

# The Role of Manually-Assigned Keywords in Query Expansion

Kazem Taghva,\* Julie Borsack, Thomas Nartker, and Allen Condit  
Information Science Research Institute  
University of Nevada, Las Vegas  
Las Vegas, NV 89154-4021  
phone: 702-895-0873 fax: 702-895-1183

---

\*taghva@isri.unlv.edu

## Abstract

We report on two types of experiments with respect to manually-assigned keywords to documents in a collection. The first type of experiment examines the usefulness of manually-assigned keywords to automatic feedback. The second type of experiment considers the potential benefits of these keywords to the user as an interactive tool. Several experiments were run and compared. The results of these experiments indicate that there is no gain in average precision when manually-assigned keywords are used for query expansion. Further, manually-assigned keywords did not aid the user as an interactive tool for document understanding.

Keywords: query expansion, document keywords, relevance feedback, retrieval effectiveness, user satisfaction

## 1 Introduction

The set of experiments described in this paper was motivated by a combination of issues including current Licensing Support Network (LSN) regulations and changes in information retrieval (IR) technology. The LSN is a planned system that will capture and track documents pertaining to the site licensing proceedings of the Nuclear Regulatory Commission for the Yucca Mountain Project (YMP) high level nuclear waste repository. This system will need to provide access to the general public with varied levels of domain expertise and search capabilities.

When the LSN was first proposed, the most commonly used IR systems were based on the classic Boolean exact match model. Studies have shown that, in general, Boolean systems can be improved if keywords are manually-assigned to documents in the collection. The initial design of the LSN included augmenting documents with keywords to improve retrieval—a time intensive and expensive task.

Other technologies (e.g. probabilistic and vector space) have since been introduced that improve precision and recall over the well-known Boolean model and may deem manual assignment of keywords unnecessary. But precision and recall are not the only factors that we felt were important to test with respect to manually-assigned keywords.

Keywords may aid users by helping them decide if the documents returned are relevant. Keywords may also be used to formulate new queries for feedback. So besides just testing the value of keywords with respect to precision and recall, we wanted to get an idea of the effect of manually-assigned keywords for other areas of user satisfaction. There are several facets to our study. We test:

- automatic query expansion with manually-assigned keywords.
- how helpful keywords may be when presented interactively as part of a document.
- whether making manually-assigned keywords available to the searcher for feedback will increase their search effectiveness.
- whether the searcher gains additional information about the documents when keywords are displayed.
- whether keywords help the user learn more about the collection over time.

Unlike user-centric studies in IR that try to characterize user behavior for improving effectiveness (see Section 2), our study's focus is solely on the *influence* of manually-assigned keywords on the user and vector space system IR performance. The availability of keywords to the user and their impact on the IR system are the only variables under investigation.

This project's significance does not just have an impact for searchers of the LSN collection. We believe that we are reporting on a more general question. *How might manually-assigned keywords aid users searching a document collection?* These studies were carefully designed to try and look at all conceivable ways in which manually-assigned keywords may be useful to a searcher and may improve IR performance.

## 2 Related Research

Several IR experiments have observed the way in which users interact with search systems and intermediary searchers to improve our understanding of interactive evaluation measures, methods of query expansion, and relevance feedback. Many studies observe users to identify techniques for improving the interactive search process. In this section, we relate our work to these previous studies and show how it augments the IR community's understanding of the use of keywords by searchers in an interactive search environment.

### 2.1 Interactive Evaluation Measures

Su points to three major approaches that underly existing performance measures: relevance, utility, and user satisfaction (Su, 1999). As mentioned in the introduction, our studies include precision and recall with respect to a baseline run. But these relevance criteria were not enough to determine "utility" or "user satisfaction" that may be attained with document-assigned keywords.

Several of the utility criteria in (Su, 1999) were not applicable to our study. Our focus was solely on the utility of keywords. We did though try to measure *value of search results as a whole* through our short survey filled out by the user after each query and again in the final questionnaire (See Section 3 and Appendix A). Again, user satisfaction was aimed at the usefulness of keywords but the final questionnaire covers several of the measures listed as important indicators of IR performance in (Su, 1999) including *searcher's contribution* and *search results measures*.

Spink discusses more precise relevance measures including levels of relevance, negative relevance, and the median effect in (Spink, 2001). The relevance judgements for our test collection were binary as is the case for many experimental collections.

Spink's paper is quite recent. It may take some time before standard experimental collections are impacted by her new approach.

The evaluation measures we use in this study are not exhaustive. But through the use of standard precision and recall, examining keyword utility, and carefully considering user satisfaction, we are able to judge the usefulness of document-assigned keywords to users of the LSN.

## **2.2 Automatic Query Expansion**

Automatic query expansion is another way to evaluate the potential usefulness of manually-assigned keywords. There are several algorithms in the literature for automatic query expansion and our goal was not necessarily directed at comparing expansion of assigned keywords to any particular method(Rocchio, 1971; Robertson & Jones, 1976; Jones & Webster, 1980; Harper, 1980; van Rijsbergen, Harper, & Porter, 1981; Smeaton & van Rijsbergen, 1983; Belkin & Croft, 1987; Efthimiadis & Robertson, 1989). Our objective was to compare keyword expansion with our baseline run (see Experiment A) and to include a well-accepted query expansion algorithm for reference (see Experiment B). Since automatic query expansion is not available in Fulcrum, we implemented a version of Rocchio(Rocchio, 1971) for this purpose.

Measuring the usefulness of manually-assigned keywords for automatic query expansion was the objective of this part of the study. Little prior research was comparable with this combination. But considering that several older collections still have keywords (descriptors), we felt it necessary to give it consideration(Fidel, 1991b).

## **2.3 Interactive Relevance Feedback**

The automatic query expansion we refer to is not completely automatic. As Efthimiadis(Efthimiadis & Robertson, 1989) points out, the query expansion we apply in our

experiments results in “a form of feedback to the user” since *user-selected* documents were used for query expansion. We’d like to distinguish it clearly though from *interactive relevance feedback* where the user plays an active role in selecting terms for query reformulation. In (Efthimiadis & Robertson, 1989), four term selection methods for query reformulation are identified:

1. use only original query terms
2. use original query terms and terms from some other source
3. use combinations of terms from original query and terms from documents judged relevant
4. abandon original query terms and use only those from the retrieved set of documents

Our research applied a combination of items 2 and 3 where the “some other source” is the document-assigned keywords. The original query terms were always used; terms from the relevant documents were always made available; the keywords were made available half the time to determine if users found them useful.

Few researchers explore the interactive component of retrieval. This is understandable when one considers that just selecting the appropriate evaluation measures can be perplexing. But studies are emerging that will help us define and analyze the interactive search process(Su, 1999; Spink, 1997b, 1997a).

## **2.4 The Human Element**

One of the best ways to determine search needs is through user observation. Studies by Spink(Spink, 1994), Jansen et al.(Jansen, Spink, & Saracevic, 2000), Lucas(Lucas & Topi, 2002), and Fidel(Fidel, 1991a, 1991b, 1991c) provide insight into a user’s search

behavior and performance in retrieval tasks. All these studies show that search term selection is crucial to retrieval performance. In (Spink, 1994), Spink goes further and analyzes the sources of terms that aid the user most. Our study compliments these studies by analyzing the usefulness of manually-assigned keywords in particular as source terms.

We focus directly on the utility of manually-assigned keywords on searcher performance. Selection of document-assigned keywords by users is the only variable under investigation. For example, all original queries run by the users were identical and all first generation documents reviewed by the users were identical. Each user viewed 20 queries where returned documents were displayed with and without keywords. Since the queries were predefined and recall and precision were known, it gave us a unique ability to measure improvements in the experimental runs with respect to these relevancy measures. We found no other relevant research that applied both these standard measures in their studies.

### **3 Research Design**

#### **3.1 Experimental Collection Set**

The nucleus of the data collected for our experiments is the sample documents chosen from the LSN. Our sample collection includes 1055 documents, 40 test queries, relevancy judgments, and document-assigned keywords. Following is a short description of these components.

**1055 Documents** The 1055 documents were selected because of their rich content and their importance to the site viability assessment. These documents in particular would be of interest to users of the Licensing Support Network. Example titles include *Recommended Sealing Requirements for the Revision of the Controlled*

Table 1: Experimental Collection Statistics

Collection Statistics	
Document count	1055
Number of pages	75,236
Average document length (pages)	71
Median document length (pages)	34

Figure 1: Sample Query

*Find documents which describe natural ore bodies that contain radioactive elements, and correlate these ore bodies to the Yucca Mountain Project in order to assess radioactive decay rates.*

*Design Assumption Document and Viability Assessment Mined Geologic Disposal System Test and Evaluation Plan.* Table 1 shows some collection statistics.

**40 Test Queries** 40 queries were developed, reviewed, and revised by a team of Yucca Mountain records personnel and ISRI staff who were familiar with the collection. An example query appears in Figure 1.

**Relevancy Judgments** Relevancy judgments were assigned by a group of 5 graduate geology students. The complete documents were scanned so that they could be read online. The students were presented the list of 40 queries (also online) and were able to select the queries to which each document was relevant. Each document was read by two geologists. If there was a discrepancy in their judgments, a third geologist would re-evaluate the document and resolve the differences. We collected binary relevancy judgments so that standard recall and precision could be measured. There are a total of 4000 relevant documents averaging 100 relevant documents per query.

**Manual Keyword Assignment** Document keyword assignment was a similar but distinct process from determining relevancy judgments. An experienced consultant

was employed to train an independent group of 6 geology students in document keyword assignment. These geologists read the 1055 documents. Each document was read by 2 geologists and keyword assignments were made. Sources for keyword<sup>1</sup> selection included: YMP developed vocabularies, geologic thesauri, document text, and their own expertise in the subject area. The union of term selections from the two geologists were assigned to the document. There are on average approximately 9 keywords assigned to each document. An example set of assigned keywords to Document MOL.19980518.0157 appears below.

repository operations, repository design, transporter, personnel management, design constraints, design specifications, heat load, working environment, ramp, shaft, tunnel, design criteria, emplacement holes

All data described above was stored in an Oracle relational database.

### **3.2 Retrieval Environment**

The retrieval system we use, Fulcrum SearchServer 3.7e(Hummingbird Communications Ltd., 1999), is an example of a system that uses the vector space model. Vector space IR systems differ from Boolean systems in both query formation and in the method of determining relevance. In a pure Boolean system, documents are represented as inverted lists of terms and queries are written as keywords connected with logical operators (AND, OR, NOT). No notion of query/document similarity is measured in a Boolean system and therefore, the returned documents are not presented in relevance order. In the vector space model, both documents and queries are represented as vectors. Many vector space models, including Fulcrum, employ statistical term weighting to represent the documents in a collection. Using these statistical methods, retrieved

---

<sup>1</sup>Note that keywords include key phrases as well.

documents can be ranked and thus, tend to give more information about a document's relevance to a query.

We apply Fulcrum's *linguistic processing* capability and its statistical ranking algorithm called *critical terms ordered*(Hum, 1999). Linguistic processing reduces terms to their uninflected forms thus expanding the query with related words; critical terms ordered uses both document term occurrence and collection frequency to determine a document's similarity to a query.

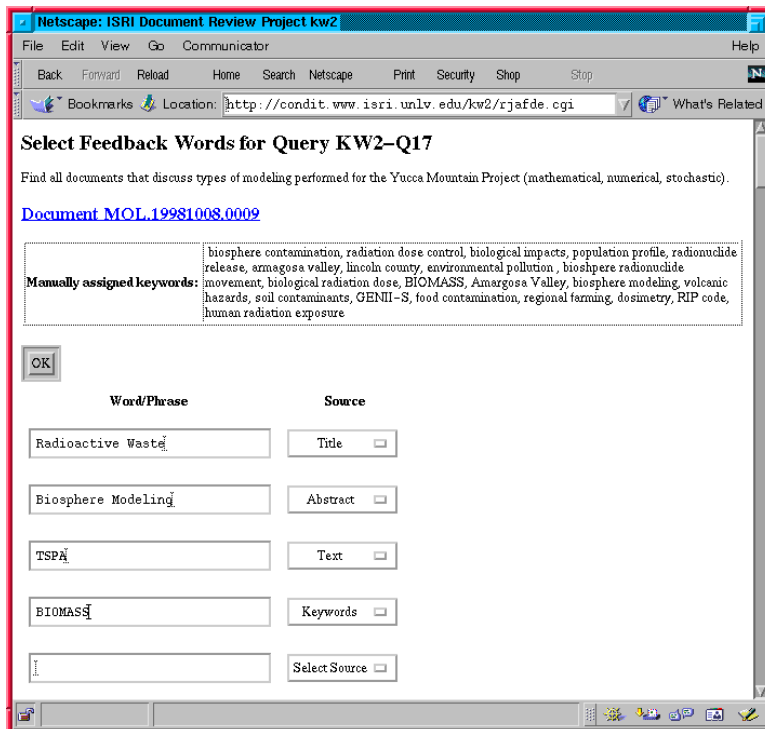
### **3.3 Experimental Data Capture**

The experiments described in Section 4 included 4 students acting as "users" of the LSN. They were all university students from varying majors but were not "experts" and had only limited knowledge of the planned Yucca Mountain Repository. They would be considered intelligent and informed citizens from the public domain. Each had an average amount of experience using retrieval systems. For each query, each user performed the following tasks:

1. They selected relevant documents from a set returned from the original query run (the top 25 documents returned by Fulcrum).
2. They selected words from the relevant set in 1 above for *interactive query expansion* (from the title, abstract, text, and if displayed, keywords).
3. After running the expanded query, they were asked to determine if any of the "new" documents retrieved were relevant.
4. They then responded to the short survey in Figure 3.

Document titles were displayed for the retrieved set and linked to the complete document for review. Figure 2 shows the interactive term feedback screen. Displayed

Figure 2: Interactive Term Feedback Screen



is the text of the query, one of the documents the user had selected as relevant, a list of the previously assigned keywords for that document, and input boxes for the user to add feedback terms to the original query.

Each user was presented with all forty queries in order. Randomly though, only half of the time keywords were displayed. Each user viewed 20 queries where the documents were displayed with keywords and 20 queries where the documents were displayed without keywords. This randomization and responses to the survey gave us the ability to see if the keywords were helpful to the user if they were displayed with the document. When the keywords were displayed, the user had the ability to select from this list when making term selections for interactive query expansion. This analysis helped us: 1) see if the user “learned more” about the documents for those documents with keywords, and 2) examine whether these words were useful for expansion.

We include a short survey after each query in hopes of acquiring some insight into any cognitive learning that a user might experience. Knowing the goals of the project, survey questions were developed by ISRI staff. Figure 3 shows the survey questions displayed to the users.

Note that none of the users were aware that the goal of this project was to evaluate the usefulness of keywords. When they were presented with a document that included them, using this survey, we tried to discover any increase in knowledge that may have been gained after reviewing queries with and without keywords.

Each question had a multiple choice response: 1 Definitely Yes, 2 Mostly Yes, 3 Neutral, 4 Mostly No, 5 Definitely No.

Again, the user environment was web-based. It was consistent for all queries and across users. Data collected were stored in Oracle.

Figure 3: Short Survey

**Question 1:** Have the documents you've reviewed for this query aided you in finding more relevant documents to this query?

**Question 2:** Were the terms available to you for query expansion helpful in finding more relevant documents to this query?

**Question 3:** After reviewing documents thus far, do you feel you have a better understanding of the topics that may be contained in this collection?

A final questionnaire was developed to obtain a broader view. It was designed by ISRI staff with an attempt to discern the user's perspective on the availability and usefulness of keywords in a more retrospective way. The complete questionnaire with user responses can be found in Appendix A. Although our goal was to compare *interactive term relevance feedback*(Spink, 1997a) with and without keywords, we do touch on the cognitive learning that takes place during the feedback cycle over time(Allen, 1994). Our purpose was to discover the following information:

**Topic 1:** How good did the users perceive the document collection to be?

**questions: 2, 3, 11**

**Topic 2:** How well did the search engine perform?

**questions: 4, 5, 10**

**Topic 3:** Were the keywords previously assigned to the documents helpful?

**questions: 6, 7, 8, 9**

**Topic 4:** Would other/better tools improve results?

**questions: 12, 15, 16**

**Topic 5:** How much did the users learn during the query retrieval process?

**questions: 1, 13, 14**

Our analysis of the questionnaire responses appears in Section 5.

## 4 Experiments

The following experiments were run to evaluate the usefulness of keywords if applied to retrieval automatically or if the keywords were supplied to a user interactively. The first experiment, experiment *A*, is a baseline run. It gives us something with which to compare the results of our other experimental runs. Experiments *B* and *C* compare two forms of automatic expansion: *B* using system selected terms and *C* using the manually-assigned keywords. Experiments *D* and *E* evaluate the usefulness of keywords to the user by having them available for review and for interactive expansion. A complete analysis of the results follows.

### 4.1 Experiment Description

**Experiment *A*** gathers baseline information for our collection using Fulcrum and the set of queries as described in Section 3. We run the queries against the document collection and calculate precision at standard recall points. This test tells us how well Fulcrum does without any assistance. Results for this baseline run appear in column A of Table 2.

**Experiment *B*** tests system implemented automatic query expansion. With *Automatic query expansion* or automatic feedback, the system uses statistical information to automatically select the “best” terms from a user’s selection of relevant returned documents. There are several techniques proposed in the literature, but most apply some form of term re-weighting and query expansion. Since some IR systems apply automatic query expansion, these runs can also be used to compare how well a system can choose words (keywords) for feedback from user-selected documents. Unfortunately, automatic query expansion is not available in Fulcrum

Table 2: 11-Point and Average Precision

Recall	Precision						
	$A$	$B_1$	$B_2$	$C_1$	$C_2$	$D$	$E$
0.00	0.72	0.55	0.57	0.63	0.65	0.82	0.78
0.10	0.55	0.41	0.43	0.46	0.50	0.66	0.65
0.20	0.44	0.33	0.35	0.36	0.39	0.54	0.52
0.30	0.37	0.27	0.27	0.29	0.31	0.44	0.43
0.40	0.30	0.22	0.22	0.23	0.24	0.36	0.36
0.50	0.25	0.18	0.19	0.19	0.20	0.30	0.29
0.60	0.20	0.16	0.16	0.16	0.16	0.23	0.23
0.70	0.17	0.13	0.14	0.13	0.14	0.18	0.18
0.80	0.14	0.12	0.12	0.12	0.12	0.14	0.14
0.90	0.12	0.11	0.11	0.11	0.11	0.12	0.12
1.00	0.10	0.10	0.10	0.10	0.10	0.10	0.11
Average:	0.3055	0.2345	0.2418	0.2527	0.2655	0.3536	0.3464

but we have implemented a version of the *Standard Rocchio* method (Salton & Buckley, 1990) (based on user document selection) and augmented the queries for this purpose. Note that this external implementation may not give the same results that an IR system would give if it had been an integral tool of the system. We run this experiment twice, first expanding the original query with the top 50 terms ( $B_1$ ) and then again with the top 20 terms ( $B_2$ ). Columns  $B_1$  and  $B_2$  of Table 2 shows the 11-point precision results and the averages for these values for the automatic query expansion runs.

**Experiment  $C$**  uses the same set of relevant retrieved documents identified in  $B$ , but in this experiment, we expand the queries automatically with the keywords that had been pre-assigned to these documents. Experiment  $C$  helps determine if we can use manually-assigned keywords for query expansion automatically. For this experiment we concatenate all manually-assigned keywords from a user's selected relevant documents (identified in  $B$ ) and rerun the query. All the keywords from all the documents are used to augment the original query causing

Figure 4: Keyword Augmented Query

**Query KW2-Q02:** *Find documents which describe natural ore bodies that contain radioactive elements, and correlate these ore bodies to the Yucca Mountain Project in order to assess radioactive decay rates.*

**Added Keywords for  $C_1$ :** chemical investigation planning coprecipitation geochemical analysis procedures geochemistry modeling program ground water chemistry hydrogeologic properties mass transfer mineral composition mineralogy and petrology probabilistic risk analysis radionuclide movement radionuclide solubility rock chemistry solute transport speciation dissolution geohydrochemistry mathematical modeling radionuclide migration

**Added Keywords for  $C_2$ :** radionuclide chemistry modeling analysis

the number of keywords added to be quite high. We noticed that in general, the meaning of the original query was overshadowed by the large number of added keywords. There is no simple way to select the “best” keywords from the list, however, we did limit them by only augmenting the query with the keywords associated with more than one document. An example query and the list of keywords that were used to augment it appear in Figure 4. The precision results without keyword limitation and with keyword limitation appear in columns  $C_1$  and  $C_2$  respectively of Table 2.

**Experiment  $D$**  applies interactive query expansion from the text of the user-selected documents. The title, abstract, and document text only are displayed to the user for term selection; keywords are *not* displayed (Spink, 1995). The user is then asked to expand his query with terms he has selected. The purpose of this experiment is to test a user’s ability to select relevant words directly from the document and see if, without keywords, he can improve his results. An example of one user’s selected terms when keywords were not displayed appears in Figure 5, item 2. 11-point precision results for experiment  $D$  appears in column  $D$  of Table 2.

Figure 5: Interactive Expansion with and without Keywords

**Query KW2-Q02:** *Find documents which describe natural ore bodies that contain radioactive elements, and correlate these ore bodies to the Yucca Mountain Project in order to assess radioactive decay rates.*

**Terms Selected/No Keywords:** Koongarra uranium ore deposit natural system ore body

**Terms Selected/with Keywords:** nuclear criticality stability of radioactive solid uranium ore deposits thorium ore deposits radioactive natural analogs

Table 3: Feedback Term Selections

Term Type	W/O Keywords	W/ Keywords
text	3061	3020
abstract	152	100
title	173	97
keywords		750
total term feedback words assigned	3386	3967

**Experiment E** also applies interactive query expansion, but this time it allows the searcher to select terms from the pre-assigned keywords as well as the parts of the document text listed in *D*. Experiment *E* is another way of evaluating the usefulness of keywords. If a user has access to terms already considered important, can he selectively use these words to improve his query? A list of the words selected by one of our user's appears in Figure 5. The 11-point precision results of experiment *E* appear in column *E* of Table 2.

By comparing experiment *E* to experiment *D*, we obtain more insight into the knowledge that may be gained by having the ability to view keywords. Looking at Table 3, note that the total number of feedback terms selected increased by approximately 600 terms. It seems in most cases, *more words* were selected for interactive expansion because the keywords were made available. This seems to suggest that having keywords augmented the selection but it did not replace

using terms derived from the document itself.

## 4.2 Interpretation of Recall/Precision Results

Our experiments can be grouped as follows: *baseline* (Experiment *A*), *automatic query expansion* (Experiments *B*<sub>1</sub>, *B*<sub>2</sub> and *C*<sub>1</sub>, *C*<sub>2</sub>), and *interactive query expansion* (Experiments *D* and *E*). The baseline result shows the coarse effectiveness given by Fulcrum and all the other results are compared against it. In what follows, we explain the impact of query expansions on retrieval effectiveness and consequently, what role the manually-assigned keywords play.

The results for automatic query expansion (*B*<sub>1</sub> and *B*<sub>2</sub> vs *A*) shows a drop in retrieval effectiveness compared to the baseline. We believe that the cause for this decrease is due to the fact that the original query terms were dominated by the expanded terms. This result is not new to the literature. Previous studies show that too many extra terms can overshadow original query terms (van Rijsbergen et al., 1981; Smeaton & van Rijsbergen, 1983). This observation is apparent when you compare the results of *B*<sub>1</sub> to *B*<sub>2</sub>. In other words, when we expand the queries with only 20 new words, we get better results than when we expand using 50 words. The same is true when we compare *C*<sub>1</sub> with *C*<sub>2</sub>. Again, what is significant is that there is no difference in retrieval effectiveness between the two automatic query expansion runs (*B*) or between the interactive query expansion runs (*C*).

Table 2 reveals improvements achieved with interactive query expansion: Experiment *D* is nearly 16% improved; Experiment *E* is over 13% improved. These experiments allowed the user to expand their queries by choosing terms from relevant documents. Experienced online searchers typically follow search techniques such as the *building block strategy* or the *citation pearl growing strategy* to build sophisticated Boolean queries (Efthimiadis, 1996; Harter, 1986; Bates, 1981). In our vector space

environment, the chosen terms were simply added to the original query for expansion.

The dramatic increase in average precision for these experiments support the importance of user observation studies done by Efthimiadis, Spink, Jansen, Lucas, and others (Efthimiadis, 2000; Spink, 1997b, 1997a; Jansen et al., 2000; Lucas & Topi, 2002). These studies show, as we show here, the importance of human/IR interaction for retrieving relevant material. What is significant in our study however, is that there is no difference if the keywords were displayed to the user or not (the average difference between  $D$  and  $E$ ). Simply put, the manually-assigned keywords played no role in improving interactive query expansion.

## 5 The User's Perspective

The goal of this project was not just to study the impact of keywords on the *traditional IR model* as defined by Spink in (Spink, 1997b), but to ascertain the usefulness of manually-assigned keywords as an interactive tool for document understanding and feedback. We tried to capture evidence of the possible interactive value through a short survey and by interviewing each user when they had completed all forty queries. The methodology and interpretation of each is described below.

### 5.1 Survey Results

As mentioned in Section 3, after completing the tasks associated with each query, the users were required to respond to a short survey.

Table 5 averages and totals the responses to the survey questions when keywords were displayed (Experiment  $E$ ) and when they were not (Experiment  $D$ ). The lower the average, the more positive the users felt after reviewing this set of queries. The queries are also broken down in Table 5 into groups of 10 to try and capture the learning that

Table 4: Averaged Short Survey Results

Question	KW2-Q01 thru KW2-Q10	
	D	E
1	3.15	3.30
2	3.35	3.50
3	2.15	2.10
	KW2-Q11 thru KW2-Q20	
1	2.15	2.20
2	3.00	2.30
3	1.85	1.75
	KW2-Q21 thru KW2-Q30	
1	2.20	3.20
2	2.60	3.35
3	2.05	1.80
	KW2-Q31 thru KW2-Q40	
1	2.80	2.40
2	3.25	2.85
3	2.20	1.95
	Totals	
1	10.30	11.10
2	12.20	12.00
3	8.25	7.60

takes place just from the experience of reviewing additional queries.<sup>2</sup>

Our survey seems to echo the same results we conclude from our analysis of the recall/precision results. The users do not seem to give any indication that manually-assigned keywords are a significant help for interactive query expansion or that they give more information about a document's content. They seem satisfied with using just the document terms for query expansion (as indicated in Table 3). Further, Table 5 shows us that over time, our users tend to get better at selecting terms for query expansion but having keywords available to them did not assist them.

## 5.2 Final Questionnaire Analysis

The final questionnaire interviewed each user at the completion of the project. The complete questionnaire with user responses can be found in Appendix A. Although our goal was to compare *interactive term relevance feedback*(Spink, 1997a) with and without keywords, we do touch on the cognitive learning that takes place during the feedback cycle over time(Allen, 1994). Our purpose was to discover the following information:

**Topic 1:** How good did the users perceive the document collection to be?

**questions: 2, 3, 11**

Reviewing questions 2, 3, and 11, the responses seem to be dependent on the user's prior knowledge within the domains covered. It seems the more knowledge the user had of the subjects, the more they felt the collection may not have been adequate. Also, the more background, the more they felt expertise was required to fully grasp the topics fully.

**Topic 2:** How well did the search engine perform?

**questions: 4, 5, 10**

---

<sup>2</sup>The queries were displayed to the user in numerical order.

In general, our users seemed to be satisfied with the search engine's performance. There seemed to be a consensus that the original query returned relevant documents in the top twenty documents returned. They seemed to realize fairly quickly that saturating the query with too many additional terms did not help the search engine find more relevant documents.

**Topic 3:** Were the keywords previously assigned to the documents helpful?

**questions: 6, 7, 8, 9**

This set of questions focused on our research objective: Are manually-assigned keywords useful as an interactive tool to the user? And although the responses seem mixed, the answers given for question 9 tell us that manually-assigned keywords were of no assistance. Every user felt that the terms they chose themselves from the content of the documents were the most useful. Although no description of keyword assignment was given to our users, their views could be attributed to uncertainty of keyword quality similar to what Fidel noted in (Fidel, 1991b). Note though that these questionnaire responses substantiate both the recall/precision and survey results as well.

**Topic 4:** Would other/better tools improve results?

**questions: 12, 15, 16**

The majority felt that a domain specific thesaurus would be helpful for the general user. Their consensus that thesauri would be useful conforms to studies by Fidel(Fidel, 1991b). They unanimously felt that the title was not enough to determine relevance; three of the four thought that displaying the table of contents would be extremely beneficial. These responses tell us that displaying more information from the document (abstract, table of contents, summaries) is extremely useful for the user.

**Topic 5:** How much did the users learn during the query retrieval process?

**questions: 1, 13, 14**

Responses to questions 1, 13, and 14 tell us users *learn* during the search and retrieval process. This learning process was also exhibited in the short survey results (see Table 5). What they learned seemed to be related to what they already knew and their personal interests. As pointed out by Spink (Spink, 2002), the learning process is an obvious by-product of the IR interaction but is not considered in the design of the traditional IR model. Capturing this gain in knowledge and incorporating it into a retrieval system should result in a much more satisfied user.

## 6 Conclusion

The introduction lists several facets of our study. We address the implications of our experiments for each below:

- automatic query expansion with manually-assigned keywords.

The results from these experiments ( $C_1$ ,  $C_2$ ) show degradation when compared to the baseline run  $A$ . These results echo previous expansion results found in the literature (van Rijsbergen et al., 1981; Smeaton & van Rijsbergen, 1983). Also, the manually-assigned keywords applied for query expansion were *document related*. It may have been the case that some of the document-assigned keywords were unrelated to the query causing divergence of the topics initially being searched.

- how helpful keywords may be when presented interactively as part of a document.

With so few users in our study we are unable to generalize our conclusions. We can state however that in this study, there was no indication that the manually-assigned keywords aided the users for query expansion or for imparting information about the document collection. Again, document-assigned keywords may introduce topics not under consideration in the initial query.

- whether making manually-assigned keywords available to the searcher for feedback will increase their search effectiveness.

Our experiments show no significant difference in search effectiveness when manually-assigned keywords are displayed to the user (compare experiments *D* and *E*). Both the recall/precision results and the comments on the final questionnaire indicate no increase in search effectiveness when users had access to the manually-assigned keywords.

- whether the searcher gains additional information about the documents when keywords are displayed.

Our results are based solely on the four users in our study and again, cannot be generalized. Both the short survey and the final questionnaire though seem to indicate that at best, the manually-assigned keywords were of little value, and at worst, they can actually cause confusion.<sup>3</sup>

- whether keywords help the user learn more about the collection over time.

Our short survey tried to discover any cognitive learning our users may have experienced; the final questionnaire queried the users more directly. Although user familiarity with the collection was apparent over the course of our study, we cannot attribute any additional learning to the manually-assigned keywords.

---

<sup>3</sup>See **user 4**'s response to question 7 and 8.

The goal of this project was to investigate the effects of manually-assigned keywords from a user's interactive perspective as well as with traditional automatic feedback. The results of the experiments, the survey, and the questionnaire imply that at least for the LSN collection, manually-assigned keywords do not aid automatic or manual query expansion. Further, manually-assigned keywords offer little information to the user as an interactive tool.

## 7 Acknowledgments

We would like to thank Amanda Spink for her suggestions and support. We would also like to thank the anonymous referees for their critical reading of this paper.

## References

- Allen, B. (1994). Cognitive Abilities and Information System Usability. *Information Processing and Management*, 30(2), 177–191.
- Bates, M. (1981). Search Techniques. *Annual Review of Information Science and Technology*, 60, 139–169.
- Belkin, W. B., & Croft, W. B. (1987). Retrieval Techniques.. In *Annual Review of Information Science and Technology*, Vol. 22, pp. 109–145.
- Efthimiadis, E. (1996). Query Expansion. *Annual Review of Information Science and Technology*, 31, 121–187.
- Efthimiadis, E. (2000). Interactive Query Expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11), 989–1003.

- Efthimiadis, E., & Robertson, S. (1989). *Perspectives in Information Management*, chap. Feedback and Interaction in Information Retrieval, pp. 257–272. Butterworths, London.
- Fidel, R. (1991a). Searchers' Selection of Search Keys: I. The Selection Routine. *Journal of the American Society for Information Science*, 42(7), 490–500.
- Fidel, R. (1991b). Searchers' Selection of Search Keys: II. Controlled Vocabulary of Free-Text Searching. *Journal of the American Society for Information Science*, 42(7), 501–514.
- Fidel, R. (1991c). Searchers' Selection of Search Keys: III. Searching Styles. *Journal of the American Society for Information Science*, 42(7), 515–527.
- Harper, D. J. (1980). *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. Ph.D. thesis, Jesus College, Cambridge, England.
- Harter, S. (1986). *Online Information Retrieval: Concepts, Principals, and Techniques*. Academic Press, Orlando, Florida.
- Hummingbird Communications Ltd., Toronto, Ontario, Canada (1999). *Fulcrum SearchSQL Reference Manual* (3.7e edition).
- Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36, 207–227.
- Jones, K. S., & Webster, C. A. (1980). Research on relevance weighting 1976-1979. British library report 5553, Cambridge.

- Lucas, W., & Topi, H. (2002). Form and Function: The Impact of Query Term and Operator Usage on Web Search Results. *Journal of the American Society for Information Science and Technology*, 53(2), 95–108.
- Robertson, S. E., & Jones, K. S. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3), 129–46.
- Rocchio, J. J. (1971). *The SMART Retrieval System*, chap. Relevance Feedback in Information Retrieval, pp. 313–323. Prentice Hall, Englewood Cliffs, NJ.
- Salton, G., & Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- Smeaton, A. F., & van Rijsbergen, C. J. (1983). The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System. *The Computer Journal*, 26(3), 239–246.
- Spink, A. (1994). Term Relevance Feedback and Query Expansion: Relation to Design. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pp. 81–90 Dublin, Ireland.
- Spink, A. (1995). Term Relevance Feedback and Mediated Database Searching: Implications for Information Retrieval Systems Practice and Systems Design. *Information Processing and Management*, 31(2), 1–11.
- Spink, A. (1997a). Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. *Journal of the American Society for Information Science*, 48(8), 741–761.

- Spink, A. (1997b). Study of Interactive Feedback during Mediated Information Retrieval. *Journal of the American Society for Information Science*, 48(5), 382–394.
- Spink, A. (2001). Regions and Levels: Measuring and Mapping Users' Relevance Judgments. *Journal of the American Society for Information Science*, 52(2), 161–173.
- Spink, A. (2002). Personal Communication..
- Su, L. T. (1999). Evaluation Measures for Interactive Information Retrieval. *Information Processing and Management*, 28(4), 503–516.
- van Rijsbergen, C. J., Harper, D., & Porter, M. F. (1981). The Selection of Good Search Terms. *Information Processing and Management*, 17, 77–91.

## A Final Questionnaire

1. *Were you familiar with some of the topics covered in these queries before you began this exercise?*

**user 1** No.

**user 2** I was familiar with about a third or more of the topics covered in the queries.

**user 3** Yes.

**user 4** I believe I have a good background in most of the given topics.

2. *Do you think that the document collection contained sufficient information for the queries you ran?*

**user 1** Yes.

**user 2** The collection of documents sometimes contained enough information and sometimes not. Many of the documents were repeated in other queries.

**user 3** For some of them, yes. For others, there was a limited number of documents that contained relevant information. (some queries contained no documents). For the most part, however, i feel that some good information was able to be obtained in the documents reviewed.

**user 4** In some cases, for example management, costs, and project information, the collection seemed more than adequate. However, for some topics, like net infiltration and water level contours, the information was more difficult to find and few documents were really relevant.

3. *Do you think that the documents were too long or too short? Was there too much information given or too little?*

**user 1** Too long, because of typical gov. (government) inability to state something briefly.

**user 2** The documents were not necessarily too long, but some "documents" did contain multiple copies or more than one document. Overall, the length was not too bad.

**user 3** Some documents were way too long (i.e. some exceeded well over 300 pages). As for whether or not there was too information, again, that was query-specific. Some documents were entirely specific to the particular query, others had no more than a sentence or two of information that probably would not have been beneficial to the query.

**user 4** Some of the topics are very complicated and extensive investigations have taken place to corroborate or refute previously held ideas. I think that the DOE and its' subsidiary contractors are diligent in cross-checking the documents for accuracy and clarity and we, as researchers, must simply work with what we receive.

4. *Were the first few top retrieved documents most relevant to the queries?*

**user 1** No.

**user 2** For the most part, yes, the top documents were most relevant.

**user 3** For the most part, yes.

**user 4** I found that the top few retrieved documents were not the most relevant documents for the query in most cases. On an average of twenty documents per query, it could take up to the sixth document to find the most relevant work. The first few documents may have mentioned the subject but were not usually the most relevant.

5. *In general, did the documents you received from the original query give you the information you were looking for? Were results better after the query was revised?*

**user 1** Yes, usually better after revised.

**user 2** The majority of the time documents from the original query contained the needed information. Overall, the results were not much better after revision.

**user 3** No, in general the original query produced the most relevant documents.

**user 4** In general, the retrieved documents for the original query contained the information I was looking for.

6. *Did you find more relevant information after expanding the queries using terms from the keyword list, title, abstract or document text?*

**user 1** Yes.

**user 2** The most relevant information was found using terms from the document text.

**user 3** In some cases, yes. In others, no.

**user 4** I found more relevant documents after expanding the queries using information I extracted originally to revise the list. I felt this was especially true if the keywords contained one or more words from the query itself especially if those words were found in the title. This technique seemed to narrow down the subject matter and give me other documents that were relevant.

7. *Do you think that documents with keywords were easier to understand and use than documents that didn't have keywords?*

**user 1** Yes.

**user 2** No, it was the same with or without keywords.

**user 3** Yes.

**user 4** For this particular exercise, I found the given keywords to be a detriment rather than a help. Sometimes, the keywords were misleading or just plain wrong. Another keyword problem that I encountered were misspelled words and duplicated words. A spelling error in a keyword renders it useless in aiding any research.

8. *Did you find the keywords helpful when they were available with the document?*

**user 1** Yes.

**user 2** Sometimes the keywords were relevant but mostly not.

**user 3** Yes.

**user 4** I believe that someone doing keywording of these documents should not be influenced in any way by someone else's work. If the keyword was wrong originally for the query, then the error may be perpetuated.

9. *Did you find more relevant documents using keywords you extracted from the document yourself or from keywords that were made available to you with the document?*

**user 1** Ones found myself.

**user 2** At times the words i chose from the text were the same as the keywords, but more relevant documents were found with keywords I chose.

**user 3** Combination of both.

**user 4** Keywords specific to the query that I extracted myself were more useful in finding relevant documents than keywords available with the document.

10. *What number of keywords seemed to be the most appropriate for bringing up the best set of relevant documents in the revised query?*

**user 1** 6

**user 2** There was no set number of keywords for best results. However, better results came from using the most specific keywords possible, which basically means using less keywords.

**user 3** In all honesty, i really didn't count.

**user 4** I found many relevant terms for each query but I think fifteen to twenty words, with an attempt to incorporate words in the title or abstract, returned the most significant results.

11. *What is your opinion of the level of the language in the document collection? Would someone need to have background in the nomenclature and subject matter to understand these documents?*

**user 1** A person would need very little background.

**user 2** The language in some documents is completely, scientific; the rest of the documents are understandable.

**user 3** Being a fairly technical subject, yes, i do think in order to truly understand the material in most of the documents, you would need to have some kind of technical, scientific background to get the most out of the documents.

**user 4** The Yucca Mountain Project documents that I viewed are very technical and an understanding of the basic concepts is essential, in my opinion. A majority of the documents are not written for the lay person but for colleagues in their field.

12. *Would a domain specific thesaurus be helpful?*

**user 1** Yes.

**user 2** For certain readers, yes, a domain specific thesaurus would be helpful.

**user 3** Yes.

**user 4** If a researcher needs a thesaurus for the topic, then a lack of background on the reader's part is assumed and a thesaurus may or may not help to qualify them to do the job.

13. *What did you learn from the documents that you retrieved?*

**user 1** Just about everything about the yucca mountain project and the selection process.

**user 2** I learned quite a lot about the yucca mountain project and geology in general.

**user 3** Several pieces of scientific information about yucca mountain including characteristics such as hydrogeology, lithology, rock characteristics, types of tests performed at the site, and much more.

**user 4** I think my biggest eye-opener about Yucca Mountain is how close they are to being in full operation.

14. *After reviewing the documents returned for these queries, do you think your knowledge of the subject has increased?*

**user 1** Yes!

**user 2** My knowledge increased reviewing the documents for each query while looking for keywords.

**user 3** In general, yes.

**user 4** I feel as if I just had a refresher course in all the science classes I have ever taken and I was pleased to find that my previous work experience was useful. My knowledge base increased and certain ideas and formulations really started to jell.

15. *Do you think that viewing the title was enough to determine if the document was relevant to the query? If not, what other content of the document would you like to have displayed in the search results?*

**user 1** No, the table of contents were very helpful.

**user 2** No, the title was not enough to determine if the document was relevant. If the table of contents is short enough, it can be displayed making searching easier.

**user 3** No, the title was not sufficient. at the bare minimum, i think you can get a pretty good idea from the table of contents (provided there is one) and from there, narrow your search to see if it is relevant.

**user 4** The title was rarely adequate to portray the contents of the document. A code differentiating the type of document might be useful. Is it a report (R), a study plan (SP), a foundation document (SCP, SAR, EIS), published by the Yucca Mountain Project Office (YMPO), draft (D), revision (R), final (F), briefing to congress (Br), public comments (PC), expert opinions (EO), etc.

16. *Did you perceive any difficulties with the methods of searching available to you? What recommendations do you have for improving the search process?*

**user 1** No, none.

**user 2** Yes, it seemed that when keyword phrases were entered the search engine looked for each word separately not in conjunct. This is why, i think, the retrieved documents were most of the time irrelevant. I also think that it should not matter from where the keyword is chosen; it seems like the search engine searches for them where we found them.

**user 3** No.

**user 4** The main difficulty with the documents, in my opinion, is the lack of correlation between the document "actual" page numbers and the number of the page as counted by the browser. This completely throws off the document table of contents and makes it very difficult to go to the exact page required.