A Review of Relevance Feedback Experiments at the 2003 Reliable Information Access (RIA) Workshop. *

Robert H. Warren
School of Computer Science
University of Waterloo
Waterloo, ON Canada N2L 3G1
rhwarren@uwaterloo.ca

Ting Liu
ILS University at Albany
1400 Washington Ave.
Albany, NY USA 12222
tl7612@albany.edu

ABSTRACT

We review here the results of one of the experiments performed at the 2003 Reliable Information Access (RIA) Workshop, hosted by Mitre Corporation and the Northeast Regional Research Center (NRRC). The experiment concentrates on query expansion using relevance feedback and explores the behaviour of several information retrieval systems using variable numbers of relevant documents.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval

General Terms

Relevance feedback

1. INTRODUCTION

The RIA workshop sought to reduce the variability in the results returned by information retrieval systems. To do this a series of experimental runs titled bf_numdocs_relonly were attempted on the systems in use at the Workshop. These runs were designed to vary the number of documents being used for query expansion with the restriction that only documents judged to be relevant were to be used. By attempting the experiment on several of the systems used at the workshop, we were able to analyse the results while taking system implementation specifics under consideration.

1.1 Systems used in the experiments

The systems which were used in this experiment were conceived or operated by the University of Waterloo, Carnegie Mellon University, Sabir Research, City University of London, SUNY Albany and Clairvoyance Corporation. The Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts also participated at the Workshop, but because of system design issues was unable to run this experiment.

*This work was sponsored by the Northeast Regional Research Center which is funded by 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.

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

2. HYPOTHESIS

The experiments were designed to test a number of different hypotheses and questions:

- That the number of relevant documents used for query expansion would affect system retrieval performance.
- That the exclusive use of relevant documents to generate query expansion terms would effect the systems positively.
- That any document judged as relevant would have a positive effect on query expansion.

3. EXPERIMENTAL METHODOLOGY

For each system, a baseline run was created using no query expansion on the TREC 6, 7 and 8 topics with TREC disks 4 and 5 without Congressional Records [3]. The results of this run were then used as a ranked source of feedback documents for the rest of the experiments, while the Mean Average Precision (MAP) scores provided a reference performance measure.

In the bf_numdocs experiments, a series of runs was created which used the top nth documents of the baseline run as feedback documents for each topic. Runs were created for 1 to 20 feedback documents and then in steps of 5 documents until $n = 100$ documents were in use. In the bf_numdocs_relonly experiments, the same methodology was taken, but the feedback documents were filtered so that only relevant documents could be used.

One of the artifacts of the experiment's design was that the results were influenced by the rate of arrival of relevant documents within the baseline run and the absolute number of documents judged as relevant to the topic (which can be less than 100). Hence in a number of instances the bf_ numdocs_relonly results are stationary when compared to other experiments because no new relevant documents appeared within the baseline run at rank n. Some preprocessing was needed to ensure data alignment across runs and topics.

4. RESULTS

Figure 1 is a plot of the averaged MAP score for each increment of feedback documents on all seven systems. A cubic-spline was used to smooth the curve plot.

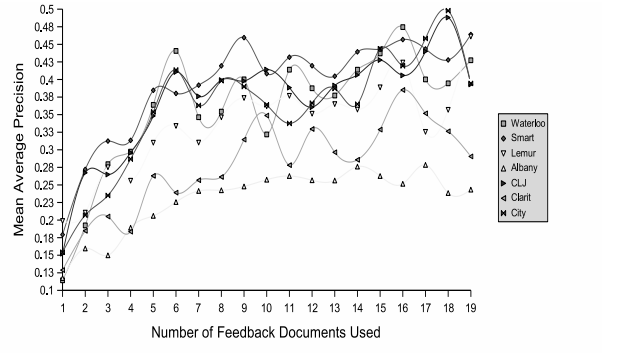

Figure 1: Averaged MAP scores for all seven participating systems.

One of the interesting elements common to all systems is a rapid initial rise in the system's score until about 6 relevant feedback documents are in use. Then the marginal benefit of additional documents seems to decrease asymptotically.

A possible explanation for this drop could be sampling error: as the value of n increases, there are fewer topics which have n feedback documents and the smaller population perturbs the averaged score. However, an inspection of the data reveals that this would only occur at about 11 relevant documents. It is more likely that this reflects the internal design optimisations of each system.

Interestingly, Albany's retrieval systems reveals an almost damped response to additional documents, while the Clarit (Clairvoyance) system can be observed to clearly oscillate in its performance with each additional new document.

Overall, there does seem to be an optimal number of relevant documents to use for query expansion. In all systems, document feedback begins to fail with below baseline MAP scores when too many (eg: $n > 40$, not graphed) relevant documents are used. A possible explanation for this advanced at the workshop [1] was that the system is attempting to "split" a query among too many different concepts. This would also explain why the same systems under the bf_numdocs experiments function well with 100 feedback documents: non-relevant documents would prevent too many coherent concepts [2].

Hence the exclusive use of relevant documents does not necessarily return a better result, especially when a large number of documents are in use.

The behaviour of the system on isolated topics also yielded some interesting results. Initially, we attributed drops in the MAP scores to the system running out of relevant documents and removing high scoring topics, lowering the average. However, this also occurred in situations where relevant documents were being added to specific topics.

Figure 2 is a plot representing the MAP scores of TREC topic number 343 for the baseline, bf_numdocs, bf_numdocs_relonly runs. It is an example of some of the dramatic changes commonly observed in the bf_numdocs_relonly experiment. Normally we would expect the addition of relevant documents to raise or maintain the score, but in this case the score lowers itself while the bf_numdocs run increases.

This points to some kind of issue in the feedback mechanism. An initial theory was that this was the result of an error in the system or that the document was forcing two different concepts onto the query at once (eg: splitting). Further inspection revealed that this was not necessarily a system specific problem, but that a number of systems were having difficulties with certain feedback documents.

An additional set of single document feedback experiments has identified at least 400 documents judged relevant which have a negative, non-zero effect on the systems of Sabir Research, Clairvoyance (CLJ), City University and CMU. While these may represent less than 4% of the topic relevance judgements, system specific results are more alarming, with feedback failures occurring with more than 40% of documents judged relevant.

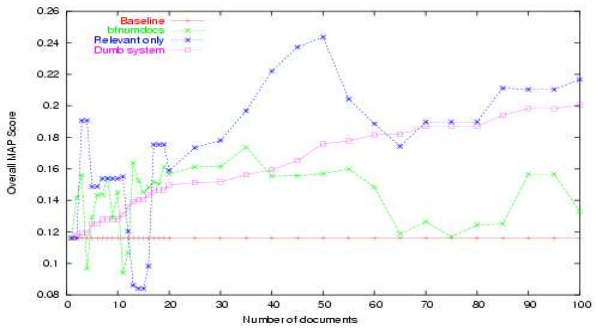

Figure 2: Comparison of experimental results for topic 343 with the Waterloo system.

The underlying reasons for this behaviour are unknown as of yet, but preliminary inspection of the query data would seem to discount the "query splitting" explanation. Hence, not all relevant documents can be used for effective query expansion.

5. CONCLUSIONS

- The incremental benefits of document feedback for query expansion seem to diminish after 6 relevant documents for most systems under study.
- The use of a large number of relevant documents for feedback documents may be counter productive and lower system performance.
- Early data indicates that some relevant documents perform poorly as feedback documents for query expansion independently of the system. This research area is currently under study [4].

6. REFERENCES

[1] C. Buckley. Why current i.r. engines fail. In *Proceedings of SIGIR2004*, Sheffield, England, 7 2004.

[2] J. Montgomery and L. Si. Effect of varying number of documents in blind feedback. In *Proceedings of SIGIR2004*, Sheffield, England, 7 2004.

[3] E. M. Voorhees and L. P. Buckland, editors. Department of Commerce, National Institute of Standards and Technology, November 2003. NIST Special Publication 500-255.

[4] R. H. Warren and E. L. Terra. Poison pills: notes on query expansion failure of relevant documents. In preparation.