Inf Retrieval (2009) 12:642–651
DOI 10.1007/s10791-009-9102-3

RELIABLE INFORMATION ACCESS WORKSHOP

A guide to the RIA workshop data archive

Ian Soboroff

Received: 18 October 2006 / Accepted: 7 January 2009 / Published online: 18 July 2009

Abstract During the course of the Reliable Information Access (RIA) workshop, a data archive was created to hold the outputs of the many experiments being done. This archive was designed to serve both as an organizational structure to support the researchers at the workshop itself and as a public archive of experimental retrieval results. This article describes the structure of the data in the archive and the ways in which the data may be accessed.

Keywords Experiment · Archive

1 Introduction

Creating and maintaining an organizational structure for experimental data and results is critical for any research effort, especially when multiple researchers are involved. It was all the more so for the Reliable Information Access (RIA) workshop, which brought together seven research teams to work for six weeks on a large, common set of information retrieval problems. Over the course of the workshop, people came and went, experiments were run and re-run, and procedures were devised, thrown out, and rewritten again. Without a strong structure, all would have been lost, or at least badly misplaced.

It would be wrong to imply that the RIA archive was designed completely before the start of the workshop. Chris Buckley (Sabir Research) did devise and implement the initial structure ahead of time to accommodate experiments which were meant to be run in the first days of the workshop in order to test all the systems involved. But the actual operational archive evolved over the course of the workshop through the tireless (or at any rate sleepless) efforts of Robert Warren and Jeff Terrace of the University of Waterloo. On top of the basic information structure, they developed a web site and scripts that permitted everyone to easily place their work into the archive.

I. Soboroff (✉)
National Institute of Standards and Technology, Gaithersburg, MD, USA
e-mail: ian.soboroff@nist.gov

The archive is currently hosted at NIST, and can be found on the web at http://ir.nist.gov/ria. This article describes the structure of the information present in the archive, and the ways in which the RIA data, experimental results, and analyses may be accessed. In this paper we often refer to the web site, and it may be helpful to the reader to explore the site while reading the paper. In particular, when we refer to "navigational links", these are listed in a sidebar on the left-hand side of the web page.

2 Information types

Although the experimental details of the RIA workshop are explained elsewhere (Harman and Buckley 2004; Buckley 2004), it is useful to summarize the key points here. The RIA workshop consisted of two main types of activities. First, a series of large-scale retrieval experiments were designed and conducted using a common test collection and similar parameter settings across all participating retrieval systems. Second, 45 search topics from the test collection were examined in detail to understand how different systems fail to perform effectively for those topics.[1]

These activities imply several classes of information:

Topic: A *topic* is an articulation of a user's information need. It contains several fields which provide varying levels of description of the need. The topics used at RIA were developed within the ad hoc track of the Text REtrieval Conferences (TREC) 6, 7, and 8 (Voorhees and Harman 2005). They include a *title* of two to four key words, a sentence-length *description*, and a paragraph-length *narrative*. Retrieval experiments (see below) use a large set of topics, and effectiveness measures are available for individual topics and averaged over the full set. Failure analyses (again, see below) consider an individual topic.

System: Seven information retrieval *systems* were used over the course of the workshop. A system refers to the piece of software itself, possibly including standard, experiment-invariant parameter settings. Because systems represent fundamental approaches to search, the system is the usual comparative axis in an experiment or a failure analysis.

Run: Following the TREC terminology, a *run* is a batch-mode search of a document collection using a particular system in response to each of a set of topics. A run is associated with particular parameter settings of the system involved as well as its output. The output of a run is the top *n* (usually 1000) documents retrieved for each topic, with a score and rank for each. Runs are conducted in the context of experiments.

Experiment: Experiments have their customary definition within the context of laboratory-style information retrieval research. RIA experiments always have the system as a dependent variable, in addition to others which are particular to the experiment. Each experiment involves a (possibly large) number of runs, each with a particular system set to the corresponding parameters dictated by the experiment. Each experiment has a short name associated with it. Two important experiments are **standard**, a baseline run from all systems which is used as the starting point for failure analysis, and **bf_base**, which consists of baseline blind feedback runs for comparison in experiments which vary feedback parameters.

Analysis: There are two principal types of analyses in the RIA archive. The first is an analysis of an experiment. The second is a failure analysis, which is related to a topic.

[1] See elsewhere in this issue for details on how these topics were chosen.

The RIA data archive is designed to allow access to these types of information from all possible starting points. It is possible to start from the analysis of an experiment, drill down to a particular system configuration, step back to look at average effectiveness for that system or several systems, or to look at particular topics within that experiment. A similar progression is possible within a topic failure analysis. By selecting particular runs, systems, and topics, it is even possible to analyze the output of a "virtual" experiment using the data already present in the archive.

The following sections describe how this information is structured in the RIA archive, and how to access it through the site.

3 Topics

Nearly all of the experiments at the RIA workshop used a single test collection: TREC topics 301–450 and the documents from TREC CDs 4 and 5.[2] These topics were developed in the adhoc track of TRECs 6, 7, and 8. This collection is referred to in the archive as **v45.301-450.d**. Another collection, **v24.251-300.d** is present in the archive but is not used in any experiments.

Selecting the "Topics" link in the RIA site navigation bar (top left in Fig. 1) and then the **v45.301-450.d** collections allows access to a descriptive web page for each topic. As an example, the page for topic 419 is shown in Fig. 1. The topic page shows the title, description, and narrative topic fields on the left, extracted keywords from those sections on the right, and several other statistical measures, some of which are referenced in the experiments and failure analyses. These measures include average word frequency, the rarest word, the Flesch-Kincaid Reading Ease Score, the number of hits at that time from Google (using the title as the query, both as a simple query and as a phrase), and the number of known relevant documents across the document subcollections. Where appropriate, these measures are given for the title and description sections and for the entire topic. If a failure analysis exists for this topic, the topic page links directly to it.

Some experiments, notably the **topic_analysis** experiment, computed several topic-specific measures and some of these are included on the topic pages as well.

4 Systems

While the systems are a focal parameter of each experiment, there is no system view in the RIA data archive. The best way to see the system baseline configurations is to look at the **standard** experiment page.

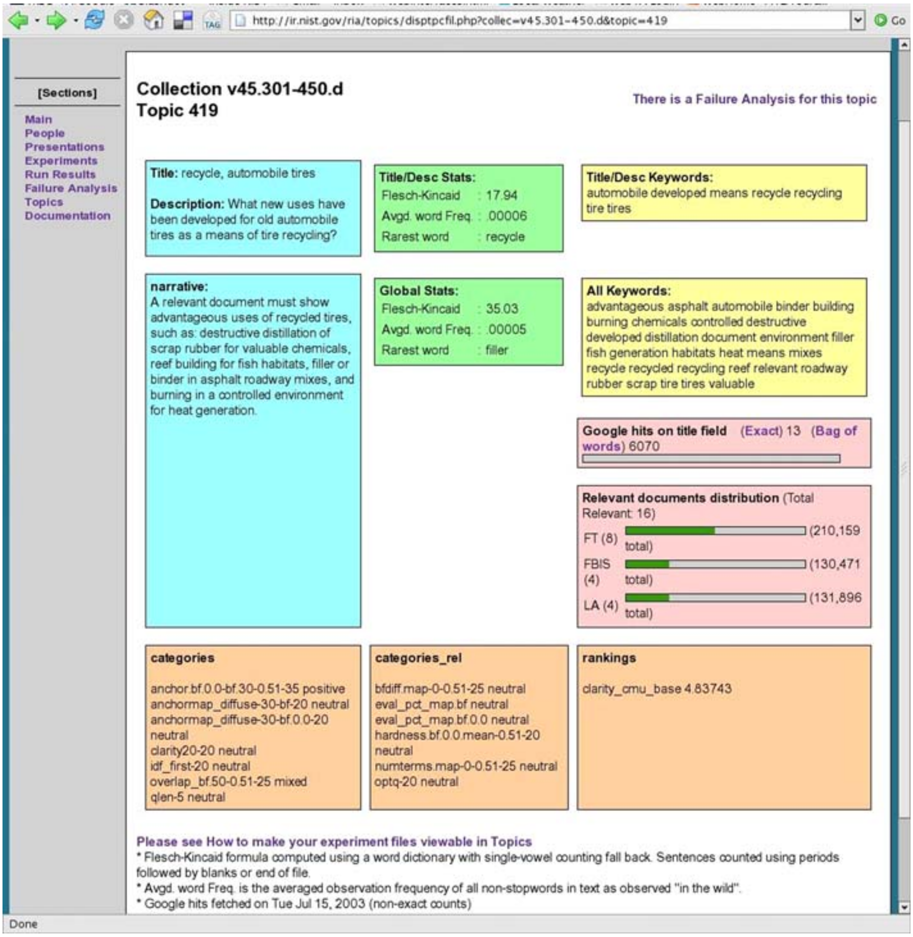
The seven systems used at RIA were:

Albany: is SMART 11.0, a commonly-available[3] vector-space system which is used as the retrieval engine in the HITIQA question-answering system.
City: is OKAPI, a well-known probabilistic retrieval system.[4]

[2] These CDs include the Financial Times (FT), Federal Register (FR), Foreign Broadcast Information Service (FBIS), and LA Times (LA) subcollections. The *Congressional Record* subcollection on those CDs is not used with these topics.

[3] ftp://ftp.cs.cornell.edu/pub/smart/.

[4] As of this writing OKAPI has been released as open-source software under the BSD license; see http://www.soi.city.ac.uk/~andym/OKAPI-PACK/.



Collection v45.301-450.d
Topic 419

There is a Failure Analysis for this topic

Title: recycle, automobile tires

Description: What new uses have been developed for old automobile tires as a means of tire recycling?

Title/Desc Stats:

Flesch-Kincaid	: 17.94
Avgd. word Freq.	: .00006
Rarest word	: recycle

Title/Desc Keywords:
automobile developed means recycle recycling tire tires

narrative:
A relevant document must show advantageous uses of recycled tires, such as: destructive distillation of scrap rubber for valuable chemicals, reef building for fish habitats, filler or binder in asphalt roadway mixes, and burning in a controlled environment for heat generation.

Global Stats:

Flesch-Kincaid	: 35.03
Avgd. word Freq.	: .00005
Rarest word	: filler

All Keywords:
advantageous asphalt automobile binder building burning chemicals controlled destructive developed distillation document environment filler fish generation habitats heat means mixes recycle recycled recycling reef relevant roadway rubber scrap tire tires valuable

Google hits on title field (Exact) 13 (Bag of words) 6070

Relevant documents distribution (Total Relevant: 16)
FT (8) (210,159 total)
FBIS (4) (130,471 total)
LA (4) (131,896 total)

categories

anchor.bf.0.0-bf.30-0.51-35 positive
anchormap_diffuse-30-bf-20 neutral
anchormap_diffuse-30-bf.0.0-20 neutral
clarity20-20 neutral
idf_first-20 neutral
overlap_bf.50-0.51-25 mixed
qlen-5 neutral

categories_rel

bfdiff.map-0-0.51-25 neutral
eval_pct_map.bf neutral
eval_pct_map.bf.0.0 neutral
hardness.bf.0.0.mean-0.51-20 neutral
numterms.map-0-0.51-25 neutral
optq-20 neutral

rankings

clarity_cmu_base 4.83743

Please see How to make your experiment files viewable in Topics
* Flesch-Kincaid formula computed using a word dictionary with single-vowel counting fall back. Sentences counted using periods followed by blanks or end of file.
* Avgd. word Freq. is the averaged observation frequency of all non-stopwords in text as observed "in the wild".
* Google hits fetched on Tue Jul 15, 2003 (non-exact counts)

Fig. 1 The RIA archive page for topic 419, “recycle, automobile tires”

Clairvoyance: is sometimes referred to as “CLARIT” or “Full CLARIT” in the archive. This is the CLARIT retrieval system within the Analyst’s Workbench product from Clairvoyance Corp.

Carnegie Mellon: is usually referred to as “CMU” in the archive. This system is their version of the Lemur, a modern retrieval system that uses language modeling.[5]

Sabir: is SMART version 14. It is substantially similar to SMART 11 but has many improvements in weighting schemes and efficiency.

UMass: is the version of Lemur from the University of Massachusetts, Amherst. The two versions of Lemur have minor differences in language-modeling algorithms for feedback, which were just being developed at that time.

[5] http://www.lemurproject.org/.

Waterloo: is the MultiText system,[6] which is designed to retrieve arbitrary passages rather than only whole documents.

Parameter settings for each system are typically given within each experiment.

5 Experiments

The experiments section of the archive, linked in the navigation bar on the far left, is the simplest way to approach the massive number of runs and evaluation results generated at the RIA workshop. The major experiments are:

standard: A "representative" run from each system, done at the beginning of the workshop. These runs were used for failure analysis.
bf_base: Baseline blind feedback runs from each system.
bf_numdocs: Varying the number of top documents used for feedback. A sub-experiment, **bf_numdocs_relonly** looked at using the top relevant documents only.
bf_numterms: Varying the number of terms added from feedback documents. The experiment **bf_pass_numterms** looked at adding terms from passages for systems that can return passages rather than documents.
bf_swap_doc: Feedback using documents retrieved by other systems. This was a complicated experiment and has several sub-experiments:

bf_swap_doc_cluster: Using documents from CLARIT's clustering algorithm.
bf_swap_doc_fuse: Using documents fused from several systems.
bf_swap_doc_hitiqa: Using documents from HITIQA.
bf_swap_doc_term: Using both documents and expansion terms from another system.

We focus our discussion on these experiments because they link to descriptions of systems and runs as well as providing experimental analysis. These experiments, which include more than two thousand runs, form the bulk of the RIA data archive. Additionally, there were several experiments that did not generate runs, but instead were analyses of data from other experiments, the topics themselves, and the failure analyses.

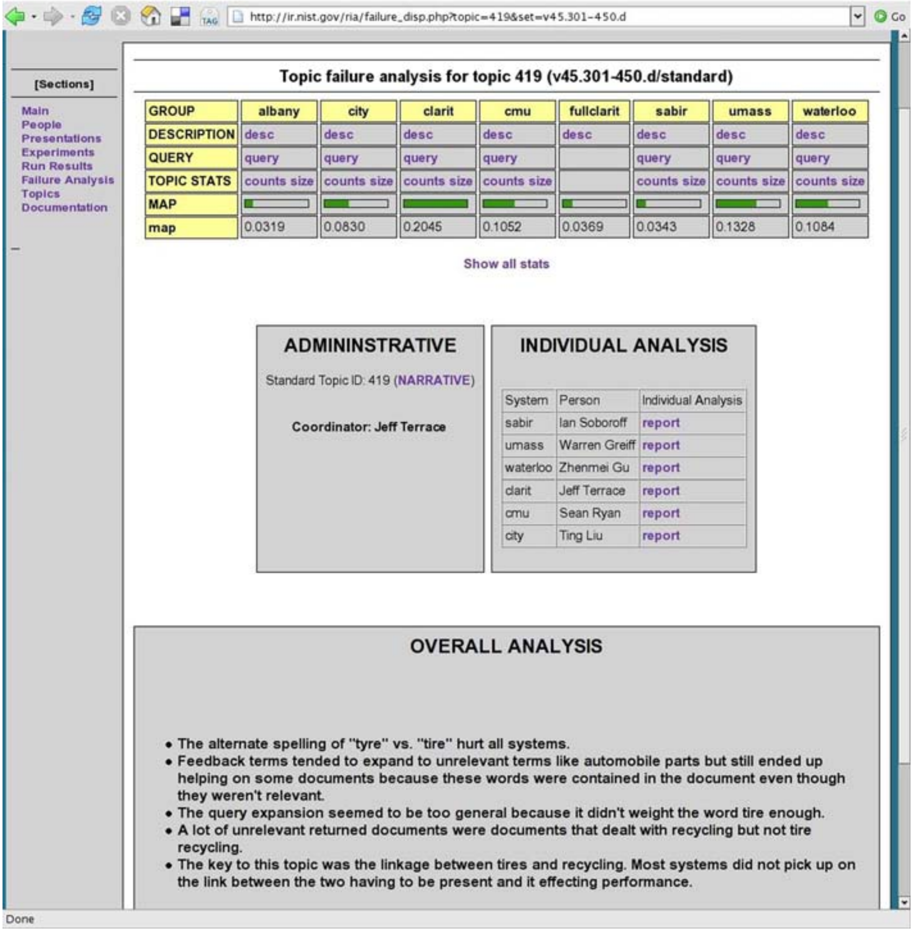
Each experiment is summarized on its own web page linked from the experiments page. For example, the experiment page for **bf_numterms** gives an overview of the experiment, its goals and hypothesis, which is that varying the number of terms added in feedback has a measurable effect on feedback effectiveness. Below this are links to run reports by system.

Each system page links to the runs for each experimental parameter setting. For example, following the link to CMU, then "bf.20.1" gives the details on the CMU run using 20 feedback documents and a single feedback term. It includes a script to recreate the run, a description of settings, and links to the generated query, term weights, retrieval output, and evaluation results.

6 Failure analyses

Failure analyses were a daily activity at the RIA workshop, sometimes extending through an entire morning. A RIA failure analysis examines a single topic, using the output of each

[6] http://multitext.uwaterloo.ca/.



Topic failure analysis for topic 419 (v45.301-450.d/standard)

GROUP	albany	city	clarit	cmu	fullclarit	sabir	umass	waterloo
DESCRIPTION	desc	desc	desc	desc	desc	desc	desc	desc
QUERY	query	query	query	query		query	query	query
TOPIC STATS	counts size	counts size	counts size	counts size		counts size	counts size	counts size
MAP								
map	0.0319	0.0830	0.2045	0.1052	0.0369	0.0343	0.1328	0.1084

Show all stats

ADMININSTRATIVE

Standard Topic ID: 419 (NARRATIVE)

Coordinator: Jeff Terrace

INDIVIDUAL ANALYSIS

System	Person	Individual Analysis
sabir	Ian Soboroff	report
umass	Warren Greiff	report
waterloo	Zhenmei Gu	report
clarit	Jeff Terrace	report
cmu	Sean Ryan	report
city	Ting Liu	report

OVERALL ANALYSIS

- The alternate spelling of "tyre" vs. "tire" hurt all systems.
- Feedback terms tended to expand to unrelevant terms like automobile parts but still ended up helping on some documents because these words were contained in the document even though they weren't relevant.
- The query expansion seemed to be too general because it didn't weight the word tire enough.
- A lot of unrelevant returned documents were documents that dealt with recycling but not tire recycling.
- The key to this topic was the linkage between tires and recycling. Most systems did not pick up on the link between the two having to be present and it effecting performance.

Fig. 2 The failure analysis page for topic 419

run from the **standard** experiment.[7] Figure 2 illustrates the analysis report for topic 419, which can be reached either via its topic page or the failure analysis navigational link.

The first part of the report links to technical information about each run. The "desc" link points to parameter settings for that system. The "query" link leads to the parsed and (for some systems) weighted query terms. "Counts" pops up a display showing relevant, relevant retrieved, and total retrieved documents from each of the four v45.301-450 subcollections. "Size" pops up a display showing average document lengths for these documents across the subcollections. The average precision for each run is also displayed, and optionally the full evaluation output produced by the trec_eval utility[8] can be shown.

[7] The other experiments were not used in failure analysis primarily because the analyses were started before the other experiments were conducted! The analyses were sometimes informed by the other experiments, and did themselves serve to drive the design of some of the experiments.

[8] http://trec.nist.gov/trec_eval/.

The body of the report contains an overall analysis written by one attendee who is designated as the coordinator for that topic. This analysis is in turn based on an examination of each individual system. Responsibility for individual analyses rotated among participants so that each person worked with each system.

The individual analyses relied upon several general tools as well as the outputs of the system. The tools included using the SMART system in interactive mode to view term weights and retrieval outputs; beadplot, a tool from NIST that presents a visualization of run rankings along with relevance judgments; wui, a web-based tool built on top of the MultiText system from Waterloo that allows exploration of the different components of the index and the ranking process; and various facilities of the Clairvoyance system which make it easy to do small experiments to test hypotheses during the analysis.

Initially, failure analysis reports were free-form. Examples of this type of analysis include those for topics 355, 368, 384, 411, and 414. Later, the workshop participants developed a failure analysis template for both individual systems and for the overall report. For individual systems, the investigator(s) examined the top relevant and non-relevant documents, unretrieved relevant documents, and the system's base and expanded queries. Some reports include beadplot displays. They also tried to identify obvious blunders of the system, what the system might have done to improve effectiveness, what additional information would be needed to gain that improvement, and occasionally quirks in the relevance judgments. Summary analyses try to list failures common to several systems, notable failures unique to individual systems, winning strategies, classes of missed relevant and retrieved non-relevant documents, testable hypotheses, and any notes on the topic statement and/or relevance judgments.

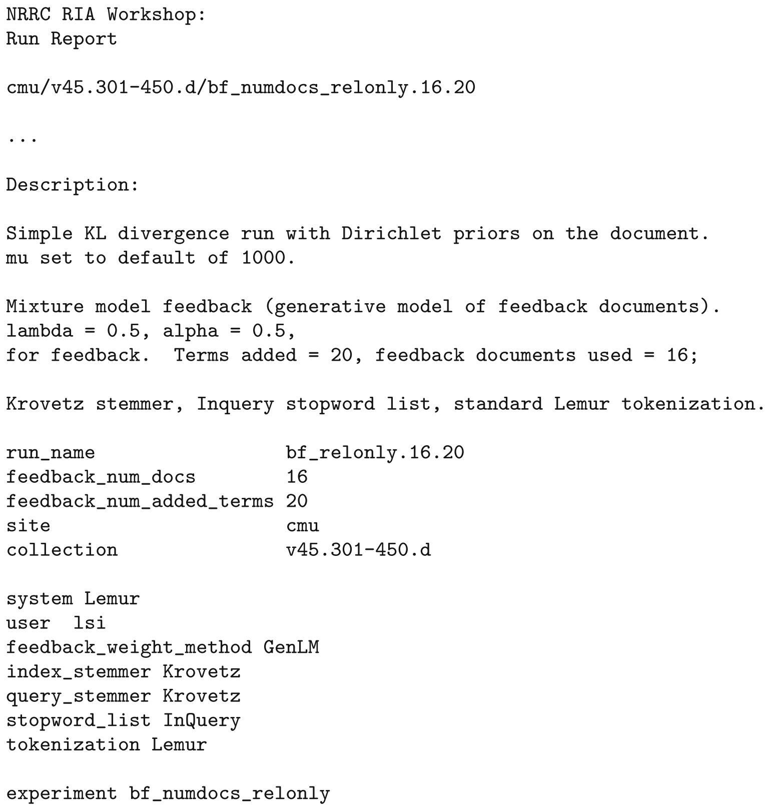
7 Runs

As we have said above, a *run*, in the TREC terminology, consists of the top *n* documents (usually 1000, but a system may return fewer if it wishes) from a given document collection for each topic in a given topic set, and represents the output of a single retrieval system with a single set of parameter settings. RIA runs are named according to the experiment in which they were produced; for example, the run "bf.15.20" is from the **bf_base** experiment and uses 15 documents and 20 expansion terms. Over the course of

the RIA workshop, hundreds upon hundreds of runs were created.

Each individual run has a "run page" which lists the parameters of the run as completely as possible, both to describe the run and to facilitate re-creating the run if needed. Often, the description is in the form of a shell script which, when executed, will regenerate the run. The run page also provides links to the run's document ranking (called the "results"), evaluation output from trec_eval, the query terms for all topics, and weighted query vectors if applicable. Figure 3 shows some information from the run page for a run of the **CMU** system in the **bf_numdocs_relonly** experiment.

The RIA run results browser, accessed via the "Run Results" link in the navigation bar, is organized to facilitate comparing runs between systems and across experiments. One can select runs from a single system, multiple systems, or from the group of "main systems" used at RIA, to compare side-by-side on a single page. Run results can be shown for individual topics or averaged across all topics in a collection.

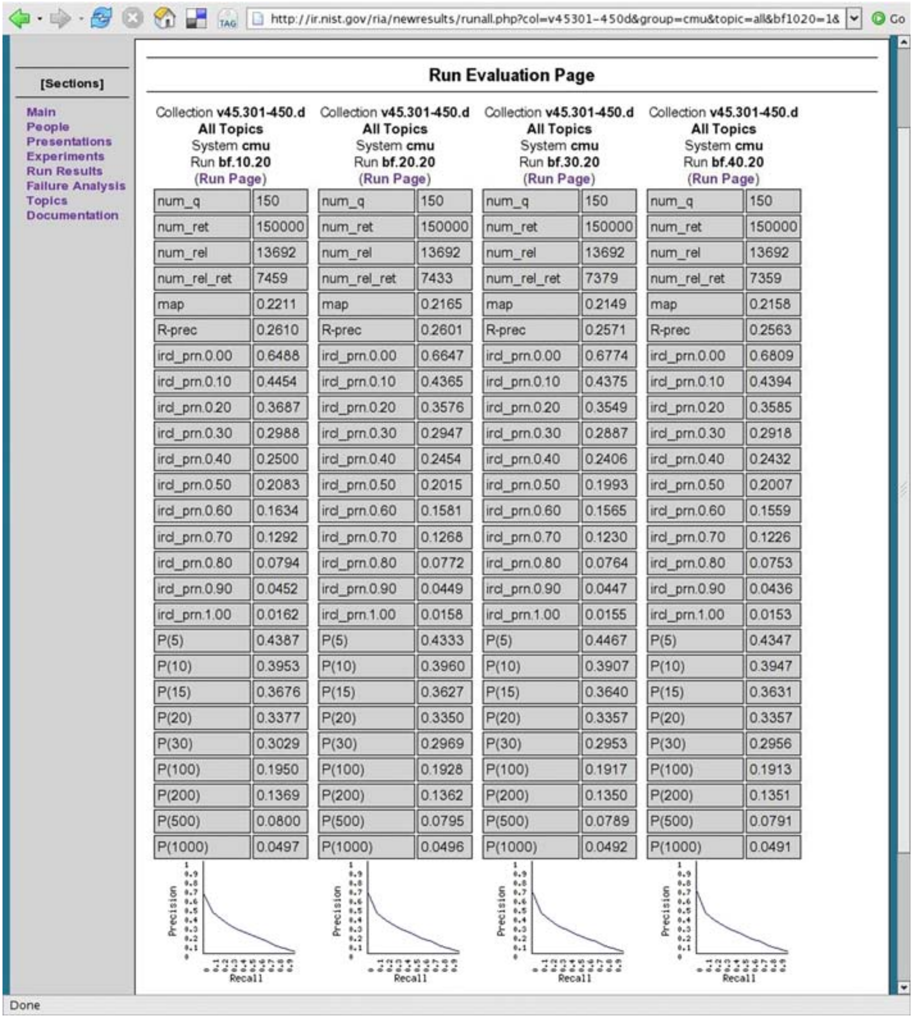


```
NRRC RIA Workshop:
Run Report

cmu/v45.301-450.d/bf_numdocs_relonly.16.20

...

Description:

Simple KL divergence run with Dirichlet priors on the document.
mu set to default of 1000.

Mixture model feedback (generative model of feedback documents).
lambda = 0.5, alpha = 0.5,
for feedback.  Terms added = 20, feedback documents used = 16;

Krovetz stemmer, Inquery stopword list, standard Lemur tokenization.

run_name                 bf_relonly.16.20
feedback_num_docs        16
feedback_num_added_terms 20
site                     cmu
collection               v45.301-450.d

system Lemur
user  lsi
feedback_weight_method GenLM
index_stemmer Krovetz
query_stemmer Krovetz
stopword_list InQuery
tokenization Lemur

experiment bf_numdocs_relonly
```

Fig. 3 A portion of a run page, describing the parameter settings for that run. Not shown is a shell script which re-creates the run

As an example of how one might use the run results browser, consider that we are interested in the effect of increasing the number of blind feedback documents given to the **CMU** system. Starting with the "Run results" navigational link, we choose the **v45.301-450.d** collection, and then the **CMU** system, and then to average across all topics. Next, from the runs page we choose "bf.10.20", "bf.20.20", "bf.30.20", and "bf.40.20" to compare this system with between 10 and 40 pseudo-feedback documents.

The end result is shown in Fig. 4. The evaluation results for a run appear in a single column. At the top of each column is a link to the run page, so we can easily see how that run was created. The evaluation scores include the standard measures produced by trec_eval, as well as a recall-precision graph. From Fig. 4, we can see that, at least from this limited exploration, that the **CMU** system seems to do best with 10 documents, and that adding more documents does not help. From here, we might choose to look at larger numbers of feedback documents, or switch gears to look at the number of feedback terms, or compare to another system, or look at a parallel experiment such as **bf_swapdocs** to see if document selection is the issue.



Run Evaluation Page

Collection **v45.301-450.d** **All Topics** System **cmu** Run **bf.10.20** (Run Page)		Collection **v45.301-450.d** **All Topics** System **cmu** Run **bf.20.20** (Run Page)		Collection **v45.301-450.d** **All Topics** System **cmu** Run **bf.30.20** (Run Page)		Collection **v45.301-450.d** **All Topics** System **cmu** Run **bf.40.20** (Run Page)	
num_q	150	num_q	150	num_q	150	num_q	150
num_ret	150000	num_ret	150000	num_ret	150000	num_ret	150000
num_rel	13692	num_rel	13692	num_rel	13692	num_rel	13692
num_rel_ret	7459	num_rel_ret	7433	num_rel_ret	7379	num_rel_ret	7359
map	0.2211	map	0.2165	map	0.2149	map	0.2158
R-prec	0.2610	R-prec	0.2601	R-prec	0.2571	R-prec	0.2563
ircl_prn.0.00	0.6488	ircl_prn.0.00	0.6647	ircl_prn.0.00	0.6774	ircl_prn.0.00	0.6809
ircl_prn.0.10	0.4454	ircl_prn.0.10	0.4365	ircl_prn.0.10	0.4375	ircl_prn.0.10	0.4394
ircl_prn.0.20	0.3687	ircl_prn.0.20	0.3576	ircl_prn.0.20	0.3549	ircl_prn.0.20	0.3585
ircl_prn.0.30	0.2988	ircl_prn.0.30	0.2947	ircl_prn.0.30	0.2887	ircl_prn.0.30	0.2918
ircl_prn.0.40	0.2500	ircl_prn.0.40	0.2454	ircl_prn.0.40	0.2406	ircl_prn.0.40	0.2432
ircl_prn.0.50	0.2083	ircl_prn.0.50	0.2015	ircl_prn.0.50	0.1993	ircl_prn.0.50	0.2007
ircl_prn.0.60	0.1634	ircl_prn.0.60	0.1581	ircl_prn.0.60	0.1565	ircl_prn.0.60	0.1559
ircl_prn.0.70	0.1292	ircl_prn.0.70	0.1268	ircl_prn.0.70	0.1230	ircl_prn.0.70	0.1226
ircl_prn.0.80	0.0794	ircl_prn.0.80	0.0772	ircl_prn.0.80	0.0764	ircl_prn.0.80	0.0753
ircl_prn.0.90	0.0452	ircl_prn.0.90	0.0449	ircl_prn.0.90	0.0447	ircl_prn.0.90	0.0436
ircl_prn.1.00	0.0162	ircl_prn.1.00	0.0158	ircl_prn.1.00	0.0155	ircl_prn.1.00	0.0153
P(5)	0.4387	P(5)	0.4333	P(5)	0.4467	P(5)	0.4347
P(10)	0.3953	P(10)	0.3960	P(10)	0.3907	P(10)	0.3947
P(15)	0.3676	P(15)	0.3627	P(15)	0.3640	P(15)	0.3631
P(20)	0.3377	P(20)	0.3350	P(20)	0.3357	P(20)	0.3357
P(30)	0.3029	P(30)	0.2969	P(30)	0.2953	P(30)	0.2956
P(100)	0.1950	P(100)	0.1928	P(100)	0.1917	P(100)	0.1913
P(200)	0.1369	P(200)	0.1362	P(200)	0.1350	P(200)	0.1351
P(500)	0.0800	P(500)	0.0795	P(500)	0.0789	P(500)	0.0791
P(1000)	0.0497	P(1000)	0.0496	P(1000)	0.0492	P(1000)	0.0491

Fig. 4 A comparison of blind feedback with the **CMU** system at 10, 20, 30, and 40 feedback documents

8 Conclusion

The RIA workshop data archive contains the results of hundreds of batch retrieval experiments across several well-known research systems and organized into a series of experiments. The archive is structured so that an outside researcher can not only benefit from the experimental results and failure analyses, but also arrange the runs in any way they wish in order to answer their own questions.

While development on the archive largely stopped after the workshop, we do hope to incorporate some improvements. As we have mentioned, it would be useful to be able to show aggregate statistics for arbitrary topic sets in the runs browser. Recently, we added automatically-generated recall-precision graphs to the runs display.

We hope that the RIA archive would grow to serve as a larger data repository for runs of all kinds, not just those created at the RIA workshop. We are aware of some other recent efforts to create these kinds of repositories, and it is likely that some broader discussion of the structural and access requirements of experiment repositories would be generally useful.

Acknowledgments and disclaimer The author is grateful for the support of ARDA and MITRE in sponsoring and hosting the RIA workshop. Any commercial products or companies mentioned in this paper are mentioned for descriptive purposes only and do not constitute an endorsement of any product or company.

References

Buckley, C. (2004). Why current IR engines fail. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004)* (pp. 584–585).

Harman, D., & Buckley, C. (2004). The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004)* (pp. 528–529).

Voorhees, E. M., & Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.