# Experiments in the Probabilistic Retrieval of Full Text Documents

*William. S. Cooper*
*Aitao Chen*

School of Library and Information Studies

*Fredric C. Gey*

U.C. Data Archive & Technical Assistance

University of California, Berkeley
wcooper@violet.berkeley.edu

**ABSTRACT**

**The experiments described here constitute a continuation of a research program whose object is to find probabilistically sound, yet simple and powerful, ways of combining search clues in full-text retrieval. The methodology investigated for ad hoc retrieval is that of logistic regression, in which the retrieval rule takes the form of a regression equation fitted to learning data. Most of the variables used in the regression take the form of means rather than the more customary sums, and it is argued that this is logically preferable. Radical manual reformulations of the topics were tried out and found to boost retrieval effectiveness. For routing retrieval, an approach based on the Assumption of Linked Dependence, involving the extraction of relevance-associated stems from feedback documents, is investigated. One characteristic of this approach is that only a very minimal use is made of the original topic formulation.**

## Introduction

The problem of full text retrieval system design can be thought of as a problem in the combination of statistical clues. The design objective is to achieve as high a level of retrieval effectiveness as possible, consistent with reasonable theoretical and computational simplicity. As its contribution to TREC-3, the Berkeley group has continued its efforts to develop a simple, statistically sound probabilistic basis for text retrieval.

Because the emphasis is on finding a good logic of combination, the statistical clues used are all based on simple conventional frequency counts. This is not because we have objections to thesauri, parsing, phrase discovery, disambiguation, and other natural language processing or AI-like approaches. To the contrary, we feel that it is a virtue of the regression procedures explored here that they are more hospitable than most to the incorporation of additional clues. However, our experience and that of others suggests that an astute use of simple stem and document frequency information immediately lifts one to a high plateau of effectiveness. Further measures to attain slightly higher levels can then be incorporated if they are deemed worth the trouble.

Three experiments were performed for TREC-3 by the Berkeley group. Two of these were experiments in retrospective ('ad hoc') retrieval in which the group continued its earlier explorations of logistic regression as a simple but powerful tool for combining statistical evidence. Specifically, the advantages of separating the 'quantitative' from the 'qualitative' variables in the regression are examined in the experiments to be described. In its single routing retrieval experiment, the group has extended its earlier approach of estimating directly the crucial probabilistic quantities. A possibly novel aspect of the approach taken is that the search topic -- that is, the original natural language question itself -- is essentially ignored in the formulation of the internal search request used by the retrieval algorithm. Instead of the usual 'query expansion', we practiced vir-

tual 'query defenestration', in which only the already-judged documents were used to formulate the internal search specification.

**The Ad Hoc Retrieval Rule**

Our first experiment in ad hoc retrieval for TREC-3 (coded Brkly6) used fully automatic query formulation, in which the natural language topics were stoplisted and stemmed but not otherwise changed except for doubling of title stems. The computational rule used to assign a probability of relevance to each document in the collection -- and thereby to impose a search order on the collection for the user -- is easily stated. The logodds of relevance of any given document to any given query is estimated by the equation

$$\log O(R \mid Q,D) \simeq c_0 + \sum_{i=1}^{6} c_i X_i \qquad (1)$$

where $R$ is the event that the document is in fact relevant to the query, $O(R \mid Q,D)$ is the odds that the event of relevance has occurred, and

$X_1 = \frac{1}{M}\sum_{j=1}^{M} \log QAF_{t_j}$, the logged absolute frequency of occurrence in the topic $Q$ of the match stems, averaged over all match stems. Thus if there are $M$ stems in common between the topic and the document, $X_1$ is calculated by noting how many times each such stem $t_j$ occurs in the query, taking the log of each of these $M$ quantities, summing the logs, and dividing by $M$;

$X_2 = \sqrt{QL}$ i.e. the square root of the total number of all (non-stopword) stem occurrences in the topic;

$X_3 = \frac{1}{M}\sum_{j=1}^{M} \log DAF_{t_j}$ the logged absolute frequency of occurrence in the document of the match stems $t_j$, averaged over all match stems.

$X_4 = \sqrt{DL}$, the square root of the document length (analogous to $X_2$);

$X_5 = \frac{1}{M}\sum_{j=1}^{M} \log IDF_{t_j}$ the logged inverse document frequency (IDF) of the match stems $t_j$, averaged over all match stems;

$X_6 = \log M$, i.e. the logarithm of the number of match stems;

and where the values of the coefficients in the equation are $c_0 = 0.679$, $c_4 = -0.0674$, $c_5 = 0.223$, and $c_6 = 2.01$. The logodds-of-relevance figures estimated by this formula and used to rank the output documents are of course easily transformed into probability-of-relevance estimates to be displayed with the output documents for the guidance of the user.

The values of the coefficients $c_0, \ldots, c_6$ in Equation (1) were obtained by fitting the equation to the empirical data by means of a logistic regression analysis (Hosmer & Lemeshow 1989). The empirical data used as the basis of this analysis consisted of the values of $X_1$ through $X_6$, and a human judgement of relevance or nonrelevance, for each topic-document pair in the TREC corpus of relevance judgements. Upon supplying this matrix of data to a standard statistical package capable of performing logistic regression (BLSS), the regression equation (1) was obtained.

Actually, so many relevance judgements were available that only a subset of the available data were used. This subsample was constructed by applying, for each topic, a crude preliminary retrieval rule to rank the set of all documents whose relevance to the topic had been judged. The top 433 judged documents were then selected for each topic as the documents whose associated data would be used in the regression analysis. It was felt that by choosing high-ranking documents for the analysis, the resulting regression equation could be made more sensitive to the appropriate ordering of the documents near the top of the output ranking where the output order matters most. The number 433 was chosen arbitrarily as the number that would yield the maximum number of data points (433×150≈65000) that the statistical package could handle.

Retrieval based on Equation (1) is readily accomplished by storing stem-specific partial sums from the right side of Eq. (1) as indexing weights on the query stems and document stems. Then, at retrieval time, each match stem's query weight is added to its document weight, the mean of these sums over all match stems is calculated, and $c_6 X_6$ is added in. The procedure is roughly comparable in computational efficiency to the operations required for simple vector processing retrieval models, and the system used for our

experiments was in fact a modified version of the SMART system.

**Separating Quality from Quantity**

In past regression experiments the Berkeley group, and so far as we know all other IR experimenters attempting a regression approach (e.g. Fuhr & Buckley 1993), have taken as their regression variables various statistics *totaled* over all match stems. For instance, the regression variable used to represent the inverse document frequency information about a topic-document pair might be the *sum* of the (logged) inverse document frequencies of the match terms. Although this seems to have worked well, we have come to believe that it is not the soundest choice of variable form, and that it might be possible to do better.

To see the problem with the use of totals, suppose that a retrieval clue has been scaled (linearly transformed) in such a way that zero and negative values of it can occur. For instance, if it has been transformed into a standardized statistical variable, zero might represent its mean value. Now, how a variable has been scaled should not matter, but under a policy of totaling over match terms it does. Suppose that for a given query, document A has just one match stem and its value for that variable is zero, while document B has several match stems and their values for the variable are also all zero. The total contribution for the variable is zero for both documents, yet B is likelier to be relevant to the topic than A simply because it has more match terms. In other words, information about number of match terms can be lost or distorted when totals are used.

The underlying difficulty is that use of totals confounds match stem *quality* with the number of match stems or match stem *quantity*. The corrective measure that has been taken by the Berkeley group is to calculate the values of the variables, not as totals, but as *means*, of their values for the different match terms. Then, to incorporate the information about the number of match terms, a separate new variable is introduced that deals with nothing but the match term count. This policy overcomes difficulties of the sort just indicated when zero and negative values of clues are allowed, and lessens distortions even when they are not. It can be viewed as a policy of separation of quality variables from quantity variables. For instance, in Eq. (1) the quality variables are $X_1$ through $X_5$, while the quantity variable is $X_6$.

It can also be argued that separating quantity from quality tends to make the resulting regression equation more robust -- that is, less sensitive to systematic errors or changes in the data and more amenable to transfer into different collections or changed conditions. To see this, consider the effect on retrieval of multiplying all IDF values (before logging) by a factor of ten (perhaps because of a systematically misplaced decimal point). With a little study it can be seen that the output ordering of the documents imposed by Eq. (1) will remain unchanged. This is very different from the effect under the conventional totaling procedure, in which the tenfold increase would cause the influence of the IDF variable to overwhelm that of the other variables, and would greatly affect the ordering. Changes by an additive constant (before logging) also cause less disruption of the output ordering when the quality/quantity dichotomy is maintained. Thus there is a lessened sensitivity to any (linear) rescaling of the data from whatever cause.

**Manual Query Modification**

There has been interest recently in the question of whether it is possible to improve retrieval results significantly by means of expert manual modification of the original natural language query submitted by the user. The Berkeley group's second ad hoc retrieval experiment (Brkly7) attempted to cast some further light on this issue. The second experiment was identical to the first in all respects, including the use of Eq. (1) without modification, with the sole exception that the natural language topics were transformed by human intermediaries into new word lists before being submitted for stoplisting and stemming.

The human intermediaries were instructed to work intuitively to include in the list any words likely to appear in documents relevant to the topic at hand. They were encouraged to consider including, for a start, any content words and phrases that appeared in the original topic, though there was no compulsion to include those that seemed unhelpful. To expand the list further, they were instructed to add to these any synonyms, broader or narrower terms, or indeed any other words they could think of that might enhance the retrieval.

The intermediaries were allowed to exploit any personal knowledge of the topic they might happen to

possess. For instance, for the topic on oil spills, a knowledgeable intermediary might add to the list 'Exxon' and 'Valdez'. As a further source of indicative words and phrases, the intermediaries were encouraged to make crude Boolean searches in other data bases and, when obviously relevant documents were found, to scan them for potentially useful new terms. The Newspaper Articles Database, freely searchable via the University of California's MELVYL catalog system, was used for this purpose. Most of the searches were made in its New York Times database, though a few were made in its Los Angeles Times or Washington Post files. Care was taken not to do any searching in sources, such as the Wall Street Journal, which were also included in the TREC corpus, as this would have contravened the TREC guidelines. The WordNet thesaurus was made available to the intermediaries but for the most part they declined to use it, finding the Boolean searches in the auxiliary collections to be a more fertile source of helpful terms. Finally, the intermediaries were encouraged to write into their lists double or multiple instances of any terms that they considered especially vital to the satisfaction of the information need.

Except for the first two topics, on which the intermediaries spent a long time getting used to the process, about twenty to thirty minutes were spent modifying each topic. This was thought to be an amount of time that a zealous librarian or other intermediary might be willing to spend on occasion to try to improve an important search request. The idea was to make the hypothetical conditions as favorable to modification as possible, to see if manual intervention is of any avail even under especially advantageous circumstances.

To illustrate the results of this process of topic modification, here is the reconstructed form of topic 154, which concerns oil spills:

> *oil spills oil spills oil spills oil spills oil spills oil spills oil spills oil spills oil spills oil spills major offshore drilling holding tank spills light crude tons gallons slick Exxon Valdez Prince William Hazelwood Unocal Guadalupe Shetlands San Juan Puerto Rico Rubis Lyria Mobil Shell Francisco Ashland Newport Hudson Latouche Delaware Narragansett Rhode Huntington*

In the first ad hoc retrieval experiment, which involved no manual query modification, the average precision over all relevant documents for all fifty routing topics was 27.8%. In the second ad hoc experiment, which differed only in the manual query modifications, it rose to 37.1%. This substantial increase suggests that extensive reworking of natural language requests by intermediaries can indeed, under favorable circumstances at least, considerably enhance a system's retrieval effectiveness.

**Topic-Free Routing Retrieval**

For its third TREC-3 experiment (Brkly8) the Berkeley group designed a routing retrieval algorithm along the following, we suspect novel, lines. It is usual to base all document retrieval, including routing retrieval, on some sort of search specification derived from the original natural language topic. In many routing systems such a search specification is then expanded in some way by the inclusion of terms taken from documents known from past user feedback to be relevant to the topic. However, under the present Berkeley approach the original topic wording can be ignored completely, provided enough feedback is available.

The unusual character of this procedure is apparent from the fact that if the original natural language topic were written in a language different from the language of the collection, it would not matter, because the method is not dependent upon finding matches between topic stems and document stems. Rather, the internal search request vocabulary is drawn entirely from previously judged documents. It is not essential that the method be applied in this radical, absolutely topic-free form; and in fact the Berkeley experiment did make one slight gesture in the direction of exploiting the original topic wording, as will be explained. However, the essence of the method is that the search specification be constructed independently of the original topic expression to the extent allowed by the amount of available feedback.

As the first step in the construction of the internal search request for a topic, every stem that occurred in at least one document in the set of judged documents for the topic was tested to see whether it was statistically related to the property of relevance to the topic. This was done by constructing a four-cell table for each such stem and conducting a chi-square test on it. The two rows of each stem's table corresponded to the presence or absence of the stem in a document, while the two

columns corresponded to relevance or nonrelevance to the topic of a document. Thus the upper left cell contained the number of documents that (i) contained at least one occurrence of the stem and (ii) had been judged relevant to the topic. The sum of the quantities in all four cells was the total number of available judged documents for the topic.

The chi-square test showed which stems had the property that their presence in a document was statistically associated (either positively or negatively) with the relevance of the document to the topic. All stems for which the hypothesis of independence (i.e. lack of association) could be rejected at the 0.05 level of statistical significance under the test were collected and used in the search specification for the topic; all others were discarded. The stem sets so collected ranged in size from 300 stems for Topic 131 to 4,114 stems for Topic 109. The mean stem set size was 1,357.

Next, for each topic, a weight was calculated for each stem that had survived the chi-square screening for that topic. This weight was an estimate of the surplus logodds (i.e. the logodds over and above the prior logodds) that a document under consideration is relevant to the topic, given that the stem is absent from the document. The motivation for using such estimates as weights will be explained in the next section; for the moment we merely state the formula used to calculate the weight. It is:

$$weight_s = \log \frac{r_s + r / (n + r)}{n_s + n / (n + r)} - \log \frac{r}{n} \qquad (2)$$

where $r_s$ is the number of documents in which the stem is absent that have been judged relevant to the topic, $n_s$ is the number of documents in which the stem is absent that have been judged nonrelevant to the topic, $r$ is the total number of documents that have been judged relevant to the topic, and $n$ is the total number of documents that have been judged nonrelevant to the topic.

For stems occurring in the original topic, weights were calculated in accordance with the foregoing formula whether the stems passed the chi-square screening test or not. In cases where these weights turned out to be negative, their magnitudes were arbitrarily increased by a factor of five. This was the only concession in the direction of using the original topic wording. It was made because it was found to improve the retrieval effectiveness of the scheme slightly. However, the tiny size of the gain suggested that this refinement is unimportant when sufficient relevance data are available.

To illustrate the form of the internal search requests, here is a small segment of the list of weighted stems constructed for Topic 102, which is about U.S. strategic defense:

| | |
|---|---|
| decid | 0.236631 |
| declin | 0.103601 |
| decoy | -0.162831 |
| defend | -5.052930 |
| deform | -0.014523 |
| defray | -0.014523 |
| degree | 0.083949 |
| deliv | 0.140666 |
| delt | -0.072431 |
| demand | 0.144963 |

A negative weight on a stem indicates that the stem's absence from the document reduces the chances that the document is relevant. Thus a stem with a negative weight is really a positive clue. For instance, the excellent clue 'defend' has a negative weight -- a large negative weight, in fact, because it occurs in the topic wording and hence had its weight multiplied by five.

The in-principle retrieval rule for estimating the logodds of relevance of a document to a topic is simply to add up the weights of all stems in the stem list for the topic that are absent from the document. In practice, however, it is quicker and produces the same result to add up the weights of all the stems in the list that are present in the document, and subtract this sum from the prestored grand total of all the weights in the stem list for the topic.

While the Brkly8 routing entry produced respectable recall and precision, there remains some question that the size of the expanded queries may be chosen as too large. The Cornell group has conducted experiments (Buckley, Salton, Allan 1994) on massive query expansion and concluded that the point of diminishing returns was an expansion by about 300 terms. Since the Berkeley tests show performance enhancement for terms chosen up to the 0.1 probability level of the Chi square test on the training set, and the 0.05 level was chosen for realistic computational times, we see evidence that evaluation on a training set may lead to the phenomenon of *overtraining* when the expansion is applied to the new test document set.

**Theoretical Motivation of the Routing Rule**

The foregoing retrieval rule is based on a statistical simplifying assumption called the 'Assumption of Linked Dependence' (Cooper 1991). Intuitively, it states that in the universe of all query-document pairs, the degree of statistical dependency that exists between properties of pairs in the subset of all relevance-related pairs is linked in a certain way with the degree of dependency that exists between the same pair-properties in the subset of all nonrelevance-related pairs. For $N$ pair-properties $A_1, \ldots, A_N$, the mathematical statement of the assumption is as follows.

ASSUMPTION OF LINKED DEPENDENCE:

$$\frac{P(A_1, \ldots, A_N \mid R)}{P(A_1, \ldots, A_N \mid \bar{R})} = \frac{P(A_1 \mid R)}{P(A_1 \mid \bar{R})} \times \cdots \times \frac{P(A_N \mid R)}{P(A_N \mid \bar{R})}$$

If one thinks of a query and document drawn at random, $R$ is the event that the document is relevant to the query, while each clue $A_i$ is some property of the topic-document pair, for example some property of the document that makes it either more or less likely to be relevant to the topic.

From the Assumption of Linked Dependence the following probabilistic modelling equation can be derived:

$$\log O(R \mid A_1, \ldots, A_N) = \log O(R) + \sum_{i=1}^{N} [\log O(R \mid A_i) - \log O(R)] \qquad (3)$$

where for any events $E$ and $E'$ the odds $O(E / E')$ is by definition $\frac{P(E / E')}{P(\bar{E} / E')}$. Using this equation, the evidence of $N$ separate clues can be combined as shown in the right side to yield the logodds of relevance based on all clues, shown on the left. Query-document pairs can be ranked by this logodds estimate, and a ranking of documents for a particular query by logodds is of course a probability ranking in the IR sense. For further discussion of the linked dependence assumption and a formal derivation of Eq. (1) from it, see Cooper (1991) and Cooper, Dabney & Gey (1992).

The properties $A_i$ in the modelling equation can in general be any clues; for instance the occurrence in a document of a particular stem could serve as such a clue, as could its nonoccurrence. In the experiment just described, each property $A_i$ of interest is the absence from the document of some particular stem from the constructed internal stem-list for the topic at hand. The weighting formula (2) estimates, for any given stem-absence, the corresponding quantity in square brackets on the right side of the modelling equation (3). The expressions $r / (n + r)$ and $n / (n + r)$ added into the numerator and denominator of the first fraction prevent either from taking on a value of zero. The retrieval rule stated in the preceding section is simply the computation of the summation indicated in the right side of (3). The initial term $\log O(R)$ appearing after the equal sign in (3) is omitted from the computation because it was found to affect the result only insignificantly in practice.

Either the presence of a stem in a document or its absence may be used as a retrieval clue. For TREC-2, as well as in preliminary TREC-3 experiments, the Berkeley group used both types of clues in tandem. Under that approach, for each stem there were two weights in the search query, one to be added into the retrieval score in case the stem was present in the document, and the other to be added in in case the stem was absent. However, subsequent experimentation showed that the two sets of weights were largely redundant, in the sense that use of just one type produced almost as effective retrieval as use of both. Stem-absence worked somewhat better than stem presence, and was virtually as effective by itself as when used in tandem with stem presence clues. So, to simplify the procedure, the stem-present weights were eliminated.

The Assumption of Linked Dependence, like all simplifying assumptions, introduces errors into the logodds computations. When the number of properties involved is large, as is the case in the present application, this distortion can become severe, taking the form of grossly exaggerated estimates for the logodds of relevance of documents near the top of the output rankings. This exaggeration does not affect the ordering of the output, but it makes the probabilities of relevance displayed for the user unrealistic. To counter this tendency, each sum calculated by the retrieval rule was divided by the square root of the number of stems in the query list for the topic of the search. This crude remedy suggested itself because of its success in compensating for similar distortions in an ad hoc experiment performed by the Berkeley group for TREC-2.

**Table I: Brkly8 Routing Entry**
Comparison between estimated and actual relevant/non-relevant

| recall level | num docs | actual relevant | actual non-rel | actual probability | estimated probability | expected relevant | expected non-rel |
|---|---|---|---|---|---|---|---|
| 0.1 | 1054 | 683 | 371 | 0.6480 | 0.7725 | 814.21 | 239.78 |
| 0.2 | 1098 | 683 | 415 | 0.6220 | 0.6547 | 718.96 | 379.03 |
| 0.3 | 1351 | 683 | 668 | 0.5056 | 0.5061 | 683.81 | 667.18 |
| 0.4 | 1685 | 683 | 1002 | 0.4053 | 0.4178 | 704.14 | 980.85 |
| 0.5 | 2188 | 683 | 1505 | 0.3121 | 0.3585 | 784.57 | 1403.43 |
| 0.6 | 1983 | 683 | 1300 | 0.3444 | 0.2610 | 517.55 | 1465.45 |
| 0.7 | 1952 | 683 | 1269 | 0.3499 | 0.2290 | 447.04 | 1504.96 |
| 0.8 | 2609 | 683 | 1926 | 0.2618 | 0.1882 | 491.08 | 2117.91 |
| 0.9 | 4330 | 683 | 3647 | 0.1577 | 0.1276 | 552.55 | 3777.45 |
| 1.0 | 6940 | 681 | 6259 | 0.0981 | 0.0344 | 238.92 | 6701.08 |

**Evaluation of Probability Estimates**

As in TREC2, the Berkeley text retrieval research group is concerned with the accuracy of probability estimates obtained through its methods. If one can, with confidence, display these estimates to the user, along with the ranking of documents presumed relevant to the topics, the user has the added flexibility of choosing where to stop the search. For example, if the probability estimate is 0.1, and the estimate has been shown to be realistic, the user can realize that, on average, (s)he will have to examine ten documents to find the next relevant one.

One way to test the calibration of probability estimates is to take the entry of 50,000 query-document pairs and sort them in descending order of magnitude of probability estimate. Using the results of the judged set of documents from the TREC3 relevance judges, one can attach relevance judgments for those pairs which have been judged. Of the 50,000 pairs submitted for Brkly8, 25,190 pairs were in the judgment set. This list can then be divided into deciles of recall, and the mean estimated probability of relevance for each level of recall can be computed. Since we know the actual number of relevant for each decile, the actual probability of relevance for that decile can be computed as well. From the estimated probabilities, expected numbers of relevant and non-relevant documents can also be computed. Table I shows the results for the ten levels of recall for Brkly8.

In evaluation of logistic regression models, Hosmer and Lemeshow (1989) have shown that the two sets of 10 pairs of actual relevant/non-relevant and estimated relevant/non-relevant documents in each decile of recall can be used to assess the fit of the model. In particular, the fit is demonstrated to follow a $\chi^2$ distribution with 8 degrees of freedom. Computing this statistic for Table 1, we have

$$\chi^2 = 1327.69$$

and a probability value of less than 0.0001. Similar computations for Brkly6 and Brkly7 result in the following statistics:

$$\chi^2 = 861.99 \quad (Brkly\,6)$$

$$\chi^2 = 758.10 \quad (Brkly\,7)$$

also with probability values of less than 0.0001. For further details on the actual formulae to compute the $\chi^2$ statistics, the reader is referred to Gey (1993).

These results would lead us to reject the null hypothesis that the models correctly fit the data, and perhaps lead us to conclude that further refinement of these models is in order.

**Summary**

The Berkeley entries to TREC3 have built upon previous research using logistic regression and the principle of linked dependence for combination of evidence. For ad hoc retrieval, we suggest that a separation of quality and quantity variables is appropriate in a regression-based methodology, with the quality variables being combined by taking means rather than sums. The

Brkly7 entry demonstrates that judicious manual intervention to reconstruct queries can, under favorable circumstances, increase retrieval effectiveness considerably. Specifically, the inclusion of proper noun *particularization* of examples of relevant real-world objects (companies, persons, geographic place names) seems to move the query closer to the relevant set.

For routing retrieval, the Brkly8 entry shows that, given a history of relevance judgments, internal requests can be constructed virtually without reference to the original topic wording and that query terms can be combined by a simple retrieval rule based on the Principle of Linked Dependence. A $\chi^2$ test on degree of dependence of the term to relevance is used to rigorously define the query expansion term set. However results seem to suggest that an "overtraining effect" may come in to play when moving from the training judgment set to the test collection.

Finally, preliminary tests to assess the effectiveness of the probability estimates generated by these methods indicate that additional refinement of the models will be necessary before the probabilistic system can, with confidence, display these estimates to the user.

**Acknowledgements**

We are indebted to Ray Larson for helpful systems advice, as well as to the several colleagues already acknowledged in our TREC-1 and TREC-2 reports. Edwin Cooper, Emily Cooper, and Nate Kurz donated their services as intermediaries in the second ad hoc experiment. The work stations used for the experiment were supplied by the Sequoia 2000 project at the University of California, a project principally funded by the Digital Equipment Corporation.

**References**

Buckley, C.; Salton G.; Allan, J. The Effect of Adding Relevance Information in a Relevance Feedback Environment. *Proceedings Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland 272-281. July 1994.

Cooper, W. S. Inconsistencies and Misnomers in Probabilistic IR. *Proceedings Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Chicago: 57-62. October 1991.

Cooper, W. S.; Dabney, D.; Gey, F. Probabilistic retrieval based on staged logistic regression. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Copenhagen: 198-210; June 1992.

Fuhr, N.; Buckley, C. Optimizing Document Indexing and Search Term Weighting Based on Probabilistic Models. *The First Text REtrieval Conference (TREC-1).* NIST Special Publication 500-207. Washington: 89-100; 1993.

Gey, F. C. *Probabilistic Dependence and Logistic Inference in Information Retrieval*. PhD Dissertation, Graduate Division, University of California, Berkeley, May 1993.

Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression.* New York: Wiley; 1989.