RELEVANCE FEEDBACK IN CHESHIRE

by
Yasar Tonta

(A term paper prepared for Professor Maron's *course on* ***"Advanced Information Retrieval Systems"*** *(L206))*

Berkeley
December 1990

CONTENTS

Introduction to Relevance Feedback

In an article published in 1977, Swanson examined the well-known information retrieval experiments and the measures used therein. He suggested that the design of information retrieval systems "should facilitate the trial-and-error process itself, as a means of enhancing the correctability of the request" (Swanson, 1977: 142).

Van Rijsbergen (1979) shares the same view when he points out that a user confronted with an automatic retrieval system is likely to have difficulties in expressing his/her information need in one go. "He is more likely to want to indulge in a trial-and-error process in which he formulates his query in the light of what the system can tell him about his query. The kind of information that he is likely to want to use for the reformulation of his query is:

(1) the frequency of occurrence in the data base of his search terms;

(2) the number of documents likely to be retrieved by his query;

(3) alternative and related terms to be the ones used in his search;

(4) a small sample of the citations likely to be retrieved, and;

(5) the terms used to index the citations in (4)" (Van Rijsbergen, 1979: 105).

"Relevance feedback" is one of the tools that facilitates the trial-and-error process by allowing the user to interactively modify his/her query based on the search results obtained during the initial run. The following quotation summarizes the relevance feedback process very well:

> "It is well known that the original query formulation process is not transparent to most information system users. In particular, without detailed knowledge of the collection make-up, and of the retrieval environment, most users find it difficult to formulate information queries that are well designed for retrieval purposes. This suggests that the first retrieval operation should be conducted with a tentative, initial query formulation, and should be treated as a trial run only, designed to retrieve a few useful items from a given collection. These initially retrieved items could then be examined for relevance, and new improved query formulations could be constructed in the hope of retrieving additional useful items during subsequent search operations" (Salton and Buckley, 1990: 288).

Relevance feedback was first introduced over 20 years ago during SMART information retrieval experiments. Earlier relevance feedback experiments were performed on small collections (e.g., 200 documents) where the retrieval performance was unusually high (Rocchio, 1971; Salton, 1971; Ide, 1971). It was shown that the relevance feedback has improved the retrieval performance markedly. Recently Salton and Buckley (1990) examined both "vector processing" and "probabilistic" relevance feedback methods, and evaluated twelve different feedback methods "by using six document collections in various subject areas for experimental purposes." The collection sizes they used varied from 1,400 to 12,600 documents. The relevance feedback methods produced improvements in retrieval performance ranging from 47% to 160%.

The relevance feedback studies performed in the past will not be reviewed in this paper. Nor will the different relevance feedback formulae and the methods used therein. More detailed information on relevance feedback formulae can be found in Salton and Buckley (1990). For mathematical explications of relevance feedback process, see Rocchio (1971); Ide (1971); and, Salton and Buckley (1990).

It should be noted that relevance feedback methods have yet to be tried on large scale operational information retrieval systems.

CHESHIRE Experimental Online Catalog

The CHESHIRE (California Hybrid Extended SMART for Hypertext and Information Retrieval Experimentation) experimental online catalog system is "designed to accommodate information retrieval techniques that go beyond simple keyword matching and Boolean retrieval to incorporate methods derived from information retrieval research and hypertext experiments" (Larson, 1989: 130). The test database for the CHESHIRE system consists of some 30,000 MARC records representing the holdings of the Library of the School of Library and Information Studies in the University of California at Berkeley. CHESHIRE uses a modified version of Salton's SMART system as the "retrieval engine" and index

manager and it runs on a Sun workstation with 320 megabytes of disk storage. Larson (1989) provides a more detailed information about CHESHIRE and the characteristics of the collection.

The CHESHIRE system uses the "classification clustering" technique which is based on the presence of identical LC classification numbers and which brings documents with the same LC classification number together along with the most frequently used LC subject headings in a particular cluster. At present, some 8400 classification clusters have been created for the above collection.

CHESHIRE accommodates queries in natural language form. The user describes his/her information need using words taken from the natural language and submits this statement to the system. First, a retrieval function within the system analyzes the query, eliminates the "buzz" words (using a stop list), processes the query using the stemming and indexing routines and weights the terms in the query to produce a vector representation of the query. Second, the system compares the query representation with each of the some 8400+ document cluster representations in order "to retrieve and rank the cluster records by their probabilistic "score" based on the term weights stored in the inverted file... The ranked clusters are then displayed to the user in the form of a textual description of the classification area (derived from the LCC summary schedule) along with several of the most frequently assigned subject headings within the cluster" (Larson, 1990b: 17). (For the theoretical basis of, and the probabilistic retrieval models used in, CHESHIRE online catalog system, see Larson (1990a).)

Once the system finds the "would-be" relevant clusters the user then will be able to judge some of the clusters as being relevant by simply identifying the relevant clusters on the screen. "After one or more clusters have been selected, the system reformulates the user's query to include class numbers for the selected clusters, and retrieves and ranks the individual MARC records based on this expanded query" (Larson, 1990b: 17).

Larson (1990b: 17) describes how it is that this tentative relevance information for the selected clusters can be utilized for ranking the individual MARC records:

> "In the second stage of retrieval in CHESHIRE, we still have no information about the relevance of individual documents, only the tentative relevance information provided by cluster selection. In this search, the class numbers assigned to the selected clusters

> are added to the other terms used in the first-stage query. The individual documents are ranked in decreasing order of document relevance weight calculated, using both the original query terms and the selected class numbers, and their associated MARC records are retrieved, formatted, and displayed in this rank order... In general documents from the selected classes will tend to be promoted over all others in the ranking. However, a document with very high index term weights that is not from one of the selected classes can appear in the rankings ahead of documents from that class that have fewer terms in common with the query."

Although the identification of relevant clusters can be thought of, quite rightly so, a type of relevance feedback, we rather consider it as some sort of system help before the user's query is run on the entire database.

After all of the above re-weighting and ranking processes, which are based on the user's original query and the selection of relevant clusters, are done, CHESHIRE will eventually come up with individual MARC records. This time the user is able to judge each individual record (rather than the cluster) that is retrieved as being relevant or nonrelevant. He/she can examine several records by making relevance judgments along the way for each record until he/she thinks that there is no use to continue displaying records as the probability of relevance gets smaller and smaller. The user's relevance judgments in this stage are recorded. When the user decides to quit because he/she is either satisfied (or frustrated) with the documents he/she has seen in the first retrieval, the system asks the user if he/she wants to perform a relevance feedback search. If the user decides to perform a relevance feedback search, then the system further revises and modifies the original query based on the documents the user has already judged relevant or nonrelevant in the previous stage, and comes up with more documents. The relevance feedback search can be iterated as many times as the user desires until he/she is satisfied with the search results.

Relevance Feedback Experiments on CHESHIRE

Although planned, no relevance feedback evaluation studies have been performed on CHESHIRE yet. The present one is a modest attempt to measure the retrieval effectiveness of the relevance feedback

search feature available in CHESHIRE.

Methodology

The retrieval effectiveness of the relevance feedback process in CHESHIRE has been tested on two sets of queries. The first set consists of 10 queries (Figure 1) and they were first used to test some advanced information retrieval techniques on CHESHIRE experimental online catalog by Larson (1990a). The second set consists of 11 queries (Figure 2) that were produced specifically for this experiment. Queries in both sets are in natural language form. Using two sets of queries created by two different persons allows one to test if there is any statistically significant difference between the two sets in regards to, say, the query types and specificity or exhaustivity of the queries.

It should be noted that the test queries used in this research have not engendered from real information needs of different users; rather, they were made up by the researchers themselves who acted as different users submitting hypothetical queries to the CHESHIRE system.

The search and evaluation procedure for each query has proceeded as follows:

The searcher would enter the query to the system in natural language form and instruct CHESHIRE to perform a search for that particular query. As was summarized in the previous section, in the first step CHESHIRE would bring classification clusters that were deemed, by the system, similar to the search terms and would satisfy the query, if judged relevant by the user, by retrieving individual documents contained in that particular cluster (or clusters). The searcher would judge each cluster as being relevant or nonrelevant to the query by pushing a select button. In this experiment, the number of clusters that should be judged before the initial retrieval has been limited to 20, although it is possible to perform an initial search by judging as few as just one or two clusters. The system, then, would revise the initial query by adding some components (i.e., class numbers) while re-weighting the search terms. This step basically enables the system to "understand" the user's query better: the documents that are similar to the query are rewarded by being assigned higher ranks, while dissimilar

documents are pushed farther down in the ranking.

After the query has been revised automatically by the system based on the feedback from clustering procedure, the individual documents are brought to the searcher's attention one after another. Again, for each document the searcher would make a relevance judgment by checking the bibliographic information (i.e, author, title, subject headings, etc.). The relevance/nonrelevance judgment for each document that the searcher has seen has been recorded on the paper so that the results can later be tallied and the precision ratio can be calculated for each query. The number of documents for which the relevance judgments were made at this stage has also been limited to 20. (It should be noted that the system retrieved different editions (or copies) of some books for some queries. In this case, we judged all editions as relevant (or nonrelevant for that matter). It is highly unlikely that this would effect the outcome of our experiment.)

Based on the relevance judgment for each individual document, the system, once more, revises the original query and tries to retrieve more relevant documents. The feedback gathered through the relevance judgments for 20 documents is called relevance feedback, and it is the effect of the relevance feedback process on the overall retrieval performance that has been tested in this experiment.

Once again, the searcher would see the new documents, one after another, that were retrieved from the database as the result of the relevance feedback process, and judge them as being relevant or nonrelevant to the query. The relevance judgment for each document that is retrieved for a given query has been recorded for further analysis and for computation of the precision ratios. The number of documents for which the relevance judgments has been made was limited to, once again, 20.

It would seem that the necessity of relevance judgments at each step prolongs the overall retrieval time: first, the searcher should judge the classification clusters; second, the individual documents that are retrieved based on the cluster feedback; and, finally, the documents that are retrieved in the relevance feedback stage. This was certainly the case in this experiment. The searcher was to go through more than 60 screenful of catalog data (20 for clusters, 20 for the first retrieval, and

20 for the relevance feedback retrieval) in order to come up with a precision ratio for a given query. Each query took approximately 20-25 minutes to run on CHESHIRE. It should be noted that the response time in CHESHIRE is quite fast (4-5 seconds). In actuality, most of the time reported above has been spent for relevance judgments. As mentioned before, although the searcher does not have to judge all the retrieved clusters/documents, we felt that judging the same number of documents for each query could provide some kind of consistency in relevance judgments. Furthermore, we wanted to see if the retrieval effectiveness changes as the searcher scans documents farther down the ranked list.

Precision and recall are the most commonly used retrieval effectiveness measures in information retrieval research. Precision is defined as the proportion of the retrieved documents that are relevant. Recall is the proportion of the relevant documents that are retrieved. Precision is relatively easier to calculate than recall. In large databases it is impossibly impractical to find the recall value for each query: one has to sequentially scan the whole database in order to find the total number of documents that are relevant for a given query. In view of the lack of time and the size of the CHESHIRE database (30,000+ documents), the recall ratios have not been calculated in this experiment. Only the precision ratios have been used to measure the retrieval effectiveness in CHESHIRE. (We could have, at least, provided the recall values for Larson's test queries as he obtained the precision and recall values at certain cutoff levels in an earlier research. However, such a comparison would have been misleading as the relevance judgments made by different persons are most likely to be different.)

After obtaining relevance judgments for all 21 queries, the tallies have been counted and the precision ratio for each query has been calculated at certain cutoff levels (5, 10, 15, and 20 documents).

Table 1 and Table 2 give the precision ratios for two sets of queries used in this experiment: for Larson's queries and for our queries, respectively. The figure in each cell represents the precision ratio for a given query at a certain cutoff level. It is simply the ratio of the number of documents

retrieved that are relevant to the number of documents retrieved.

The precision ratios on the left hand-side of each table are for the initial retrieval results while the ones on the right-hand side are for relevance feedback retrieval results. The average precision values at each cutoff level for all queries in each set have also been calculated along with their standard deviations.

The two-sample *z-test* has been carried out so as to find out if there was any difference between Larson's test queries and those produced by me. The two-sample *z-test* has been applied in four cutoff levels (i.e., for 5, 10, 15, 20 documents retrieved) for both initial retrieval and retrieval with relevance feedback. This test allows us to compare the two samples of the test queries used in the research, the assumption being that the test queries produced by myself and by Larson are not significantly different from each other.

The formula for the two-sample *z-test* is:

z= (1st average - 2nd average)/ standard error for difference

Here the first average is the average precision ratio obtained for my test queries, and the second average is that which was obtained for Larson's test queries. Standard error (SE) for difference can be found using the square root law (i.e., taking the square root of the sum of the squares of standard deviations).

Findings

The major findings obtained in this study are given below.

As can be seen from the Table 1, the average precision values at certain cutoff levels that were obtained during the initial retrieval stage for Larson's 10 test queries are as follows:

at 5 documents: 74% (sd = 28%);

at 10 documents: 64% (sd = 29%);

at 15 documents: 60% (sd = 28%);

at 20 documents: 59% (sd = 27%).

The average precision values at the same cutoff levels for my test queries are as follows (see Table 2):

at 5 documents: 69% (sd = 23%);

at 10 documents: 53% (sd = 23%);

at 15 documents: 47% (sd = 25%);

at 20 documents: 45% (sd = 27%).

The first thing that draws immediate attention in regards to the above figures is that the precision values decrease (as one would expect) as the number of documents retrieved increases. For Larson's test queries the average precision values range from 74% (at 5 documents) to 59% (at 20 documents) while the precision values for my queries range from 69% (at 5 documents) to 45% (at 20 documents).

Although the average precision values for my queries are about 10% lower than those obtained for Larson's queries, the result of the two-sample *z-test* indicates that the difference is not statistically significant at 0.01 level. In other words, the types of queries created by Larson and myself appear to be similar.

The relatively large standard deviations in both cases indicate that the precision values for individual queries vary a great deal from query to query.

The main objective of this study is to compare the improvement, if there is any, in retrieval effectiveness between the initial retrieval and the retrieval based on relevance feedback. The following precision values after relevance feedback are obtained for Larson's test queries (see Table 1):

at 5 documents: 54% (sd = 40%);

at 10 documents: 54% (sd = 37%);

at 15 documents: 48% (sd = 33%);

at 20 documents: 47% (sd = 33%).

The average precision values after the relevance feedback at the same cutoff levels for my test queries are as follows (see Table 2):

at 5 documents: 32% (sd = 21%);

at 10 documents: 38% (sd = 22%);

at 15 documents: 36% (sd = 21%);

at 20 documents: 36% (sd = 22%).

Again, the difference between the precision values for Larson's test queries and those for my queries do not seem to be real. The two-sample *z-test* result shows that the difference is not statistically significant at 0.01 level.

Compared to the precision values obtained in the initial search, it would first seem that the precision values after relevance feedback have not actually improved. Rather, the precision values are much lower in the relevance feedback stage. Such a conclusion, however, would be misleading, as explained below.

Relevance feedback process helps in refining the original query and finding more relevant materials in the second try. The true advantage gained through the relevance feedback process can be measured in two different ways:

1) By changing the ranking of documents and moving the documents that are judged by the user as being relevant up in the ranking. With this method documents that have already been seen (and judged as being relevant) by the user will still be retrieved in the second trial, although they are somewhat ranked higher this time. "This occurs because the feedback query has been constructed

so as to resemble the previously obtained relevant items" (Salton and Buckley, 1990: 292). This effect is called "**ranking effect**" (Ide, 1971) and it is difficult to distinguish artificial ranking effect from the true feedback effect (Salton and Buckley, 1990). Note that the user may not want to see the documents second time because s/he has already seen them during the initial retrieval.

2) By eliminating the documents that have already been seen by the user in the first retrieval and "freezing" the document collection at this point for the second retrieval. In other words, documents that were judged as being relevant (or nonrelevant) during the initial retrieval will be excluded in the second retrieval, and the search will be repeated only on the frozen part of the collection (i.e., the rest of the collection from which user has seen no documents yet). This is called "**residual collection**" method and it "depresses the absolute performance level in terms of recall and precision, but maintains a correct relative difference between initial and feedback runs" (Salton and Buckley, 1990: 292).

The different relevance feedback formulae are basically based on the variations of these two methods. In this experiment, the feedback weight for an individual query term *i* was computed according to the following probabilistic relevance feedback formula:

$$\log\left(\frac{p_i\,(1 - q_i)}{q_i\,(1 - p_i)}\right)$$

where

$$p_i = \frac{rel\_ret + (freq \,/\, num\_doc)}{num\_rel + 1.0}$$

$$q_i = \frac{freq - rel\_ret + (freq \,/\, num\_doc)}{num\_doc - num\_rel + 1.0}$$

where

freq is the frequency of term *i* in the entire collection;

rel_ret is the number of relevant documents term *i* is in;

num_rel is the number of relevant documents that are retrieved;

num_doc is the number of documents.

This formula takes into account only the "feedback effect," not the artificial "ranking effect" (i.e., documents retrieved in the first are not included in the second run).

The findings of this experiment can now be interpreted in more precise terms. Although the precision dropped about 15% during the relevance feedback process, this drop actually does not represent a decrease. It is because the relevant documents that have been retrieved in the initial retrieval are not included in the second retrieval: improvement in retrieving more relevant documents in the second trial is based on the "feedback effect" only. In other words, although the precision values are somewhat lower than the first retrieval, the system was still able to retrieve relevant documents. As pointed out earlier, this method "depresses the absolute performance level in terms of recall and precision." Nevertheless, this should be seen as success rather than failure.

Precision values obtained in the relevance feedback stage exhibit some interesting findings. Large standard deviations indicate that the precision values vary greatly from query to query. In fact, for three queries no relevant documents were retrieved during the relevance feedback (so the precision was zero). This may be due to the fact that the success (or failure) in relevance feedback very much depends on the judgments during the initial retrieval. It could be that the evaluations did not make the desired effect on the rest of the documents because the formula depends on a couple of characteristics such as terms in titles and subject headings. When the terms are re-weighted for the second retrieval, this process may sometimes behave in unexpected ways.

It is also plausible that the density of relevant documents in the collection for these three queries was not as good as the others. It might be that the system was able to find the relevant ones in the initial retrieval. The second retrieval, on the other hand, might have failed to retrieve any relevant documents because there were not any left in the collection. In fact, we believe that most queries are quite specific in nature and there were not that many relevant documents in the collection. For

instance, the system was able to retrieve two relevant documents in the first run for the following query (#10 Table 2):

"open systems interconnection reference model: applications in libraries and information centers"

Among twenty documents retrieved in the first run, these two documents were the only ones that were relevant. No relevant documents were retrieved in the second run (hence the precision was zero). However, we believe that these were the only relevant documents in the entire collection for this query. (I have written a paper on this topic recently: there are only a few monographs on OSI Reference Model.)

Three queries have been canceled because the system was not able to come up with relevant documents in the first place. These queries were:

1. teaching programming languages in library schools
2. authority control in marc records
3. statistics for librarians and information professionals

The first two queries were canceled because the documents retrieved were not "on the topic." The first query retrieved many programming language textbooks but only one of them ("*ABC's of BASIC for Librarians*") seemed relevant, and it was retrieved as the 12th document in the first run. The same for the second query: most documents were about MARC in general. We are not aware of any specific titles on authority control in MARC records.

The third query that was canceled exhibits more interesting results. We know that there are a few titles on this topic in the collection. It seems that the system failed to analyze the query properly. The retrieved documents were mostly about library statistics!

Conclusion and Further Research

The effect of relevance feedback on retrieval performance has been tested on CHESHIRE using one

of the probabilistic relevance feedback formulae based on "residual collection" method. It was found that the relevance feedback process helped retrieve more relevant documents from the collection that were unknown to the searcher during the initial search. The average precision (at various cutoff levels) for retrieved documents during the relevance feedback process range from 32% to 54%.

The success of relevance feedback seems to be closely associated with the specificity/exhaustivity of the search queries. The queries used in this experiment were specific in nature. In other words, the relevant documents in the collection were not densely distributed. This resulted in zero retrievals for some queries during the relevance feedback retrieval process.

The effectiveness of the retrieval performance in CHESHIRE has been tested using only the precision ratio. The retrieval performance should also be tested using the recall measure as well. This, however, requires more time since identifying all the relevant documents in the collection for each query is a labor-intensive task.

Different relevance feedback formulae can be tested in CHESHIRE. This will allow us to compare the retrieval performances obtained through different formulae and choose the one that improves the performance most. A stack of relevance feedback formulae is already available in CHESHIRE and further tests can be performed in the future.

In conclusion, relevance feedback process provides users some flexibility in formulating search queries interactively. It also improves the retrieval performance by retrieving further relevant documents from the collection.

Bibliography

Buckley, Chris. (1987). *Implementation of the SMART Information Retrieval System.* Ithaca, N.Y.: Cornell University, Department of Computer Science.

Ide, E. (1971). "New Experiments in Relevance Feedback." in Salton, Gerard, ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N.J.: Prentice-Hall. pp. 337-354.

Larson, Ray R. (1989). "Managing Information Overload in Online Catalog Subject Searching." in *Managing Information and Technology: Proceedings of the 52nd Annual Meeting of the American Society for Information Science, Washington, D.C., October 30-November 2, 1989.* Edited by Jeffrey Katzer and Gregory B. Newby. Medford, N.J.: Learned Information. pp. 129-135.

Larson, Ray R. (1990a). "Evaluation of Advanced Retrieval Techniques in an Experimental Online Catalog." (Research paper).

Larson, Ray R. (1990b). "Classification Clustering, Probabilistic Information Retrieval and the Online Catalog." (Research paper).

Rocchio, Jr., J.J. (1971). "Relevance Feedback in Information Retrieval." in Salton, Gerard, ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N.J.: Prentice-Hall. pp.313-323.

Salton, G. (1971). "Relevance Feedback and the Optimization of Retrieval Effectiveness." in Salton, Gerard, ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N.J.: Prentice-Hall. pp.324-336.

Salton, Gerard, ed. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, N.J.: Prentice-Hall.

Salton, Gerard and Chris Buckley. (1990). "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* 41(4): 288-297.

Swanson, Don R. (1977). "Information Retrieval as a Trial-and-Error Process," *Library Quarterly* 47(2): 128-148.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* 2nd ed. London: Butterworths.

Appendix

TEST QUERIES (by Larson)

1. censoring and suppression of school books
2. subject searching in online catalog systems
3. thesaurus and indexing language design, construction and maintenance
4. early english printers including william caxton
5. database management systems for library automation
6. artificial intelligence and expert systems in information retrieval
7. using the colon classification for special library collections
8. public library services for children and young adults
9. information systems and services for medical and health care professionals
10. history of the Library of Congress

Average number of terms in test queries (excluding stop words) = 5.4 terms

Figure 1. Test Queries by Larson

TEST QUERIES (by Tonta)

1. role of linked systems project in bibliographic control
2. impact of technology on academic and research libraries
3. access to information in printed and computerized sources: comparative studies
4. principles of managing library collections
5. use of cd-rom's in reference services
6. microcomputer-based circulation control systems
7. book in renaissance
8. bibliotherapy
9. design and implementation of user interfaces in online catalogs
10. open systems interconnection reference model: applications in libraries and information centers
11. transliteration and transcription of foreign language materials

Average number of terms in test queries (excluding stop words) = 4.9 terms

Figure 2. Test Queries by Tonta

PRECISION RATIOS FOR TEST QUERIES USED IN LARSON'S RESEARCH

	First Retrieval (without relevance feedback)				Retrieval with Relevance Feedback			
	Number of documents retrieved				Number of documents retrieved			
Query #	5	10	15	20	5	10	15	20
1	1.00	0.60	0.47	0.55	0.00	0.00	0.00	0.00
2	0.40	0.20	0.20	0.25	0.20	0.20	0.20	0.15
3	1.00	1.00	0.93	0.85	0.40	0.50	0.53	0.50
4	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.95
5	1.00	0.90	0.73	0.65	1.00	1.00	0.87	0.80
6	0.40	0.20	0.13	0.10	0.40	0.40	0.40	0.40
7	1.00	0.80	0.67	0.60	0.00	0.00	0.00	0.00
8	0.40	0.30	0.40	0.35	0.40	0.50	0.40	0.40
9	0.40	0.70	0.73	0.80	1.00	0.90	0.87	0.85
10	0.80	0.70	0.73	0.70	1.00	0.90	0.63	0.65
Ave. Precision Ratio (for 10 queries)	0.74	0.64	0.60	0.59	0.54	0.54	0.48	0.47
Std Dev	0.28	0.29	0.28	0.27	0.40	0.37	0.33	0.33

Table 1. Precision Ratios for Larson's Test Queries

PRECISION RATIOS FOR TEST QUERIES USED IN THIS RESEARCH

	First Retrieval (without relevance feedback)				Retrieval with Relevance Feedback			
	Number of documents retrieved				Number of documents retrieved			
Query #	5	10	15	20	5	10	15	20
1	0.40	0.30	0.27	0.20	0.20	0.20	0.20	0.30
2	0.40	0.30	0.40	0.45	0.20	0.30	0.40	0.40
3	0.80	0.70	0.60	0.55	0.60	0.60	0.53	0.50
4	0.80	0.60	0.73	0.80	0.40	0.60	0.60	0.65
5	0.80	0.40	0.27	0.20	0.40	0.30	0.20	0.15
6	0.40	0.30	0.20	0.25	0.20	0.10	0.07	0.05
7	0.80	0.70	0.47	0.35	0.80	0.80	0.67	0.70
8	1.00	1.00	1.00	0.95	0.20	0.40	0.40	0.50
9	1.00	0.70	0.73	0.80	0.20	0.40	0.47	0.40
10	0.40	0.20	0.13	0.10	0.00	0.00	0.00	0.00
11	0.80	0.60	0.40	0.30	0.40	0.50	0.47	0.35
Ave. Precision Ratio (for 11 queries)	0.69	0.53	0.47	0.45	0.32	0.38	0.36	0.36
Std Dev	0.23	0.23	0.25	0.27	0.21	0.22	0.21	0.22

Table 2. Precision Ratios for Tonta's Test Queries