Rutgers Interactive Track at TREC-5

N.J. Belkin* , A. Cabezas, C. Cool, K. Kim, K.B. Ng,
S. Park, R. Pressman, S. Rieh, P. Savage, H. Xie

School of Communication, Information & Library Studies
Rutgers University
4 Huntington Street
New Brunswick, NJ 08901-1071
belkin@scils.rutgers.edu

Abstract

The Interactive Track investigation at Rutgers concentrated primarily on three factors: the searchers' uses and understandings of relevance feedback and ranked output, and the utility of relevance feedback for the interactive track task; the searchers' understandings of the interactive track task; and performance differences based on topic characteristics and searcher and order effects. Our official results are for twelve searchers, each of whom did searches on six different topics.

1. Introduction

The Rutgers TREC-5 Interactive Track group had some substantial difficulties in performing the investigation that we originally wanted to do. We had originally intended to compare a system with relevance feedback based on positively-judged documents only, with a system which used both positive and negative judgements in the feedback process. For both systems, following the results of Koenemann (1996), we intended to have user control over which terms were added to the query through the relevance feedback process. For a variety of reasons, we were unable to construct the system which could take account of negative relevance judgements. Our fall-back position was then to test user controlled feedback versus ordinary relevance feedback. For this, we ran into incompatibility snags between the interface structure that we had available for doing this, and the new version of the search system that we wanted to use. Our next, and final fallback position, then, was to use our existing TREC-4 system, in order to investigate:

- the searchers' understanding of the TREC-5 interactive track task;
- the understanding, use and utility of positive-only relevance feedback for this task;
- effects of topic order, difficulty and domain on performance; and,
- the range of performance by different searchers on the same topics.

Because we got to out final fall-back position only just before our experiment had to begin, we did not have time to modify our existing system (INQUERY 2.1p3) to index two of the new databases which were added for the adhoc task this year whose format differed somewhat from previous years (Congressional Record and Federal Register 1994). Therefore, the results that we report here can only be considered partial, since they exclude any access to these two databases.

Our excuses having been made, we can now get to what we *were* able to do, and how we did it.

---

* To whom all correspondence concerning this paper should be addressed

2. Methods

For a description of the system which we used for TREC-5 (INQUERY 2.1p3, with a local interface, called RU-INQUERY) see Belkin, *et al.* (1996), the report of our TREC-4 study.

We attempted to follow the guidelines for the TREC-5 interactive track quite strictly, to the extent that this was possible. Those guidelines call for each searcher to perform three searches in a control condition, which was to be a version of PRISE, and another three searches in the test condition of the local study. The search topics were organized in four blocks of three searches each, and for each searcher, the order of blocks was specified in a permuted design (*i.e.* B(block) 1 then B2; B2 then B3; B3 then B4; B4 then B1), leading to a minimum of four searchers to complete the design. Unfortunately (sorry, excuses again) we were unable to get PRISE running in time for the beginning of the study, and so the control versus test aspect of the study was no longer relevant. However, since we were interested topic order effects, we maintained this general design, and duplicated the entire sequence three times (*i.e.* using twelve searchers).

We recruited fourteen volunteer subjects to take part in the study, from the community of information professionals in New Jersey, and from the students and faculty of the School of Communication, Information and Library Studies and of the Computer Science Department at Rutgers University. Results from the first two subjects are not reported, however, because they used a slightly different version of the tutorial than the other twelve searchers. Details of the subject characteristics are in the Rutgers Interactive Track "Rich Format" description.

The experimental sessions were held at our lab at Rutgers University, and took about 3 1/4 hours total time. On arrival, the subjects were administered a questionnaire concerned with demographic and educational factors, and previous experience with IR systems. We then applied a structured interview, aimed at identifying their understanding of the TREC-5 interactive track task, and at how they would go about performing this task in a system with which they were familiar. They then did a hands-on structured tutorial in the use of RU-INQUERY, which incorporated an example of the interactive track task. Then, for each search topic, the subjects were handed a sheet of paper with the general instructions for the task, and the specific instructions for that topic, and were told that they had twenty minutes to perform the task for that topic. They were also given, at this time, a separate sheet on which they were asked to list the "aspects" of the topic as they identified them. At the same time, the subjects were instructed to "think aloud" during the search, and this talk was recorded on a videotape of the monitor during the search. In addition, the entire search interaction was logged. Having finished the search, the subjects were asked to complete a brief questionnaire about that search, giving self-reports on familiarity with topic, difficulty of search, satisfaction with search results, confidence in aspect identification, and time enough to do an effective search. This process was repeated for the next two searches in the block, after which the subjects were asked if they wanted to take a break. After the break, the same process was repeated for the three searches of the second block. After all six searches were completed, the subjects were administered an exit questionnaire and interview, whose foci were: the understanding and use of relevance feedback in their searches, and its utility for the interactive track task; and their understanding and experience of the task itself. All of the data-collection instruments are shown in Appendix I.

3. Results

The basic performance result for the interactive task is the so-called *aspect recall*. The task itself was to identify as many aspects of the specific topic as possible, on the basis of the literature, saving at least one document which discussed one or more of the aspects. The basic result that was returned to TREC then, was the list of saved documents, in the order saved, plus the total time for the search (as measured from the time the subjects were given the topics, until they said they were finished). The TREC assessors, bless them, were asked to consider the documents which had

Search	Searcher	Time	Saved	Prec	Recall
254I-003-1	c	1095	12	75	78
254I-005-2	e	1260	13	15	44
254I-007-1	g	1075	3	67	33
254I-011-1	k	1108	9	78	44
254I-012-2	l	1246	7	71	44
254I-014-2	n	1043	7	71	44
254 mean		1137.8	8.5	62.8	47.8
255I-003-2	c	1141	16	38	81
255I-004-1	d	1280	1	100	15
255I-006-2	f	1214	5	20	19
255I-008-1	h	1242	2	100	23
255I-009-2	i	1330	2	100	19
255I-012-1	l	1289	6	50	15
255 mean		1249.3	5.3	68	28.7
256I-004-2	d	1222	5	20	14
256I-005-1	e	1132	3	0	0
256I-007-2	g	1013	10	40	71
256I-009-1	I	1163	2	50	29
256I-010-2	j	617	11	9	14
256I-013-1	m	1092	2	50	43
256 mean		1039.8	5.5	28.2	28.5
258I-003-2	c	1182	28	64	79
258I-004-1		1138	7	29	38
258I-006-2	f	1225	7	43	38
258I-008-1	h	1094	17	53	67
258I-009-2	i	1231	5	80	54
258I-012-1	l	1214	10	90	67
258 mean		1180.7	12.3	59.8	57.2
260I-006-1	f	1333	1	100	50
260I-008-2	h	1211	8	62	83
260I-010-1	j	1195	12	17	50
260I-011-2	k	1178	2	50	50
260I-013-2	m	751	3	33	33
260I-014-1	n	1304	0	0	0
260 mean		1162	4.3	43.7	44.3
264I-006-1	f	1217	3	67	6
264I-008-2	h	1182	11	100	29
264I-010-1	j	1344	15	87	47
264I-011-2	k	1127	13	100	47
264I-013-2	m	830	3	100	18
264I-014-1	n	1298	15	100	71
264 mean		1166.3	10	92.3	36.3

Search	Searcher	Time	Saved	Prec	Recall
274I-004-2	d	1047	8	100	64
274I-005-1	e	1385	6	83	73
274I-007-2	g	712	14	100	73
274I-009-1	i	1247	5	100	73
274I-010-2	j	1243	36	94	100
274I-013-1	m	1191	6	100	64
274 mean		1137.5	12.5	96.2	74.5
284I-003-1	c	1188	22	50	52
284I-005-2	e	1157	13	69	44
284I-007-1	g	1015	9	44	24
284I-011-1	k	1018	17	41	44
284I-012-2	l	1201	18	67	56
284I-014-2	n	818	10	80	36
284 mean		1066.2	14.8	58.5	42.7
286I-006-1	f	1226	4	50	44
286I-008-2	h	1121	14	7	44
286I-010-1	j	1013	9	44	56
286I-011-2	k	1251	11	82	56
286I-013-2	m	590	5	60	44
286I-014-1	n	1087	15	73	56
286 mean		1048	9.7	52.7	50
292I-003-1	c	1067	19	21	56
292I-005-2	e	1209	22	36	41
292I-007-1	g	1254	16	31	47
292I-011-1	k	1213	19	47	47
292I-012-2	l	1212	22	32	44
292I-014-2	n	945	11	45	38
292 mean		1150	18.2	35.3	45.5
293I-004-2	d	1357	0	0	0
293I-005-1	e	1261	11	45	100
293I-007-2	g	1337	1	100	17
293I-009-1	i	1136	5	20	17
293I-010-2	j	964	6	0	0
293I-013-1	m	1088	2	50	17
293 mean		1190.5	4.2	35.8	25.2
299I-003-2	c	1195	18	39	73
299I-004-1	d	1251	3	33	47
299I-006-2	f	1337	6	7	7
299I-008-1	h	1140	9	44	87
299I-009-2	i	1236	2	50	40
299I-012-1	l	1157	7	57	60
299 mean		1219.3	7.5	38.3	52.3
Overall		1180.3	9.4	56	43.8

Table 1. Performance data by topic and searcher.

already been judged relevant to the topic in the adhoc task relevance assessment, as well as those retrieved by the interactive track participants, and to identify all of the aspects of the topic, and the documents in which each aspect was discussed. Then, aspect recall was computed by comparing the list of saved documents for each search with the list of document-aspect tuples identified by the assessors. If documents were saved which, *in toto*, discussed all of the aspects, aspect recall was 100. No penalty, nor advantage, was assigned for multiple identification of the same aspect. Precision was computed in the normal manner: as the percentage of saved documents which treated at least one aspect, *i.e.* were relevant.

Table 1 shows the results for all six searches for each topic, with mean values for each topic and for the study overall.

Other analyses of the data will be presented at the meeting, as will some thoughts that they will have engendered about the nature of the interactive track task, the evaluation methodologies, and what one can hope to learn from this type of evaluation study.

4. References

Belkin, N.J., Cool, C., Koenemann, J., Ng, K.B., Park, S. (1996) Using relevance feedback and ranking in interactive searching. In: D. Harman, ed. *TREC-4. Proceedings of the Fourth Text Retrieval Conference*. Washington, D.C., GPO: in press.

Koenemann, J. (1996) *Relevance feedback: Usage, usability, utility*. Ph.D. Dissertation, Graduate Program in Psychology, Rutgers University, New Brunswick, NJ

Appendix I. Experimental Materials

I.1 Entry questionnaire

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY
SEARCHER QUESTIONNAIRE

Please list all the college/university degrees that you have (or expect to have):

Degree	Major	Date
Degree	Major	Date
Degree	Major	Date
Degree	Major	Date

What is your age?

❑ Under 21	❑ 31-40	❑ 51-60
❑ 21-30	❑ 41-50	❑ Over 60

What is your gender?

- ❑ Female
- ❑ Male

Please circle the appropriate number...

How much experience have you had...	None		Some		A great de
searching on computerized library catalogs	1	2	3	4	5
searching on CD ROM systems, e.g., Infotrac, Grolier	1	2	3	4	5
searching on commercial online systems, e.g., Dialog, Lexis, BRS Afterdark	1	2	3	4	5
searching on world wide web browsers, e.g., Mosaic, Netscape	1	2	3	4	5
searching on other systems, please specify the system: ___________________________	1	2	3	4	5
searching full-text databases	1	2	3	4	5
searching in ranked-output information retrieval systems	1	2	3	4	5
searching in information retrieval systems that provide automatic relevance feedback	1	2	3	4	5
using a mouse-based interface	1	2	3	4	5

Overall, for how many years have you been doing online searching? __________ years

Who have you performed searches for?
- ❏ Yourself only
- ❏ Others only (e.g., as an intermediary)
- ❏ Yourself and others

Have you participated in previous TREC Searching Studies?
- ❏ Yes
- ❏ No

I.2. Pre-search interview

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY
PRE-SEARCH INTERVIEW

In order to understand the different searching experiences of our participants, we would like you to answer a couple of questions about the methods that you typically use when you do online searching. When you answer these questions, please try to give us as much detail as you can. Please use this worksheet for any notes that you wish to make. *(Hand worksheet to searcher.)*

Imagine that you are interested in learning about the different alternative sources of energy for automobiles. You decide to investigate the literature by using a computerized database of newspaper articles that is available for your use. Since you are interested in identifying as many "aspects" of this topic as possible, you will want to identify each one of the different alternative sources of energy for automobiles, including gasoline additives that decrease pollution or reduce oil consumption.

1. What steps would you take in order to perform a search that would identify as many "aspects" as possible for this topic? Describe your overall approach by listing what would do first, and then describe each of the steps that you would follow after that.

2. How would you decide that your search is finished?

I.3. Search evaluation form

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY
SEARCH EVALUATION FORM

TOPIC NUMBER _____________

Please answer the following questions, as they relate to this specific topic.

Please circle the appropriate number...

To what extent...	Not at all		Marginally		Extremely
are you familiar with this topic?	1	2	3	4	5
was it difficult to do this search?	1	2	3	4	5
are you satisfied with your search results?	1	2	3	4	5
are you confident that you identified all the possible aspects for this topic?	1	2	3	4	5
did you have enough time to do an effective search?	1	2	3	4	5

I.4. Topic description and task instructions

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY

TOPIC DESCRIPTION

Topic 256i
Negative reactions to reduced requirements for college undergraduate core studies

GENERAL INSTRUCTIONS

Now we would like you to identify as many aspects as possible for each topic that will be presented to you. You will be given 20 minutes to search for each topic's aspects. Please save one document for each of the aspects that you identify. If you save one document that contains many aspects, try not to save additional documents that contain only those aspects, unless a document contains additional aspects as well.

Carefully read each description and narrative for each topic because the interpretation of "aspects" changes from topic to topic. For example, aspects can refer to different developments in a field, to different instances in which an event can occur, or to different kinds of treatments -- as it did in the high blood pressure example.

SPECIFIC INSTRUCTIONS

<u>Topic</u>
Colleges for a long time have been reducing their requirements in such core subjects as history, literature, philosophy, and science. Criticism of this trend has occurred.

Aspects
Please save at least one document that identifies EACH DIFFERENT criticism of this trend. If one document discusses several criticisms, then you need not save other documents that repeat those aspects, since your goal is to identify the different criticisms that have been made.

Narrative
To be relevant, a document will provide negative opinions/facts concerning the fact that colleges have reduced their basic requirements for the granting of degrees to undergraduates.

I.5. Aspect identification worksheet

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY
SEARCHER WORKSHEET

Searcher # __________

Please use this sheet of paper to write down any notes that you'd like to make during your search and to list the aspects as you identify them for this topic.

Topic 254i:

I.6. Post-search interview

Searcher #_________

RUTGERS TREC-5 INTERACTIVE SEARCHING STUDY
POST-SEARCH INTERVIEW

So we can have a better understanding of your overall searching experience, I'd like to ask you some final questions about your experiences today. In order to answer the first set of questions, I'd like you to use this scale (*hand scale sheet to participant*). In this scale a "1" means "not at all", a "3" means "marginally", and a "5" means "to a great extent".

1. To what extent did you understand how to use Relevance Feedback?

Not at all		Marginally		To a great extent
1	2	3	4	5

Why is that?

2. To what extent did you use Relevance Feedback during your searches?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: Why didn't you use it more?
rating = 4 or 5: Why did you use it so much?

3. To what extent did you find Relevance Feedback useful during your searches?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: Why wasn't it useful? What would have made it more useful?
rating = 4 or 5: Why did you find it useful?

4. To what extent did Relevance Feedback improve your ability to identify different aspects of the topics?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: What would have made it more useful to identify different topic aspects?
rating = 4 or 5: Why did you find it useful in improving your ability to identify different topic aspects?

5. To what extent was it helpful to have Ranked Output in your searches?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: What would have made Ranked Output more helpful?
rating = 4 or 5: Why did you find Ranked so helpful?

6. To what extent did you find this task different from other searching tasks that you typically perform?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: In what ways was this task different from other searching tasks that you typically perform?
rating = 4 or 5: In what ways was this task similiar to other searching tasks that you typically perform?

7. To what extent were you able to understand the nature of the task?

Not at all		**Marginally**		**To a great extent**
1	2	3	4	5

rating = 1, 2 or 3: What did you find confusing?
rating = 4 or 5: Why did you find this easy to understand?

8. Do you have any other comments about your experiences with RU-INQUERY?