examining retrieved documents of a system. It offered a GUI controlled front-end that constructed database queries to be sent to a Waterloo back-end database engine. For example, four mouse clicks and typing in the topic id 301 might start showing all non-relevant documents retrieved by Sabir's standard run with rank less than 40 for that topic. Each individual document in the initial display of documents would give whether the document was relevant, what rank the document was retrieved at for each of the six standard runs, the system's best guess at the most relevant excerpt from the document, and several user-defined buttons and text boxes with which a user could make notes. Clicking on the excerpt would give the full text of the document, with any user designated words or patterns highlighted.
  Most participants used WUI for about 75% of the time they spent on individual failure analysis.
- Beadplot: Freely available from NIST. Beadplot represents each retrieved document of a target retrieval ranking by a color bead on a rank axis. Each other retrieval ranking can be visually compared to the target ranking by seeing whether the color patterns of the two rankings are similar. Beadplot was used by some participants on all topics, but it was not as useful as anticipated. For most topics, the ordering of retrieved documents was just not similar enough between systems, with the occasional exception of the two Lemur systems (CMU and UMASS). The workshop version of Beadplot was adapted by Sean Ryan of Albany to automatically work with the six standard runs.
- AWB: The Analyst Workbench from Clairvoyance is a very nice package allowing grouping of documents by arbitrary pipe-lined filters, with easy analysis or clustering of the groups. For example, it is easy to tell that a certain term occurs in 80% of the top relevant retrieved documents, but only 5% of the relevant documents that were not

retrieved. Unfortunately, despite tutorials and substantial efforts by Jesse Montgomery of Clairvoyance at setting up installations of AWB working in the failure analysis environment, only a couple of non-Clairvoyance participants were able to use it. The learning curve was too high for the very limited amount of workshop time available. The AWB is part of commercial products available through Clairvoyance.

- smart_std: The SMART system itself offers the ability to look at documents and retrieval results of the six standard runs. Most of the capabilities were much more nicely available through WUI. One feature that was occasionally used by participants was the ability to get a table giving the ranks at which each relevant document for a topic was retrieved for each of the six runs, sorted by collection. This enabled easy determination of whether systems had a collection dependent bias to their retrieval for a topic (such biases existed for several topics and systems.)
- smart_retro: An adaptation of the SMART DFO approach performed on the retrospective collection where relevance information is known for all documents. For a given topic, smart_retro attempts to construct the optimal simple vector query given the relevant documents. Over all topics, smart_retro averages a MAP score of .83. Participants were told not to look at the smart_retro optimized query until after they performed their individual failure analysis, since it could very easily bias the analysis to simply presence or absence of key terms. Three noteworthy points of the optimized queries:

  1. Often the manual failure analysis would indicate that poor performance was due to a certain term being missed or weighted lightly. The optimized query served to verify or reject that hypothesis. It was surprising how often the hypothesis was rejected.
  2. An indication of the overall difficulty of the topic. If the optimized topic had a low MAP score (for example, below .5), that was evidence that a bag of words approach was not ideal for the topic and that relationships between terms would be needed.
  3. An indication of importance of sub-collections. It was surprising how highly smart retro weighted termsthat only indicated the source of the document. For example, all Financial Times documents containedthe term "FT" in the headline. That term was very often among the top 5 terms in the optimized query.

- Web interface: The web interface to the failure analysis topics was an enormously helpful tool by the end of the workshop. Rob Warren and Jeff Terrace built the web interface to the entire workshop from scratch. By the end of the workshop, the failure analysis page for a topic would bring up the topic itself, all evaluation scores of the six standard runs for the topic, the full text of the topic including narrative, lists of words occurring in the description field suitable for cutting and pasting into the WUI highlighting tool, similar lists for the narrative field, standard measures of textual difficulty of the topic text, any categorization of the topic that had been performed in the workshop, count of relevant documents that occurred in each sub-collection, and links to any failure analysis that had already been done on the topic. Having all that information collected in one place made the intellectual task of failure analysis that much easier. The web interface is freely available; attempts were made to make it portable, but there is no experience yet at re-installing it elsewhere.
- Individual systems: Many participants used their own system as a tool for looking at results and trying alternative query formulations. There was general agreement that

their own systems could have been much more helpful at failure analysis if there had been enough time and expertise to adapt the systems before the workshop started.

3 Topic failure analysis categorization

As discussed previously, there was not any attempt to construct explicit categories of topic failures as the topics were being analyzed. However, during the final week of the workshop, after topic failure analysis had finished, Chris Buckley constructed categories which in his opinion represented the different sorts of failures seen as important during the summer.

1. General success—present systems worked well
   1 topic, primary category
   Sample Topic *368 Identify documents that discuss in vitro fertilization*. This was the only topic examined in failure analysis on which systems worked well and is included just to be complete. The blind feedback expansion systems had noticeable improved performance, but the participants had no other suggestions or observations.
2. General technical failure (stemming, tokenization)
   2 topics, primary category
   Sample Topic *353 Identify systematic explorations and scientific investigations of Antarctica, current or planned*. Almost all systems did not stem "Antarctica" and "Antarctic" to the same stem Sample Topic *378 Identify documents that discuss opposition to the introduction of the euro, the European currency*. All systems preferred matches to all the various hyphenated forms of "euro-something" (like "euro-centric") instead of preferring the currency "euro".
3. All systems emphasize one aspect; missing another required term
   7 topics
   Sample Topic *422: What incidents have there been of stolen or forged art?* All systems would do much better if the term "art" was emphasized.
4. All systems emphasize one aspect; missing another aspect
   14 topics
   Sample Topic *355 Identify documents discussing the development and application of spaceborne ocean remote sensing*. All systems needed to emphasize the aspect of "ocean". Much like Category 3, except some collection of expansion terms related to "ocean" would likely be needed to improve performance.
5. Some systems emphasize one aspect; some another; need both
   5 topics
   Sample Topic *363 What disasters have occurred in tunnels used for transportation?* Some systems emphasized disasters and others emphasized tunnels. Both were needed.
6. All systems emphasize one irrelevant aspect; missing point of topic
   2 topics
   Sample Topic *347 The spotted owl episode in America highlighted US efforts to prevent the extinction of wildlife species. What is not well-known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?* All systems emphasized spotted owl and US efforts.
7. Need outside expansion of "general" term (*Europe* for example)
   4 topics, primary category

4 topics, secondary category (topic was assigned a different primary category)
Sample Topic *398 Identify documents that discuss the European Conventional Arms Cut as it relates to the dismantling of Europe's arsenal*.
Sample Topic *448 Identify instances in which weather was a main or contributing factor in the loss of a ship at sea*. Systems needed to expand the concept of weather.

8. Need QA query analysis and relationships
   2 topics, primary category
   1 topic, secondary category (topic was assigned a different primary category)
   Sample Topic *414 How much sugar does Cuba export and which countries import it?* Need notions of quantity and relationships between query terms.
9. Systems missed some difficult aspect that would need human help
   7 topics, primary category
   1 topic, secondary category (topic was assigned a different primary category)
   Sample Topic *413 What are new methods of producing steel?* "New methods" was very difficult for systems
   Sample Topic *393 Identify documents that discuss mercy killings*. mercy killings was a difficult concept that was often expressed only in other language (for example, system needs to match "right-to-die").
10. Need proximity relationship between two aspects
    1 topic, primary category
    5 topics, secondary category (topic was assigned a different primary category)
    Sample Topic *438 What countries are experiencing an increase in tourism?* As a primary focus, need aspect of increase. As a secondary focus, want aspects of increase and tourism to be close together.

The categories above are roughly sorted in order of increasing Natural Language Understanding (NLU) being needed to improve performance once it is understood a topic belongs in that category. Topics were placed in the least restrictive category (towards the top of the list) that would give substantial improvement if the problem could be addressed.

The assignment of topics to these categories does not address the problem of how difficult it is to automatically discover what category the topic belongs to. Thus it may be extremely difficult and require full NLU and world knowledge to distinguish those topics in, for example, category 9 from those in category 4. But if the system can distinguish those categories, possibly by using more information than is available in just the topic, it should be able to attack the missing aspect problem of category 4 in a straightforward fashion, while the missing aspect of category 9 topics will still be very difficult to attack.

It should be noted that these categories and assignments of topics to categories are the results of one person's efforts. If 100 experts were asked to do the same exercise and given the same information (the filled in topic templates and the resulting discussions), there would undoubtedly be 100 different definitions of categories, and assignment of topics to categories. Nonetheless, some conclusions can be reached from this particular categorization.

The first conclusion is that the root cause of poor performance on any one topic is likely to be the same for all systems. Except for the six topics of categories 1 and 5, all systems failed for the same reasons (modulo the rare individual system blunders). Beadplot and other tools show that the systems were retrieving different documents from each other in general, but all systems were missing the same aspect in the top documents.

The other major conclusion to be reached from these category assignments is that if a system can realize the problem associated with a given topic, then for well over half the

topics studied (at least categories 1 through 5), 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.

4 Lessons learned about failure analysis

This was the first large-scale failure analysis to be done on multiple automatic retrieval systems. As such, the lessons learned *about* doing failure analysis may be as valuable in the future as the lessons learned *from* doing the analysis.

4.1 Examining documents is central

Of course, looking at documents (or parts of documents) is critical to failure analysis, but the extent to which our time was dominated by documents was still surprising. That may be due in part to the relative maturity of the available tools to the task at hand. We had good tools to examine documents, but had less experience with effective use of our other analysis tools.

The Waterloo User Interface (WUI) was essential to most analyses. Since it was built on top of a database system, it allowed us to easily access exactly the documents we wanted, in both snippet and full text form, with desired features highlighted.

Our other tools occasionally gave valuable information, but were less generally useful than we expected. The Beadplot tool for visually observing document ranks works well on smaller collections, but there are too many documents too similar to each other to really gain much insight on larger collections. The Analyst Workbench (AWB) is a very general, flexible tool, but suffered from its flexibility in the workshop: we needed to first know a particular problem of a topic/ranking to investigate. It might have been more useful as a confirmation tool to support/deny the hypotheses we reached. The Web interface was a great time saver, collecting all the known information about a topic in one place.

4.2 Templates are useful

Having templates for both the individual system analysis and the group discussions were beneficial. Both were developed during the workshop to reflect what we were actually doing.

We held a mid-term review of the failure analysis process. David Evans and Paul Kantor observed several sessions of the group discussions, both with and without templates. Their general conclusion was that the template discussions more completely summarized performance on a topic.

The templates we used should be expanded in the future. In particular, a categorization of the conclusions should be added to allow group input into the categorization. Our conclusions on individual topics were too "free-form" to be able to form general conclusions across topics. However, be aware that categories will change, depending on system and state-of-the-art. At the beginning of the workshop, we were expecting a different categorization than we ended up with—one more focused on types of relationships between topic terms. In the end, term relationships were not the cause of many failures.

4.3 Comparative failure analysis was beneficial

When doing an individual system failure analysis, it was quite helpful to be able to compare that system to other systems. Indeed, at the start of the workshop we were comparing only pairs of systems, but we quickly expanded that to compare a system against all other systems. There was a major benefit in being able to focus on documents on which other systems either ranked consistently higher or lower.

Even though systems in general failed for the same reasons as other systems, they often failed to a greater or lesser extent. A system might not emphasize a secondary aspect of a topic enough, but still emphasize it more than other systems. That was easily detectable when looking at documents and their comparative rankings across systems.

4.4 Verification of failure analysis hypotheses is desirable

We did very little testing of our various hypotheses for individual system failures; we had planned to do much more. Unfortunately, the individual retrieval system modifications that would enable systematic testing of some of the hypotheses were not finished until near the end of the workshop. For other experiments in the workshop, we added the ability to import weights and/or terms from outside each system. We had hoped to be able to piggyback hypothesis testing on these changes (e.g., increase the weight on terms that failure analysis said were underweighted, or manually expand a topic with synonyms), but were unable to. We both ran out of time, and ran into the obstacle that the systems that had been modified for experiments throughout the workshop were all running different variations than they had been running when the failure analysis runs were frozen. Ideally in the future, system modifications to enable failure analysis hypothesis testing should be done before the failure analysis starts. Not only would that enhance confidence in the results, but those modifications would be a valuable failure analysis tool in their own right.

5 Conclusion

Most of the incoming hypotheses stated in Sect. 2 turned out not to be true. Restating the hypotheses:

1. *Systems are failing on individual topics in different ways*. Despite the fact that systems retrieve different documents, all systems tended to fail in the same way. In 39 out of the 45 topics examined, the failures were put in an *almost-all* systems category. Every system had two or three "blunders"; e.g., the SMART system had a couple of topics with over-aggressive stemming, the Clarit system expanded with inappropriate narrow terms a couple of times. Those are the expected system tradeoffs; what was surprising is how few topics had these system failures as opposed to topic failures.
   Note that this high agreement between systems as to failure cause was measured on a subset of the topics that were specifically chosen to have high variation between systems. Those were the topics that we expected to have different reasons for failure. We would expect even greater agreement on a random set of poorly performing topics.
2. *Systems are emphasizing different aspects of each topic*. Again, while this happened (5 out of 45 topics), it did not happen to the extent we expected. Systems in general agreed on which aspects to emphasize.

3. *Systems are looking at different relationships between aspects in a topic*. We saw very little evidence of this. However, this may have been due to the particular systems involved.
4. *Systems will need to look at relationships between aspects to improve and these relationships can be categorized*. We still expect this to be true eventually, but at the moment the problem is more basic: systems need to recognize the importance of the aspects of the topic. The fact there is a relationship in the topic may be helpful in emphasizing aspects, but there is little evidence that matching relationships between topic and document will be immediately helpful (categories 8 and 10).

One interesting retrieval approach suggested by these results is to try to do retrieval ranking via automatic failure analysis. The presence or absence of terms in the top documents of a retrieved set can be automatically determined. If an aspect is not being represented in those top documents, then it can be given an increased weight or can be expanded by synonyms. The query can be reformulated and re-run until the retrieved top documents reflect the important aspects of the topic. As collection size continues to increase, general collection idf becomes a very blunt tool for weighting terms. This reformulation of the query allows for much more precise relative weighting of topic aspects.

Overall, the type of semantic relationship between aspects is not yet the primary cause of failure. There should be a lot of improvement possible without understanding relationships, though in the long-term, relationships will be necessary. Finally, understanding the semantics of the topic well enough to just identify the important aspects would seem to be crucial for many topics.

Acknowledgements This research was funded in part by the Advanced Research and Development Activity in Information Technology (ARDA), a US 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.

References

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

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)* (pp. 25–48). October 1996. NIST Special Publication 500-236.

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* (pp. 358–365). September 2001.

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)* (pp. 137–150). NIST Special Publication 500-215.

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

Lemur (2009). The lemur toolkit for language modeling and information retrieval. Available at http://www.lemurproject.org/

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). NIST Special Publication 500-240.

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 21st annual*

international ACM SIGIR conference on research and development in information retrieval (pp. 275–281). August 1998.

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 *Overview of the third text retrieval conference (TREC-3) [Proceedings of TREC-3]* (pp. 109–126). NIST Special Publication 500-225.

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)*.

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)*.