Probabilistic Retrieval Based on Staged Logistic Regression

by
William S. Cooper, Fredric C. Gey
S.L.I.S., University of California, Berkeley

and
Daniel P. Dabney
G.S.L.I.S., University of California, Los Angeles

Abstract

The goal of a probabilistic retrieval system design is to rank the elements of the search universe in descending order of their estimated probability of usefulness to the user. Previously explored methods for computing such a ranking have involved the use of statistical independence assumptions and multiple regression analysis on a learning sample. In this paper these techniques are recombined in a new way to achieve greater accuracy of probabilistic estimate without undue additional computational complexity. The novel element of the proposed design is that the regression analysis be carried out in two or more levels or stages. Such an approach allows composite or grouped retrieval clues to be analyzed in an orderly manner -- first within groups, and then between. It compensates automatically for systematic biases introduced by the statistical simplifying assumptions, and gives rise to search algorithms of reasonable computational efficiency.

15th Ann Int'l SIGIR '92/Denmark-6/92

Introduction

By a *probabilistic* search system design we mean one in which the system ranks the documents for the searcher in decreasing order of their statistically estimated probability of usefulness (or 'relevance') to the searcher's information need. Thus the user is always guided to examine next the document that has been explicitly calculated by the system to be next-most-likely to be satisfactory. Since the time when probabilistic retrieval theory was conceived over thirty years ago (counting from Maron & Kuhns (1960)), there has been a steady development of the approach (surveyed e.g. in Maron 1984, Bookstein 1985, Cooper 1991a). A few probabilistic or semi-probabilistic systems are now in actual operation (Harman 1992). However, difficulties have been encountered in attempts to find methods that are both reasonably well-calibrated in their estimates of the relevance probabilities, and at the same time computationally efficient.

In this paper we wish to suggest a scheme for synthesizing existing ideas -- a unifying methodology that alleviates some of the more vexing difficulties. The novel feature of the proposed method is that it supplements the customary appeal to probabilistic simplifying assumptions with two or more applications of

multiple logistic regression. In the case of a two-stage regression -- the simplest application of the approach and the one with which we shall be principally concerned here -- the function of the first stage of the analysis is to estimate an individual predictive statistic for each composite retrieval clue. The second stage then combines the individual statistics obtained in the first stage. In more complicated circumstances, as when a relational thesaurus is to be used in the retrieval, three or even four levels of regression may be called for.

The Potential Benefits of Probabilistic Retrieval

In a thoroughgoing probabilistic design methodology, serious use is made of formal probability theory and statistics to arrive at the estimates of probability of relevance by which the documents are ranked. Such a methodology is to be distinguished from looser approaches -- for instance the 'vector space' retrieval model -- in which the retrieved items are ranked by a similarity measure (e.g. the cosine function) whose values are not directly interpretable as probabilities.

When sound statistical methods are applied to make the probability-of-relevance estimates, one has grounds for expecting retrieval effectiveness that is near-optimal relative to the evidence used. This is a principal (hoped-for) advantage of the probabilistic approach, for according to the 'Probability Ranking Principle' a document retrieval system will rank its output in an optimally useful way if it responds to each query by ordering the documents of the collection in decreasing probability of relevance to that query, estimating all probability estimates as accurately as possible on the basis of the available clues (Robertson 1977). That is to say, to the extent that the probability estimates are as sound as possible, the system may be expected to perform as well as possible. The Probability Ranking Principle is not an absolute law (Cooper 1973, Gordon & Lenk 1991), and is moreover only a special case of the more general Expected Utility Ranking Principle (Cooper & Maron 1978). But as a simple and approximately valid design guide it lends strong support to the probabilistic design approach.

A second hoped-for advantage of probabilistic retrieval is that near-optimal probabilistic designs can be developed more efficiently, with a less exclusive reliance on traditional trial-and-error retrieval experiments. Under traditional approaches, the optimization of a retrieval rule typically requires that a series of retrieval trials be conducted to discover the parameter values that result in the best performance (Sparck Jones 1981). In contrast, under the probabilistic approach the statistical procedures employed in the course of the regression analysis can under favorable circumstances be expected to yield optimal or near-optimal retrieval parameters directly and immediately. In such cases there may be little need for conventional testing except possibly as a final check or for comparative testing against other ranking rules developed outside the regression framework.

Another point of contrast is that in traditional IR experimentation, effectiveness is usually evaluated using such problematic measures as precision, recall, fallout, and the like, using tests of statistical significance whose sensitivity may be unimpressive. But when logistic regression (say) is used to derive the optimal probabilistic retrieval rule, an array of more powerful statistical indicators of predictivity and goodness-of-fit become available as an intrinsic part of the regression analysis. Standard statistical software packages provide much of the needed significance testing and other analytic detail. None of this is meant to imply that a probabilistic approach in any way lessens the investigator's dependence upon empirical data. Rather, the point is that it can shift much of the burden of the analysis of that data into a realm where more highly developed statistical methods can be brought to bear.

The third potential advantage of a probabilistic design is that each retrieved document's probability-of-relevance estimate can be reported to the user in the ranked output. Such probability estimates can provide information to the user that is of value in deciding whether to examine some particular document in detail, as well as in deciding whether or not to continue the search down the output ranking. True, some nonprobabilistic designs also report a numeric 'retrieval coefficient' along with each ranked document, but such numbers are apt to be more difficult for the average user to interpret. For instance, it would presumably be easier for most users to understand and base their stopping behavior upon a retrieval coefficient describable simply as a 'probability of relevance' than one described as 'the cosine of the angle between the query vector and the document vector in term space'.

The History of Previous Research

If the potential benefits of the probabilistic approach are so worth while, why have they not yet been more widely realized? In considering this question it may be helpful to review briefly some of the lines of probabilistic research that have been pursued in the past. One recurrent idea about how probability theory might be applied to the retrieval problem is that statistical simplifying assumptions, for example independence assumptions, might be invoked to facilitate the probability computations. Such simplifying assumptions allow joint probabilities involving many events to be broken down into simpler expressions involving only one event. The best-known theories of this sort are those of Maron and Kuhns (1960), in which conditional independence is assumed within a set of users for whom a document of interest would be relevant; and the approach of Robertson and Sparck Jones (1976) and others (Yu & Salton 1976, van Rijsbergen 1979) in which it is assumed that there is conditional independence within both the relevant and nonrelevant sets of documents. The latter assumptions give rise to what has been widely referred to as the 'binary independence model' of retrieval. (Actually that description is misleading, because the model in question is really based on assumptions of linked dependence, not independence (Cooper 1991b).)

No one supposes that the usual statistical simplifying assumptions are precisely true in any real retrieval environment. (The question of just how far they depart from the truth is currently being investigated empirically by one of the authors (F.C.G.).) Because of these doubts about the soundness of their underlying assumptions, all retrieval theories that use them naively -- that is, without correcting explicitly for their deficiencies -- are open to question. By the same token it can be doubted whether the above-discussed advantages of the probabilistic approach have yet been fully realized by any pure independence-assumption approach.

One possible avenue of escape from the problem might be to weaken or eliminate the simplifying assumptions, and to pay the price for so doing by gathering empirical evidence about the statistical dependencies among the retrieval clues. The possibility of exploiting cooccurrence information to accomplish this has been studied extensively, and the cause is far from hopeless (see e.g. van Rijsbergen 1977, 1979; Harper & van Rijsbergen 1978; Cooper & Huizinga 1982; Robertson & Bovey 1982; Cooper 1983; Yu, Buckley, Lam, & Salton 1983; Kantor 1984; Lee & Kantor 1990). However, it has to be admitted that deducing joint probabilities from scanty collection data is a delicate undertaking, and none of the schemes so far suggested is especially convenient computationally.

Another important stream of thought in probabilistic retrieval has involved the application of statistical regression analysis or pattern recognition. For this a 'learning set' is used that consists typically of a sample of documents and queries, together with indications of which of the documents are in fact relevant to which of the

queries. When multiple regression analysis is applied to the sample, the result is a predictive equation capable of estimating, for any future query-document pair, the probability that the document is relevant to the query. This method was used by Fox (1983) to combine measurements of subject-term similarity, citation-link similarity, and other kinds of similarity, into what was in effect a crude estimate of overall relevance probability. A more extensive examination of the application of standard polynomial regression to the retrieval problem was contributed later by Fuhr (1989). These investigations made no use of independence postulates or other statistical simplifying assumptions particular to the retrieval task. They were 'model-free' in the sense that the only probabilistic assumptions involved were those implicit in the statistical regression theory itself.

For some IR investigations this model-free approach seems justified, but its range of application is limited. Its most serious drawback is that it treats one piece of retrieval evidence as an independent variable on a par with every other, with no recognition given to the patterns that give structure to the data. This can make its use problematical when the clues are grouped or organized in some special way. Consider for instance the situation, ubiquitous in full-text retrieval using discursive natural language requests, where for each word stem occurring in both the document and the request there is an associated set of facts. The facts might include the term's relative frequency of occurrence in the document, its relative frequency in the request, its inverse document frequency, whether it appears in the title, and so forth. In such a case a 'match' on a term is really a *composite* clue -- a whole constellation of elementary clues that collectively describe the character of the term match.

To cope with structured data of this kind, a traditional statistical simplifying assumption may be used to divide a complex joint probability expression into simpler ones, each of which itself pertains to some composite clue. Each composite clue is then evaluated on the basis of its own constellation of elementary clues using regression techniques. This way of teaming traditional modelling assumptions up with regression analysis has been shown to be feasible by Fuhr and Buckley (1991), who used standard polynomial regression techniques to evaluate expressions for composite clues obtained with the help of the independence assumption due to Maron & Kuhns (1960). This approach is promising insofar as it combines the divide-and-conquer virtues of the earlier simplifying-assumption approaches with the empirical realism of regression on a learning set.

Unfortunately such an approach shares the failing of the earlier modelling approaches, namely that the statistical simplifying assumption inevitably distorts the final probability-of-relevance estimates. Moreover, there is in this context a serious problem with the use of standard polynomial regression methods. Standard regression theory is based on the assumption that the sample values taken on by the dependent variable are drawn from a continuum of possible magnitudes. In the retrieval application, however, the dependent variable values in the learning set are usually dichotomous: a document is either relevant to a query or it is not. So it is, strictly speaking, logically inappropriate to use standard regression methods to estimate probabilities of relevance in IR.

A more appropriate tool is available in *logistic regression*, a statistical methodology specifically developed for use with dichotomous (or n-chotomous) dependent variables. Logistic analysis was first introduced into the retrieval arena by Robertson and Bovey (1982). Its applicability to the retrieval problem was also noted by Bookstein (1985), and Bookstein *et al* have recently provided a readable introduction to the subject (1991). Related techniques have been employed to advantage by Fuhr and Pfeifer (1991), and one of us (D.P.D.) has had some

success in applying a model-free logistic regression analysis to encyclopedia data kindly supplied by researchers at Cornell University.

The SLR Methodology

The method to be presented here represents an attempt to bring into a harmonious confluence several of these streams of ideas. The guiding notion is to treat composite clues on at least two levels -- an intra-clue level in which a predictive statistic is estimated separately for each composite clue, and an inter-clue level in which these separate statistics are combined to obtain a probability-of-relevance estimate for the given query and document. Because the analysis proceeds in successive stages, we shall refer to the general method as *Staged Logistic Regression* (or for short *'SLR'*).

The procedure for a two-stage SLR analysis can be outlined as follows. *Step 1:* A statistical simplifying assumption is used to break down the complex joint probabilistic expression of ultimate interest into simpler expressions for the composite clues. *Step 2:* A logistic regression analysis on a learning sample is used to obtain a predictive equation (the 'first-level' regression equation) for estimating the values of the latter expressions. *Step 3:* A second logistic regression analysis, based on the same learning sample, is used to obtain another predictive rule (the 'second-level' regression equation) for combining the composite clues and correcting for any systematic biases introduced by the simplifying assumption.

The SLR method requires that two or more regression analyses be performed at the research stage. Although this sounds laborious, once the researcher has set up shop to do regression two analyses of the same data require little more effort than one. And during actual retrieval the computations required to carry out the search algorithm are comparable in simplicity and efficiency to other retrieval processes of proven computational feasibility such as vector-space searching.

Step 1: The Linked Dependence Model

The probabilistic approach requires that it be possible, for any query that might be submitted to the system and any document in the document collection, to estimate the probability that the document will be relevant to the query. Suppose then that a particular query and document are in hand, and the problem is to estimate the probability of the event R that the two are relevance-related. The evidence on which the estimate is to be based will consist in general of various machine-detectable properties $A, B, C, \ldots$ of the query-document pair. For this purpose any properties or relationships may be used that could serve as clues to the possible relevance-relatedness of the pair.

What is wanted is a method of estimating $P(R/A,B,C,\ldots)$ -- that is, of deducing the probability that the document is relevant to the query, given all known clues. To break down this complex expression into a combination of simpler ones we shall introduce a statistical simplifying assumption of *linked dependence,* which asserts that if two or more retrieval clues are statistically dependent in the set of all relevance-related query-document pairs, then they are statistically dependent to a corresponding degree in the set of all nonrelevance-related pairs. Thus the degrees of dependency in the relevant and nonrelevant sets are assumed to be 'linked'.

Considering first for simplicity the special case where there are only two clues A and B, the assumption may be formulated as follows:

LINKED DEPENDENCE ASSUMPTION: There exists a positive real number K such that the following two conditions hold true:

$$P(A,B/R) = K\ P(A/R)\ P(B/R)$$

$$P(A,B/\sim R) = K\ P(A/\sim R)\ P(B/\sim R)$$

That is to say, the joint probability of A and B is assumed to differ from its independence value by the same factor K whether attention is confined to the relevant or to the nonrelevant query-document pairs. The magnitude of K may be conceived as a crude measure of the statistical dependency that exists between A and B.

If $K = 1$ the Linked Dependence Assumption reduces to the familiar 'binary independence' assumptions often adopted in the literature on probabilistic retrieval (Robertson & Sparck Jones 1976, Yu & Salton 1976). As those assumptions are usually stated, the retrieval clues are posited to be independent in both the relevant and nonrelevant sets -- an implausible presumption. Here, in contrast, we have not assumed that $K = 1$, but only that the value of K (whatever it may be) is the same in the relevant set as in the nonrelevant. This weaker assumption, though still debatable, has at least the virtue of not denying the existence of dependencies.

From the Linked Dependence Assumption it follows immediately that

$$\frac{P(A,B/R)}{P(A,B/\sim R)} = \frac{P(A/R)}{P(A/\sim R)}\ \frac{P(B/R)}{P(B/\sim R)}$$

For nonzero probabilities this condition is in fact a mathematically equivalent restatement of the Linked Dependence Assumption. Generalizing from two to N properties, the assumption becomes

$$\frac{P(A_1,\ldots,A_N/R)}{P(A_1,\ldots,A_N/\sim R)} = \prod_{i=1}^{N} \frac{P(A_i/R)}{P(A_i/\sim R)}$$

The *odds* of occurrence of an event E, symbolized *'O(E)'*, is by definition $O(E) = P(E)/P(\sim E)$. From this definition and the usual definition of conditional probability it follows that

$$\begin{aligned}\frac{P(E_1/E_2)}{P(E_1/\sim E_2)} &= \frac{P(E_1,E_2)}{P(E_1,\sim E_2)}\ \frac{P(\sim E_2)}{P(E_2)} \\ &= \frac{P(E_2/E_1)}{P(\sim E_2/E_1)}\ \frac{P(\sim E_2)}{P(E_2)} \\ &= \frac{O(E_2/E_1)}{O(E_2)}\end{aligned}$$

Applying this identity to both sides of the earlier equation gives

$$\frac{O(R/A_1,\ldots,A_N)}{O(R)} = \prod_{i=1}^{N} \frac{O(R/A_i)}{O(R)}.$$

Multiplying through by $O(R)$ and taking the logarithm of both sides, one obtains

$$Log\ O(R/A_1,\ldots,A_N) = Log\ O(R) + \sum_{i=1}^{N} [Log\ O(R/A_i) - Log\ O(R)]$$

Equation 1

The logged quantities in the equation are called 'logodds'. When at retrieval time the system receives a query, Eq. (1) may be used to

calculate the logodds *Log* $O(R/A_1,\ldots,A_N)$ of relevance for each document in the collection. (Or better, the modified form of (1) to be introduced later may be used.) The documents are then ranked for the user in decreasing order of these estimated logodds. Because logodds are monotonic transforms of probabilities, ranking the collection by descending order of logodds of relevance has the same ordering effect as a ranking by descending order of probabilities of relevance. Of course, for documents or citations actually retrieved or displayed for the user, the corresponding logodds may be transformed back into probabilities for ease of interpretation by the user.

Step 2: Estimation of Composite Clue Logodds

In order to estimate the values of the terms on the right side of Eq. (1), empirical data must be gathered concerning the retrieval environment for which the system is being designed. Ideally, such data would consist of a large random sample of documents (or document representations) from the collection in which the system is to operate, a large random sample of real queries (or query representations) submitted by members of the prospective user population, and for each query-document pair consisting of a query from the query sample and a document from the document sample, a judgement by the user who submitted the query as to whether that document is in fact relevant to the information need that prompted the query. Such perfectly formed learning sets do not yet exist, but some of the so-called 'test collections' constructed by information retrieval experimenters probably approach the ideal sufficiently closely to be useful.

The quantities in Eq. (1) that have to be estimated are of two forms, *Log* $O(R)$ and *Log* $O(R/A_i)$. The estimation of *Log* $O(R)$ from a learning set is a straightforward matter, for the simple proportion of query-document pairs in the learning set that are relevance-related can be used to estimate $P(R)$ and the complementary proportion provides an estimate of $P(\sim R)$, from which the logodds follows. *Log* $O(R/A_i)$ can be estimated analogously for any property A_i for which there exist sufficiently many representative pairs in the learning set.

Complications arise when there are too few query-document pairs in the learning set with the property A_i to yield reliable estimates of $P(R/A_i)$ and $P(\sim R/A_i)$. Unfortunately this will be the usual case when A_i is a composite clue. A method of analysis that goes beyond simple averaging is then needed. We recommend logistic regression. Multiple logistic regression analysis is a statistical tool for deriving from a learning sample an equation that predicts the value of a certain 'dependent' variable as well as possible on the basis of the values of one or more 'independent' variables. The dependent variable to be predicted is in the IR application the value of *Log* $O(R/A_i)$ for the composite retrieval clue A_i. The independent variables -- call them $X_1,\ldots,X_M$ -- are the elementary clues that make up this composite clue. If for instance A_i is a cluster of facts about a certain word stem that occurs in both the query and the document, then X_1 might be (say) the relative frequency of the stem in the query, X_2 its relative frequency in the document, and X_3 its inverse document frequency in the collection.

More generally, for any composite clue about any query-document pair in the learning sample there will be a set of values $X_1,\ldots,X_M$ representing the elementary clues that make up the composite clue, and also a record of the observed relevance/nonrelevance relationship that exists between that query and that document. If all composite clues are of the same type, the set of all such data derivable from the learning set will have the form of a matrix with $M + 1$ columns. Submitting it (or a sample drawn from it) to a statistical program package capable of performing multiple logistic regression yields a series of coefficients $c_0,\ldots, c_M$. These specify a predictive equation of the form

$$Log\ O(R/A_i) = Log\ O(R/<X_1,\ldots,X_M>) = c_0 + c_1\ X_1 + \ldots\ c_M\ X_M$$

Equation 2

where $X_1,\ldots,X_M$ are the elementary attributes of A_i. When during retrieval a query and document are compared, for each composite clue A_i Eq. (2) is applied to get an estimate of *Log* $O(R/A_i)$. The logodds so obtained are then combined with an estimate of *Log* $O(R)$ in Eq. (1) (actually the modified version of Eq. (1) still to be presented) to obtain the desired logodds of relevance.

Convenient computational arrangements are possible. To illustrate, suppose each composite clue consists of three elementary clues about a matching term: its relative frequency X_1 in the request; its relative frequency X_2 in the document, and its inverse document frequency X_3 in the collection. When the system is initially set in operation, every index term for every document is assigned a weight $c_0 + c_2 X_2 + c_3 X_3$ - *Log* $O(R)$. This weight is stored permanently with the term in the document's index record. When at run time the system receives a search query, each term in the query is assigned a weight $c_1\ X_1$. Then, in comparing the query against a document during the search, each matching term's contribution to the sum in Eq. (1) becomes simply the sum of its document weight and its query weight.

There is, if anything, a modest gain in computational efficiency over what would be required for a comparable vector space retrieval search. A vector processing scheme (that calculates, say, a cosine or a vector product) must compute for every matching term the product, not just the sum, of its document weight and its query weight.

Discussion of Logistic Regression

To understand the role of logistic regression, it is helpful to suppose that the observed relevance relationships for all query-document pairs in the learning set have been coded as 1's (for judgements of relevance) and 0's (for nonrelevance). These 1's and 0's may be conceived as pseudo-probabilities of relevance; they have to be either 1 or 0 because the relevance status of each pair in the learning set is already known.

Under this representational convention the information in the learning set can be graphed as shown in Figure 1. There it has been assumed for simplicity that all relevance probabilities are to be estimated on the basis of only one independent variable X which might be, say, the frequency of occurrence in the document of a matching term. Each instance of a term match between a query and a document is shown as an asterisk whose X-coordinate indicates the term's document frequency and whose Y-coordinate is either 1 or 0 depending upon whether the document is relevant or not relevant to the query. The logistic regression methodology fits an S-curve to this array of data points in the manner shown, where the height of the curve at any given X-value represents the estimated probability of relevance for that X-value.

This curve conforms to the following equation, shown here in its general form for any number M of independent variables:

$$P(R/<X_1,\ldots,X_M>) = \frac{e^{c_0 + c_1 X_1 + \ldots + c_M X_M}}{1 + e^{c_0 + c_1 X_1 + \ldots + c_M X_M}}$$

This logistic regression equation specifies the maximum likelihood solution based on the set of points in the learning sample. Conveniently enough, taking the logodds of both sides reduces this equation to the earlier Eq. (2), as the reader can readily verify. Since it is a logodds estimate

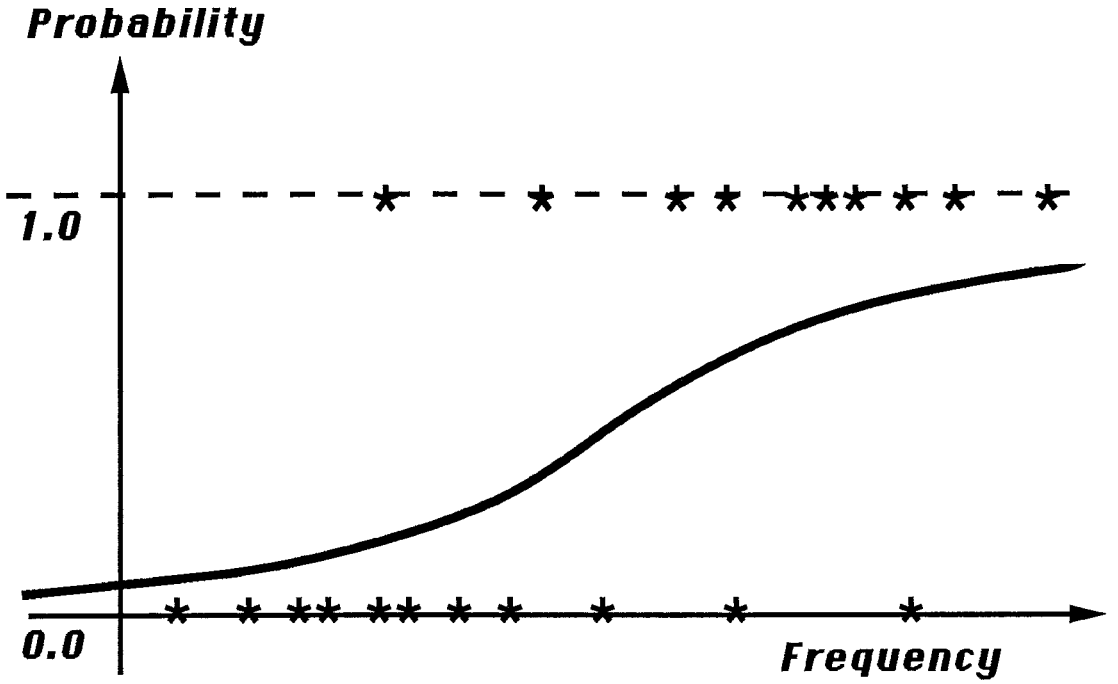

Figure 1. Graph of a hypothetical logistic regression equation. The height of the curve indicates the estimated probability of relevance for clue values along the X-axis.

rather than a probability estimate that is wanted anyway for substitution into Eq. (1), the exponentiation in the above formula need never actually be performed. The simpler linear formula of Eq. (2) is all that is needed for computational purposes.

From Fig. 1 it can be seen why standard (nonlogistic) linear regression would be inappropriate. A standard linear regression analysis would produce a sloping straight line cutting across both horizontal lines of data points; clearly such a straight line would be a less appropriate fit to the data for purposes of probability estimation than an S-curve. Moreover, the awkward possibility would arise that some probabilities might be estimated to be smaller than zero or larger than one -- a logical absurdity.

As if this were not enough, standard regression theory requires that within any given region of values of the independent variables, the differences between the actual and predicted values of the dependent variable be normally distributed. As can be seen from the figure, this assumption is far from true in the present application, where the deviations actually follow a binomial distribution. Also, in the standard theory the variance of these deviations is assumed to remain constant over the full range of the independent variables -- another dubious assumption. Finally, there is no good reason to suppose that for a dichotomous dependent variable minimizing the sum of the squared deviations or errors of prediction (as is done in standard regression) is an appropriate way to obtain a best fit.

The early chapters of Hosmer & Lemeshow's *Applied Logistic Regression* (1989) provide a readable introduction to logistic regression. Statistical packages with logistic regression capabilities include the most recent versions of BMDP, EGRET, Genstat, GLIM, S, SAS, SPSS, and SYSTAT. Collett (1991 Chapt. 9) offers a comparative survey of several of these.

Step 3: Correcting for Modelling Distortion

Having discussed how the quantities in the right side of modelling Eq. (1) might be estimated, it remains to consider how the equation as a whole might be modified to compensate for systematic biases caused by the Linked Dependence Assumption from which it was derived.

The Linked Dependence Assumption tends to inflate probability estimates produced by the modelling equation for documents near the top of the output ranking whenever the clues on which the estimates are based are strongly interdependent. This can be seen by considering the extreme case in which two clues are so strongly associated as to be coextensive. Recall that for two clues *A, B* the Linked Dependence Assumption can be formulated as the approximation

$$\frac{P(A,B/R)}{P(A,B/\sim R)} = \frac{P(A/R)}{P(A/\sim R)} \frac{P(B/R)}{P(B/\sim R)}$$

But if *A* and *B* are actually coextensive, the two fractions on the right are equal to each other and to the fraction on the left. The quantity on the left is being approximated by its own square! Although it would be unusual for two retrieval properties to be associated so strongly, the effects tend to be cumulative: the larger the number of clues involved, the greater the error of estimate is apt to be.

To help correct for this and other possible biases introduced by the simplifying assumption, the SLR technique calls for a second-level logistic regression analysis to be performed on the results of the first. A suitable regression equation for the purpose might take the form

$$Log\ O(R/A_1,\ldots,A_N) = d_0 + d_1\ Z + d_2\ N$$

Equation 3

where Z is the N-term summation expression in the right side of Eq. (1). The rationale for seeking a regression equation of this form is that the first part of the formula, $d_0 + d_1 Z$, might be expected to act as a simple corrective linear transformation on the raw estimate of Z. The third term $d_2 N$ is included because, as discussed, there is reason to suspect that the bias to be corrected tends to grow more severe as the number of composite clues to be combined grows larger. More elaborate equations, e.g. equations involving interaction terms, might also be considered; Eq. (3) is intended only as an illustration of the possibilities.

To find values for the coefficients d_0, d_1, and d_2 in Eq. (3), Eq. (2) is used to obtain an estimate of Z for all, or a sample of, the query-document pairs in the learning set. For all such pairs the values of Z, N, and the recorded relevance judgement are submitted to the statistical regression program, and maximum likelihood estimates of d_0, d_1, and d_2 are forthcoming. When during retrieval a query is submitted and a document is compared against it, Eq. (2) is first applied to evaluate each logodds in the sum in the right side of Eq. (1). Next, the sum of these is taken to obtain Z. Finally, Eq. (3) is applied to Z to obtain the adjusted estimate of the logodds of relevance for the document.

Further Elaborations

The method of staged logistic regression is an accommodating one that can be readily extended in any of several directions. Suppose for example that the uniform overall correction for term dependency bias offered by the equations presented in this paper is deemed inadequate, and an additional correction tailored to the individual degree of interdependency present in each particular case is desired. To obtain at least some predictive improvement along those lines, all that is needed is a way of measuring the average pairwise amount of term interdependency that is present among the terms involved in the analysis of any given query-document pair. Any of the well known measures of statistical association might serve this purpose, playing the role of an additional independent variable in the second-stage regression equation. Of course, such an elaboration might have a significant cost in terms of additional computational complexity.

The introduction of a relational thesaurus suggests further interesting possibilities. In the illustrative analysis of this paper, the function of the first-level regression equation was to estimate a logodds of relevance for any query-document-term triple for which the (same) term occurs in both the query and the document. But when a thesaurus is present, the lowest-level regression equation must estimate the logodds for query-document-term-term 4-tuples where the first term is from the query and the second a thesaurus-related term from the document. Depending on just how the details of the analysis are handled, this approach can lead to a regression procedure that is three or even four stages deep. For this reason we have referred to the general approach simply as 'staged logistic regression', leaving open the question of the number of stages that may be needed for any particular set of clue-types.

Speaking more generally, a two- or multi-stage SLR methodology would seem flexible enough to handle almost any type of probabilistic retrieval clue likely to be of interest: categorical as well as continuous independent variables, variables that describe document properties alone (recency, citedness, etc.), variables that describe query properties alone, and variables that describe the character of term nonmatches as opposed to matches.

Summary

We have attempted to sketch a flexible methodology for designing probabilistic retrieval rules -- one that offers the potential of yielding more reliable probability-of-relevance estimates than those attainable by many previous methods, yet not cumbersome at run time. Based on the technique of 'staged logistic regression' on a learning set or test collection, the method exploits a statistical simplifying assumption but corrects for the general upward bias it introduces. The approach is especially appropriate in retrieval environments in which the retrieval clues are grouped or composite, as in the case of subject term matches with several associated properties.

It remains to explore empirically the effectiveness of the methodology as compared with other retrieval methods, to demonstrate how it can in practice reduce the need for traditional full-scale testing, to test whether it is indeed capable of producing estimates of relevance probability that are sufficiently well-calibrated to be reported to the users without embarrassment, and to see whether it is robust enough so that parameters derived for one document collection and user population can in a pinch be used in other similar retrieval environments. We are presently investigating some of these issues, and invite others to join us in doing so.

Acknowledgements

Some of the ideas in this paper emerged from discussions with J. Allen, C. Buckley, S.

Robertson, G. Salton, and others, whose contributions are gratefully acknowledged. Part of the research was conducted in the supportive environment kindly provided by the Computer Science Department of Cornell University.

REFERENCES

Bookstein, A. Probability and fuzzy set applications to information retrieval. In M. Williams (ed.), *Annual Review of Information Science and Technology,* 20, White Plains, NY: Knowledge Industry Publications. 1985.

Bookstein, A. Implications of Boolean structure for probabilistic retrieval. *Proceedings Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Montreal, Canada: 11-17. June 1985.

Bookstein, A., M. Dillon, and D. Stephens. Application of loglinear models to informetric phenomena. *Information Processing and Management*, 28(1): 75-88; 1992.

Collett, D. *Modelling Binary Data.* London: Chapman & Hall; 1991.

Cooper, W. S. *The Inadequacy of Probability of Usefulness as a Ranking Criterion for Retrieval System Output.* Xeroxed report, School of Library and Information Studies, University of California, Berkeley, CA 94720, 1973.

Cooper, W. S.; Maron, M. E. Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association for Computing Machinery,* 25(1): 67-80; 1978.

Cooper, W. S. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, 34(1): 31-39; 1983.

Cooper, W. S. Probability theory as the basis of text retrieval. *Proceedings of the 54th Annual Meeting of the American Society for Information Science, vol. 28. Washington, D.C:* 366-369. October 1991a.

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 1991b.

Cooper, W. S.; Huizinga, P. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development,* 1(2): 99-112; 1982.

Fox, Edward A., *Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types,* Ph.D. Dissertation, Computer Science, Cornell University, 1983.

Fuhr, N. Optimal polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems* 7(3): 183-204; 1989.

Fuhr, N.; Buckley, C. Probabilistic document indexing from relevance feedback data. *ACM Transactions on Information Systems,* 9(2). 1991 (in press).

Fuhr, N.; Pfeifer, U. Combining Model-Oriented and Description-Oriented Approaches for Probabilistic Indexing. *Proceedings Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Chicago: 46-56. October 1991.

Gordon, M. D.; Lenk, P. A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the American Society for Information Science* 42(10): 703-714; 1991.

Harman, D. User-Friendly Systems Instead of User-Friendly Front-Ends. *Journal of the American Society for Information Science*, 43(2): 164-174; 1992.

Harper, D. J.; Van Rijsbergen, C. J. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3): 189-216; 1978.

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

Kantor, P. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Development,* 3(2): 88-94; 1984.

Lee, J. J.; Kantor, P. A study of probabilistic information retrieval systems in the case of inconsistent expert judgement. *Journal of the American Society for Information Science,* 42(3), 1990.

Maron, M. E. Probabilistic Retrieval Models. In B. Dervin and M. Voigt (Eds.), *Progress in Communication Sciences*, Vol. V, Ablex, 1984, pp. 145-176.

Maron, M. E.; Kuhns, J. L. On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery*, 7(3): 216-244; 1960.

Robertson, S. E. The probability ranking principle in IR. *Journal of Documentation:* 33, 294-304; 1977.

Robertson, S. E; Bovey, J. D. *Statistical problems in the application of probabilistic models to information retrieval.* British Library Research and Development Department, Report No. 5739, November 1982.

Robertson, S. E.; Sparck Jones, K. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3): 129-146; 1976.

Sparck Jones, K. (ed.) *Information Retrieval Experiment.* London: Butterworths; 1981.

van Rijsbergen, C. J. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2): 106-119; 1977.

van Rijsbergen, D. J. *Information Retrieval* (2nd ed.). London: Butterworth & Co. Ltd; 1979.

Yu, C.T.; Buckley, C.; Lam, H.; Salton, G. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2: 129-154; 1983.

Yu, C.T.; Salton, G. Precision Weighting -- An Effective Automatic Indexing Method. *Journal of the Association for Computing Machinery,* 23(1): 76-88; 1976.