# A Multi-System Analysis of Document and Term Selection for Blind Feedback

Thomas R. Lynam[†] Chris Buckley[*] Charles L. A. Clarke[†] Gordon V. Cormack[†]

[†]School of Computer Science, University of Waterloo, Canada
[*]Sabir Research Inc., Gaithersburg, Maryland

## ABSTRACT

Experiments were conducted to explore the impact of combining various components of eight leading information retrieval systems. Each system demonstrated improved effectiveness with the use of *blind feedback*, in which the results of a preliminary retrieval step were used to augment the efficacy of a secondary retrieval step. The hybrid combination of primary and secondary retrieval steps from different systems in a number of cases yielded better effectiveness than either of the constituent systems alone. This positive combining effect was observed when entire documents were passed between the two retrieval steps, but not when only the expansion terms were passed. Several combinations of primary and secondary retrieval steps were fused using the CombMNZ algorithm; all yielded significant effectiveness improvement over the individual systems, with the best yielding a an improvement of 13% ($p = 10^{-6}$) over the best individual system and an improvement of 4% ($p = 10^{-5}$) over a simple fusion of the eight systems.

## Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

## General Terms

Experimentation, Measurement, Performance

## Keywords

Blind Feedback, Pseudo-Relevance Feedback, Fusion

## 1. INTRODUCTION

We report a unique set of experiments that directly compare the document and term selection methods used for blind feedback in a number of information retrieval systems: CLARIT, CLJ, HITIQA, Lemur, MultiText, Okapi, and SMART. In all cases, principals involved in the development or deployment of these systems cooperated in the experiments and provided technical advice. Experts from each group worked together at a single location to exchange documents, terms and other information generated by various stages of the query evaluation process, allowing the individual sub-components of each system to be combined and compared in a controlled manner. Our primary goal was to achieve a better understanding of the blind feedback process, with the broader aim of improving adhoc retrieval performance. In addition, we hoped to determine if benefits could be gained through an exchange of information between systems at intermediate stages of the retrieval process.

To many observers, IR system performance on the traditional adhoc retrieval task reached a plateau several years ago. In their overview of the eighth Text REtrieval Conference (TREC), Voorhees and Harman [26] plot the performance of the SMART system — one of the better performing systems at TREC — from TREC-1 in 1992 to TREC-8 in 1999. The performance of the 1998 SMART system, as measured by mean average precision (MAP), is up to twice that of the 1991 system. While the past five years have seen the development of several novel IR techniques [1, 13, 18, 30, 31], most notably the large body of work on language modeling, we have not seen the dramatic performance improvements of the 1990's. Nonetheless, evidence provided in this paper indicates that additional performance gains may be available, and we outline areas of future work suggested by this evidence.

As witnessed by the diversity of systems participating in these experiments, the IR community has not converged on a single of set of techniques. The systems represented here vary widely in many respects, including blind feedback. Blind feedback was selected as a target area for careful examination, since it has received less attention than other areas and involves many tuning parameters and design choices. The work reported here differed from TREC in the level of control that could be exerted over the participating systems. Terms and documents from one system could be transfered to another, with experts guiding the process. Stopword lists, stemming algorithms and system parameters could be adjusted on demand.

The experiments reported in this paper were conducted as part of a workshop on Reliable Information Access (RIA) funded by the intelligence community through its Advanced Research and Development Activity (ARDA). The workshop was held at the ARDA Northeast Regional Research Center (NRRC) during the summer of 2003. Individuals and groups


*CIKM'04,* November 8–13, 2004, Washington, DC, USA.

from City University (London), Clairvoyant Corporation, Carnegie Mellon University, the MITRE Corporation, the National Institute of Standards and Technology, Sabir Research, SUNY Albany, the University of Massachusetts, and the University of Waterloo (Canada) participated in the workshop. Not all of the participants contributed experimental retrieval runs, but all were involved in the design and conduct of the experiments.

This paper reports the details of three NRRC-RIA experiments related to blind feedback. The first experiment (section 3) focuses on the exchange of feedback documents, the second (section 4) on term selection, and the third (section 5) on the fusion of intermediate results from multiple systems. In addition to the experiments reported here, the workshop conducted comparative experiments on other aspects of both information retrieval and question answering. These experiments will be reported and analyzed in concurrent and future publications.

## 1.1 Blind Feedback

Blind feedback (or "pseudo-relevance" feedback) incorporates two retrieval stages. In the first stage, the IR system formulates a set of query terms from the user's initial information request (or "topic") and retrieves the top $k$ documents. In the second stage, the system automatically selects additional query terms from this *initial document set* and adds them to the original query. Characteristics of the retrieved documents may be used to adjust weights for both the new and old query terms. The system then evaluates this augmented query to generate the final result set.

Figure 1 illustrates the blind feedback process. The diagram has been divided into two sections, one for each of the feedback stages. The first section is responsible for retrieving initial documents or sub-documents for the feedback section. The second section takes the initial documents or sub-documents and analyzes them to find additional terms. Each section is further separated into different components. All systems participating the NRRC-RIA workshop can be mapped to this general approach. In most systems the IR techniques used for the initial and final retrieval were identical, but several systems used a passage (or "sub-document") retrieval technique in the first stage and a document retrieval technique in the second.

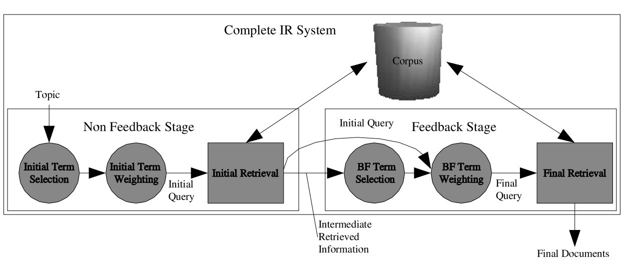


**Figure 1: Blind Feedback**

## 1.2 Evaluation Framework

The experiments utilize the description fields of 150 topics created for the TREC 6, 7 and 8 adhoc retrieval tasks [24–26]. Each definition field briefly describes documents relevant to the topic, often in a single sentence. For example, the description for topic 355 from TREC-6 is:

```
Identify the development and application of
spaceborne ocean remote sensing.
```

All participating systems formulated initial queries from these descriptions by eliminating stopwords and stemming the remaining words. Some systems identified phrases appearing in these topics and added these phrases to the query as additional terms, and some systems used term frequency within the description as a factor when computing term weights.

Queries were executed over a target corpus consisting of disks 4 and 5 of the TREC collection minus the *Congressional Record* documents. This corpus consists of 528,155 documents taken from the *Financial Times*, the *Federal Register*, the *Foreign Broadcast Information Service* and the *LA Times*. Relevance judgments for these documents are available for all 150 topics.

Following the conventions established at TREC, each experimental run consists of the top 1000 documents for each of the topics. Mean average precision (MAP) is our primary evaluation measure; the MAP values reported in this paper were computed by a variant of the `trec_eval` program. Further information on the TREC test collection is available from the TREC website[1].

# 2. PARTICIPATING SYSTEMS

This section briefly describes each system's method for accomplishing adhoc retrieval with blind feedback. Appropriate references may be consulted for further details. At the end of this section we summarize the performance of the individual systems at the start of the experiments, to establish a baseline for the remainder of the paper.

## 2.1 CLARIT and CLJ, Clairvoyant Corp.

For their participation in the NRRC-RIA workshop, Clairvoyance Corporation used two systems: CLARIT, a commercial information management toolkit written in C++, and CLJ (CLARIT Java), a recently developed IR research toolkit built on top of CLARIT. CLJ was included in the experiments primarily because it is a more suitable environment to make the rapid modifications required for the workshop.

CLARIT uses NLP for tokenization, storing individual words, noun phrases, and sub-phrases as index terms. It indexes on paragraph-sized sub-documents 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-document. Passages of different (often smaller) sizes can be re-created on the fly for query expansion.

CLARIT retrieval in the NRRC-RIA experiments used a traditional frequency-normalized TF-IDF algorithm. For query expansion based on the top $k$ documents, CLARIT used the method of Evans & Lefferts [8, 9]: Feedback is based on all sub-documents ranked equal to or higher than the highest scoring sub-documents found in the $k$-th retrieved document. The term selection (expansion) algorithm is based on either Rocchio or a customized method called "Prob2" (or "CLProb"), a variant of the traditional probabilistic term weighting algorithm [32]. CLJ runs with a slightly modified version of Okapi BM25 for standard retrieval and Prob2 for query expansion. In both CLARIT

[1] `trec.nist.gov`

and CLJ, expansion terms added to the source query are given differential weights based on the score/rank they receive under the selection algorithm.

## 2.2 HITIQA, SUNY Albany

SUNY Albany's contribution to the workshop was the HITIQA system, an analytic question answering system developed under the ARDA AQUAINT program [22, 23]. The HITIQA system is an interactive open-domain question answering technology designed to accept complex analytic questions in natural language. The HITIQA QA capabilities were utilized during the NRRC-RIA workshop to test the impact of different retrieval approaches on the effectiveness of question answering. HITIQA uses a document retrieval engine to fetch an initial set of documents from a database. At the RIA workshop, HITIQA used an older version of SMART for this purpose. The QA capabilities of HITIQA were not used in the experiments reported here, and the HITIQA results are essentially those generated by SMART.

## 2.3 University of Massachusetts Amherst

For the experiments reported in this paper the Center for Intelligent Information Retrieval at UMASS (CIIR) utilized a hybrid of query-likelihood and Relevance Models [13]. designed to fit the feedback strategies utilized by the other systems at the workshop. They performed a query-likelihood initial ranking of documents [18]. This ranking 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.

## 2.4 Lemur, Carnegie Mellon University

Carnegie Mellon's contribution used the KL-divergence based language modeling approach to text retrieval implemented in the Lemur Toolkit [14]. 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 [30]. 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.

## 2.5 MultiText, University of Waterloo

For participation in the NRRC-RIA workshop, the MultiText Project adapted passage-retrieval and term-extraction methods from their QA system to the task of blind-feedback query expansion. The MultiText QA passage-retrieval algorithm locates "hotspots" within the corpus where many query terms cluster in close proximity [5]. 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 same technique was used for their Robust track runs [29].

| System | # Docs | # Terms | Sub-Docs | No BF | BF | p value |
|---|---|---|---|---|---|---|
| CLARIT | 6 | 30 | Y | 0.169 | 0.188 (+11%) | 0.003 |
| CLJ | 6 | 20 | Y | 0.185 | 0.210 (+13%) | 0.020 |
| HITIQA | 20 | 100 | N | 0.126 | 0.154 (+22%) | $10^{-7}$ |
| UMASS | 30 | 100 | N | 0.196 | 0.235 (+20%) | $10^{-8}$ |
| Lemur | 10 | all | N | 0.201 | 0.225 (+12%) | $10^{-8}$ |
| MultiText | 25 | 25 | Y | 0.198 | 0.223 (+12%) | 0.001 |
| OKAPI | 10 | 20 | N | 0.186 | 0.216 (+12%) | $10^{-5}$ |
| SMART | 20 | 60 | Y | 0.204 | 0.226 (+11%) | 0.006 |
| Fusion | | | | 0.220 | 0.255 (+16%) | |

**Table 1: Baseline System Performance**

## 2.6 Okapi, City University London

City University's contribution to the NRRC-RIA workshop used the Robertson/Sparck Jones Probabilistic model [20], implemented in the Okapi System [17]. 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 [21]. The Term Selection Value devised by Robertson [19] was used for term selection in pseudo relevance feedback experiments. The full range of the Okapi BSS was used to support the experiments including passage processing and term extraction.

## 2.7 SMART, Sabir Research

SMART is a flexible IR research engine based on the vector space model as developed by Gerard Salton [2, 28]. 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 NRRC-RIA. The only slightly non-standard setting used in NRRC-RIA was the use of SMART statistical phrases.

## 2.8 Baseline System Performance

At the start of the workshop each group generated a blind feedback run using the standard parameters and methods of their own system. The details of these runs are summarized in Table 1. The "# Docs" column lists the number of document the system used for feedback. The "# Terms" column lists the number of terms added during the feedback process; original term are not included in this count. A system using "Sub-Docs" may use less than an entire document for feedback.

In the next two columns the table gives MAP values for two runs: a non-feedback run ("No BF") and a blind feedback ("BF") run. The "No BF" runs represents each system's default behavior when blind feedback is disabled. These are not simply the documents returned by the first retrieval

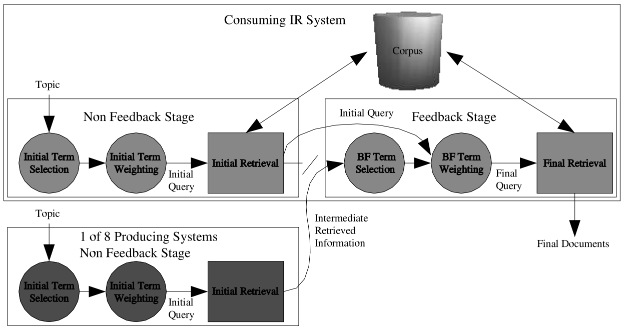


**Figure 2: Swapping Feedback Documents**

stage when blind feedback is enabled, since some systems use different tuning parameters when blind feedback is not used. These runs illustrate the impact of blind feedback, which is greater than 10% for all systems and more than 20% for some. All differences between the non-feedback and blind feedback runs are significant (Wilcoxon signed-rank test, $p < 0.01$) [2] None of the differences between the top-four systems is significant.

The last row of the table gives the results of fusing the runs using the well-established CombMNZ algorithm [10]. It is well known that combining results from more than one system can produce a higher performance than any system alone [7]. CombMNZ is a simple but effective fusion algorithm, often used as a baseline in fusion experiments [15]. The CombMNZ algorithm first normalizes the relevance scores from each system for each topic. A new score is computed for each document by summing the document's normalized score from each system and then multiplying by the number of systems that retrieved the document.

The NRRC-RIA workshop provided an unusual opportunity to experiment with fusion at early stages of the blind feedback process. These experiments will be reported in a later section. The CombMNZ algorithm is used in all fusion experiments reported in this paper.

# 3. SWAPPING DOCUMENTS

Each of the eight systems that participated in the workshop can be divided into the two stages of our feedback model (Figure 1). This experiment (the "swapdoc" experiment) investigated the effect of interchanging the two stages among the systems as shown in Figure 2.

## 3.1 Swapdoc — Experimental Design

For swapdoc, the initial stage of each system was used to produce a fixed number of documents, which then became the input to the second stage of other systems.

Some systems use sub-documents instead of full documents. Fortunately these systems used only the best sub-document from any given document; therefore, the document containing the best sub-document could be provided to the other systems. Systems using sub-documents needed an additional step after receiving documents from another system, to find the best sub-document from the documents provided.

[2] We use the Wilcoxon signed-rank test throughout.

| | SMART |
|---|---|
| 1 | **Method for Realizing Spaceborne SAR Digital Imaging** |
| 2 | International Telecommunications Advisory Committee |
| 3 | Exports, Reporting and recordkeeping requirements |
| 4 | Sino-Canadian 'Global Radar Remote Sensing Plan |
| ... | ... |
| 20 | Applications for Duty-Free Entry of Scientific ... |
| | Lemur |
| 1 | Prospects for Satellite Remote Sensing in Oceano-... |
| 2 | **Method for Realizing Spaceborne SAR Digital Imaging** |
| 3 | 1993 AIST Laboratory Research Plans |
| 4 | Environmental Research in the Interests of the Armed |
| ... | ... |
| 20 | Germany, Netherlands Cooperate on ENVISAT |

**Table 2: Topic 355 Top Document Titles**

| System | CLARIT | CLJ | HITIQA | UMASS | Lemur | MultiText | OKAPI | SMART |
|---|---|---|---|---|---|---|---|---|
| CLARIT | 100% | 28% | 22% | 34% | 34% | 29% | 32% | 38% |
| CLJ | | 100% | 23% | 37% | 36% | 25% | 36% | 31% |
| HITIQA | | | 100% | 30% | 29% | 19% | 28% | 42% |
| UMASS | | | | 100% | 88% | 34% | 55% | 42% |
| Lemur | | | | | 100% | 33% | 53% | 40% |
| MultiText | | | | | | 100% | 30% | 26% |
| OKAPI | | | | | | | 100% | 37% |
| SMART | | | | | | | | 100% |

**Table 3: Initial Document Overlap (top 60)**

Systems vary in the number of documents or sub-documents required for blind feedback; the maximum number of documents used by any system was 60. Therefore, each system was required to provide 60 documents per topic. Systems were permitted to choose their own best parameters for the number of documents and feedback terms, but were required to keep these values constant for all topics. Otherwise, documents provided by other systems were processed using each system's normal term selection and weighting methods.

Overlap, defined as the percentage of documents in one run that are also found in the other, is an indicator of the similarity of the runs. As an example, Table 2 shows the document titles for the top initial documents from Lemur and SMART for topic 355. The average overlap across topics for the 60 initial documents is detailed in Table 3. Average overlap ranges from 22% to 88%.

## 3.2 Swapdoc — Results

Since systems are tuned to their own documents, we expected systems to achieve their best performance with their own documents and lower performance with documents from other systems. However, the experiment revealed that using another system's documents can often increase performance, and that the choice of the producing system can have an enormous performance impact. All sixty-four combinations of producing and consuming systems are presented in Table 4. All systems do as well or better when consuming documents produced by SMART than they do with their own documents.

| Producing System | Consuming System CLARIT | CLJ | HITIQA | UMASS | Lemur | MultiText | OKAPI | SMART |
|---|---|---|---|---|---|---|---|---|
| CLARIT | 0.192 | 0.203 | **0.179+** | 0.245 | 0.231 | 0.237 | 0.215 | 0.209 |
| CLJ | **0.201+** | 0.207 | **0.189+** | 0.235 | 0.227 | 0.238 | 0.216 | **0.203-** |
| HITIQA | **0.145-** | 0.204 | 0.154 | **0.220-** | **0.171-** | 0.221 | **0.173-** | 0.207 |
| UMASS | *0.212+* | 0.216 | **0.187+** | 0.235 | 0.205 | 0.237 | 0.220 | 0.216 |
| Lemur | *0.211+* | 0.213 | **0.188+** | 0.235 | 0.225 | 0.238 | 0.218 | 0.217 |
| MultiText | **0.183-** | 0.197 | **0.173+** | *0.233-* | 0.223 | 0.219 | 0.201 | **0.199-** |
| OKAPI | **0.218+** | 0.222 | **0.184+** | 0.234 | *0.235+* | 0.226 | 0.217 | 0.216 |
| SMART | **0.214+** | 0.215 | **0.191+** | 0.241 | **0.244+** | *0.234+* | 0.224 | 0.226 |

"+" significant increase, "-" significant decrease ($p < 0.05$ , $\mathbf{p < 0.01}$ ) when compared to own initial documents

**Table 4: Document Swapping — Mean Average Precision**

| Producing System | Consuming System CLARIT | CLJ | HITIQA | UMASS | Lemur | MultiText | OKAPI | SMART |
|---|---|---|---|---|---|---|---|---|
| CLARIT | 0.182 | 0.208 | **0.139+** | 0.195 | 0.215 | 0.224 | 0.215 | 0.207 |
| CLJ | *0.190+* | 0.211 | **0.145+** | 0.191 | 0.222 | 0.226 | 0.215 | 0.208 |
| HITIQA | 0.172 | 0.215 | 0.127 | 0.191 | **0.224+** | 0.216 | 0.208 | **0.187-** |
| UMASS | **0.200+** | *0.218+* | **0.140+** | 0.191 | 0.218 | 0.224 | 0.221 | 0.213 |
| Lemur | **0.196+** | 0.222 | **0.139+** | 0.192 | 0.214 | 0.224 | 0.215 | 0.216 |
| MultiText | 0.189 | 0.215 | **0.137-** | *0.182-* | **0.212-** | 0.210 | 0.220 | *0.202-* |
| OKAPI | **0.201+** | 0.221 | **0.140+** | 0.192 | **0.221+** | 0.217 | 0.211 | 0.208 |
| SMART | **0.162-** | 0.228 | **0.148+** | 0.204 | **0.227+** | 0.222 | 0.219 | 0.221 |

"+" significant increase, "-" significant decrease with ($p < 0.05$ , $\mathbf{p < 0.01}$ ) when compared to own initial documents

**Table 5: Document Swapping with Fixed Parameters (20 documents, 5 terms) — Mean Average Precision**

| BF System | Own | Producing Average | | Consuming Average | |
|---|---|---|---|---|---|
| CLARIT | 0.192 | 0.218 | (+2.1%) | 0.198 | (+2.9%) |
| CLJ | 0.222 | 0.218 | (+4.0%) | 0.211 | (-5.1%) |
| HITIQA | 0.154 | 0.192 | (-12.7%) | 0.187 | (+21%) |
| UMASS | 0.235 | 0.213 | (+2.6%) | 0.235 | (+0.0%) |
| Lemur | 0.225 | 0.216 | (+3.3%) | 0.219 | (-2.5%) |
| MultiText | 0.219 | 0.201 | (-4.2%) | 0.233 | (+6.4%) |
| OKAPI | 0.217 | 0.219 | (+4.1%) | 0.209 | (-3.5%) |
| SMART | 0.226 | 0.223 | (+6.8%) | 0.210 | (-7.3%) |

**Table 6: Swapping Document Averages**

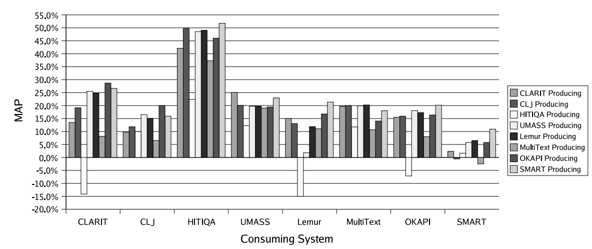


**Figure 3: Swapping Documents Feedback Improvements**

Table 6 shows that some systems are good consumers and others are good producers. "Producing Average" is the average performance of all eight systems on the given system's document set. "Consuming Average" is an average of the given system's performance using all the document sets produced by the eight systems. HITIQA has an average increase of 21% using other systems' documents while other systems have an average decrease of 12.7% when using HITIQA documents. This would indicate that HITIQA (which incorporates an older version of SMART) is a poor producer of documents. In contrast, consuming SMART documents gives an average increase of 6.8%, suggesting that SMART produces superior documents.

The swapdoc experiment demonstrates that blind feedback improves performance independent of consuming and producing systems (Figure 3). The graph reveals feedback improvement of systems consuming documents produced from other systems. HITIQA has the largest improvement in performance from feedback, but it was also the lowest performing system. SMART, which produced the best initial documents, had the lowest improvement from feedback.

## 3.3 Swapdoc with Fixed Parameters

For the swapdoc experiment each system used its standard blind feedback settings, including number of documents and number of terms. In order to facilitate the term swapping experiment discussed in the next section, we re-ran the experiment with fixed parameters: 20 initial feedback documents and 5 terms. The results appear in Table 5. The performance of many systems is seriously harmed by fixing the parameters. Once again, SMART is the best producer of feedback documents.

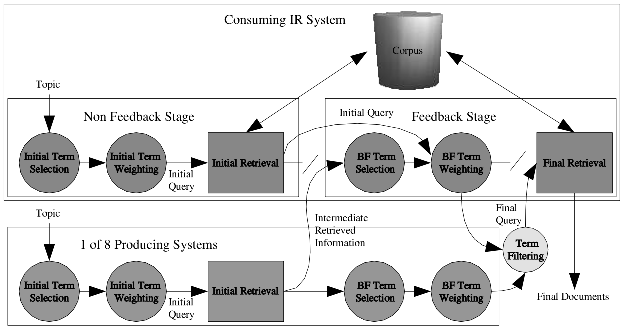


**Figure 4: Swapping Feedback Documents and Terms**

| System | CLARIT | CLJ | HITIQA | UMASS | Lemur | MultiText | OKAPI | SMART |
|---|---|---|---|---|---|---|---|---|
| CLARIT | 100% | 12% | 2% | 5% | 6% | 8% | 9% | 9% |
| CLJ | | 100% | 6% | 14% | 20% | 26% | 32% | 27% |
| HITIQA | | | 100% | 8% | 7% | 7% | 9% | 8% |
| UMASS | | | | 100% | 25% | 18% | 19% | 20% |
| Lemur | | | | | 100% | 20% | 24% | 26% |
| MultiText | | | | | | 100% | 29% | 25% |
| OKAPI | | | | | | | 100% | 30% |
| SMART | | | | | | | | 100% |

**Table 7: Term Overlap (top 5 terms)**

# 4. SWAPPING TERMS

The focus of the second experiment (the "swapterm" experiment) was to determine the term selection component's effect on performance for different systems. For this experiment, each system produced a set of terms using its blind feedback method and consumed terms generated by other systems. The process diagram for the swapterm experiment appears in Figure 4.

## 4.1 Swapterm — Experimental Design

Swapping terms is far more complicated than swapping documents. Externally, documents may be represented by document identifiers, and a document swap requires only an exchange of these identifiers. On the other hand, systems represent terms in different ways, and there is no simple, common representation. Due to these differences, one system may produce terms that cannot be consumed by another.

Stemming is the most common reason why a system produces a term that another cannot consume. For example, some systems stem the word "antarctic" to "antarct-" while others do not stem it at all. This problem is compounded by the fact that many systems only index the stemmed version. Therefore, the unstemmed version would not appear in an stemmed corpus or, vice versa. To solve this problem, systems were required to produce the most common unstemmed version of each term; each system then applied its normal stemming method.

Phrases also had to be considered, since not all systems could consume phrases. If a system could consume a phrase it was kept, otherwise it was discarded. In the majority of cases, single-word terms could be consumed by all systems. In the few cases where terms could not be consumed by all systems, the terms contained unusual features, such as numbers, or were stopwords in the consuming system.

There is surprisingly little overlap in the term sets. Term overlap percentages are given in Table 7. For the example topic 355, the following table shows the top 5 terms produced by Lemur and SMART:

| | Lemur | SMART |
|---|---|---|
| 1 | research | **satellite** |
| 2 | **satellite** | radar |
| 3 | technology | surface |
| 4 | basic | image |
| 5 | data | science |

Another problem we needed to overcome was that each system uses a different term weighting scheme. To solve this problem, each system used it's own scheme to compute weights for consumed terms. To ensure that weights could be computed for these terms, the producing system also provided the documents from with the terms were taken. To simplify this process, we fixed the feedback parameters for the experiment at 20 documents and 5 terms.

Given documents and terms from a producing system, the consuming system stemmed the terms and generated weights by processing the documents provided. These weighted terms and original query were combined into a new query and executed to retrieve the final result set.

## 4.2 Swapterm — Results

The results of the swapterm experiments appear in Table 8. The baseline for this table is a system's own performance when feedback parameters are fixed at 20 documents and 5 terms (Table 5).

In all but a few cases, consuming terms generated by another system harms performance. One exception is CLARIT consuming SMART's terms which gives a 24% improvement. The terms provided by CLARIT harm performance in most cases, but this effect is most likely due to the numerous phrases produced by CLARIT, which could not be consumed by other systems.

# 5. SWAPPING AND FUSING

Fusing results from multiple systems often produces better performance than that of the individual systems alone [7]. In a third experiment (the "swapfuse" experiment) we examined the impact of fusion on feedback, both by fusing initial document sets and by fusing the results from a single system when consuming different sets of initial documents. Having access to intermediate data allowed us to explore fusion over a wide range of system parameters.

Figure 5 summarizes the results of the swapfuse experiment. The first two bars in each group represent a system's baseline performance from Table 1, with and without blind feedback. The fusion results from that table are also reproduced. For each system, the last three bars in the group represent three different methods of fusing it with the other systems. In "Initial Result Fusion" the system consumes an initial result with is the fusion of the initial results from all of the eight systems. "Fused Common Consumer" fuses the results of the system over eight different runs, each using the initial documents produced by a different system.

| Producing System | Consuming System | | | | | | |
|---|---|---|---|---|---|---|---|
| | CLARIT | HITIQA | UMASS | Lemur | MultiText | OKAPI | SMART |
| CLARIT | 0.181 | **0.131-** | *0.174-* | 0.215 | **0.208-** | 0.202 | **0.183-** |
| CLJ | **0.207+** | *0.140-* | 0.190 | 0.216 | 0.222 | 0.210 | 0.203 |
| HITIQA | 0.182 | 0.125 | **0.155-** | *0.221-* | *0.202-* | **0.192-** | *0.182-* |
| UMASS | 0.190 | 0.141 | 0.191 | 0.208 | 0.216 | **0.199-** | 0.203 |
| Lemur | 0.192 | 0.138 | **0.155-** | 0.214 | *0.215-* | **0.196-** | **0.200-** |
| MultiText | 0.186 | 0.137 | 0.163 | 0.212 | 0.211 | **0.202-** | *0.195-* |
| OKAPI | 0.194 | 0.141 | 0.176 | 0.219 | **0.198-** | 0.202 | **0.201-** |
| SMART | **0.201+** | 0.144 | *0.179-* | **0.229+** | 0.224 | 0.212 | 0.220 |

"+" significant increase, "-" significant decrease with ($p <$ *0.05* , $\mathbf{p < 0.01}$ ) when compared to own documents

**Table 8: Swapping Terms — Mean Average Precision**

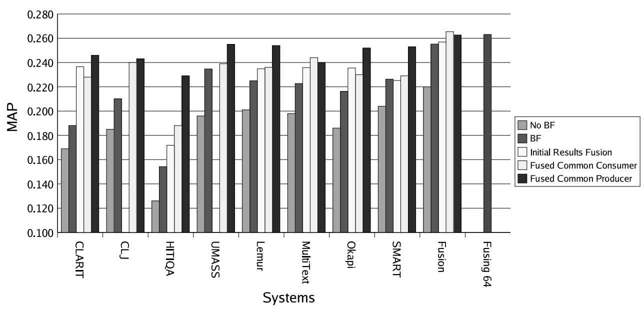


**Figure 5: Fusion Results**

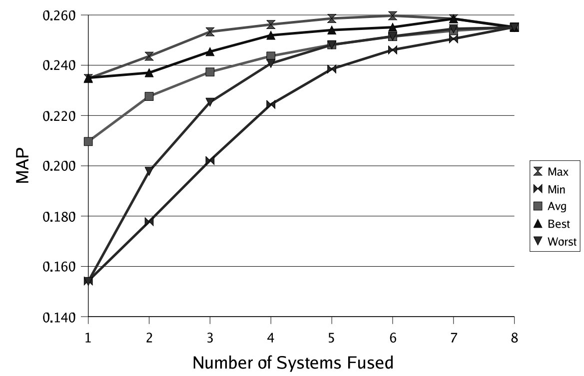


**Figure 6: Fused Systems**

"Fused Common Producer" combines the results of eight different runs, each using the results of the system's initial run consumed by a different system.

The group titled "fusion" gives "fusion of fusion" results. Each column represents the fusion of results corresponding to the eight corresponding columns to the left. Finally the lone column labelled "Fusing 64 Combinations" gives the fused result of 64 runs representing all combinations of producer and consumer components.

Of all the results, "Fused Common Consumer" gives the best MAP of 0.2654, 13% better ($p = 10^{-6}$) than the best BF system, and 4% better ($p = 10^{-5}$) than the simple fusion of the 8 BF runs. "Fused Common Consumer" is marginally better than "Fusing 64 Combinations" ($p = 0.04$) and "Fused Common Producer" ($p = 0.04$).

## 5.1 Fusing Subsets

We investigated the possibility that one or more of the systems might produces a deleterious effect on the system, and that fusion would be improved by removing these systems. To this end, we created all 255 non-empty subsets of the 8 BF runs and computed the MAP for the fusion of the results from each subset.

Figure 6 shows several views of these results as a function of the size of the subsets. The curve "Max" shows the best result that could have been achieved had we an oracle to identify the particular subset that would achieve the best score. This curve provides an upper bound on what we might have achieved. At $n = 6$, this value exceeds that for $n = 8$ (the fusion of all systems) by only a small amount. The complementary curve "Min" shows the worst result that could have been achieved.

"Best" presents the result of using the best $n$ systems, as ranked by their individual MAP results. At least for this set of systems, little is gained from excluding the worst-performing system. "Worst" presents the complimentary result. From this curve we may conclude that excluding one or two of the best systems has little impact on the effectiveness of fusion.

"Avg" presents the average MAP over all subsets. From this curve we may conclude that picking a random subset of the systems produces intermediate results for small subsets, but results comparable to "Worst" for larger subsets.

# 6. CONCLUDING DISCUSSION

In an effort to better understand the blind feedback process, researchers from nine institutions compared eight different information retrieval system on a component by component basis, swapping and combining information after each retrieval stage. By decomposing the original systems into their individual components and examining each component separately, we discovered several areas where additional performance improvements might be obtained.

From the document swapping experiments (Section 3), we were surprised to discover that one system can outperform another as a producer of feedback documents, even when there is no significant difference in the original systems. For example, there is no significant difference between the baseline performance of SMART, Lemur and MultiText, with or without blind feedback (Table 1). Nonetheless, the performance of both Lemur and MultiText improves significantly when documents from SMART's first retrieval stage are used for blind feedback. Indeed, the best overall performance in the document swapping experiment was obtained by Lemur when consuming documents from SMART.

Overall, this experiment reveals that initial document selection greatly affects performance. One system consuming feedback documents produced by another system can give better performance than either system can achieve alone. In general, a system that produces initial documents that increase performance for one consuming system tends to increase performance for all consuming systems. Conversely, a system that produces initial documents that decreases performance for one consuming system tends to decreases performance for all consuming systems.

First-stage retrieval (for blind feedback) and second-stage retrieval (for presentation to the user) should be treated as different problems. The characteristics that make a document good for feedback are not necessarily those that make it more likely to be relevant. For example, a document may discuss one facet of a request in detail but ignore another, making it a rich source of terms related to the first facet, but at the same time, not relevant. Some systems, such as CLARIT/CLJ and MultiText already reflect this observation, by using a passage retrieval technique in the first stage.

Term swapping (Section 4) did not generate the same level of improvement as document swapping. In most cases, swapping terms hurts system performance, though there were cases where swapping terms greatly increased performance (Table 8). However, we note that the average overlap among the top terms is fairly low, between 2% and 32% (Table 7). In future work we will exploit this observation by exploring ways of combining term lists generated by difference feedback techniques. A term ranked highly by many techniques may be a better expansion term than one ranked highly by a single technique.

The best overall performance was achieved by swapping documents and fusing the results (Section 5). Despite the simplicity of our fusion technique, this result was 13% better than that achieved by any system on its own. Like the term swapping experiment, the results of the fusion experiment suggest that a systematic approach to combining intermediate results may produce substantial benefits.

In this paper, we have discussed three of the main experiments undertaken at the NRRC-RIA workshop. In later papers we hope to explore the ideas arising from these experiments, some of which we have outlined above. A number of short papers about other NRRC-RIA experiments were presented at SIGIR 2004 [3, 4, 6, 11, 12, 16, 27] and additional publications are planned for the future. By the end of 2004 we expect the raw data from the workshop, including thousands of experimental runs, to be made available to general information retrieval research community.

## 7. ACKNOWLEDGMENTS

We thank Donna Harman for her leadership in organizing and running the NRRC-RIA workshop. The workshop participants, all of whom contributed to the experiments reported in this paper, included: Jamie Callan, Kevyn Collins-Thompson, Andres Corrada-Emmanuel, David Evans, Warren Greiff, Zhenmei Gu, David Hull, Paul Kantor, Ting Liu, Andy MacFarlane, Luo Ming, Jesse Montgomery, Paul Ogilvie, Sean Ryan, Luo Si, Sharon Small, Ian Soboroff, Tomek Strzalkowski, Egidio Terra, Jeff Terrace, Ellen Voorhees, Robert Warren, and Yi Zhang.

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.

## 8. REFERENCES

[1] Gianni Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Tranactions on Information Systems*, 20(4):357–389, October 2002.

[2] Chris Buckley. Implementation of the SMART information retrieval system. Technical Report 85-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.

[3] Chris Buckley. Topic prediction based on comparative retrieval rankings. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[4] Chris Buckley. Why current IR engines fail. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[5] Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *SIGIR 2001*, New Oreans, Louisiana, 2001.

[6] Kevyn Collins-Thompson, Jamie Callan, Egidio Terra, and Charles L.A. Clarke. The effect of document retrieval quality on factoid question answering performance. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[7] W. Bruce Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*. Kluwer Academic Publishers, 2000.

[8] David A. Evans and Robert G.Lefferts. Design and evaluation of the CLARIT-TREC-2 system. In *Second Text REtrieval Conference*, Gaithersburg, MD, 1994.

[9] David A. Evans and Robert G.Lefferts. CLARIT-TREC experiments. In *Third Text REtrieval Conference*, Gaithersburg, MD, 1995.

[10] E. A. Fox and J. A. Shaw. Combintation of multiple searches. In *Second Text REtrieval Conference*, 1994.

[11] Zhenmei Gu and Ming Luo. Comparison of using passages and documents for blind relevance feedback in information retrieval. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[12] Donna Harman and Chris Buckley. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[13] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. In *SIGIR 2001*, pages 120–127, New Orleans, 2001.

[14] Lemur. The Lemur toolkit for language modeling and information retrieval. Available at http://www-2.cs.cmu.edu/ lemur/.

[15] Mark Montague and Javed A Aslam. Condorcet fusion for improved retrieval. In *Eleventh International Conference on Information and Knowledge Management (CIKM 2002)*, pages 538–548, McLean, Virginia, 2002.

[16] Jesse Montgomery, Luo Si, Jamie Callan, and David A. Evans. Effect of varying number of documents in blind feedback: analysis of the 2003 NRRC RIA workshop "bf_numdocs" experiment suite. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[17] Okapi. Details of okapi and its functionality. http://www.soi.city.ac.uk/ andym/OKAPI-PACK/.

[18] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR 1998*, pages 275–281, Melbourne, Australia, 1998.

[19] S.E. Robertson. On term selection for query expansion. *Journal of Documentation*, Vol 46:p359–364, 1990.

[20] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, Vol 27(p129-146), May-June 1976.

[21] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Third Text REtrieval Conference*, Gaithersburg, MD, 1995.

[22] Small et al. *A Data Driven Approach to Interactive Question Answering (in preparation).* AAAI/MIT Press, 2004.

[23] Strzalkowski et al. *Advances in Open-Domain Question Answering (in preparation).* Kluwer Academic Publishers, 2004.

[24] Ellen Voorhees. Overview of the Seventh Text REtrieval Conference (TREC-7). In *7th Text REtrieval Conference*, Gaithersburg, MD, 1998.

[25] Ellen Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *6th Text REtrieval Conference*, Gaithersburg, MD, 1997.

[26] Ellen Voorhees and Donna Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *8th Text REtrieval Conference*, Gaithersburg, MD, 1999.

[27] Robert H. Warren and Ting Liu. A review of relevance feedback experiments at the 2003 reliable information access (RIA) workshop. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom, 2004.

[28] D. Williamson, R. Williamson, and M. Lesk. The Cornell implementation of the SMART system. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 2, pages 43–44. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[29] David L. Yeung, Charles L. A. Clarke, Gordon V. Cormack, Thomas R. Lynam, and Egidio L. Terra. Task-specific query expansion. *Text REtrieval Conference 2003*, 2003.

[30] ChengXiang Zhai and John Lafferty. Model-based-feedback in the language modeling approach to information retrieval. *Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, 2001.

[31] ChengXiang Zhai and John Lafferty. Two-stage language model for information retrieval. In *SIGIR 2002*, pages 49–56, Tampere, Finland, 2002.

[32] ChengXiang Zhai, Xiang Tong, Peter Jansen, and David A. Evans. The CLARIT system TREC-6 report. In *Sixth Text REtrieval Conference*, Gaithersburg, MD, 1998.