ON THE LIMITATIONS OF DOCUMENT RANKING ALGORITHMS IN INFORMATION RETRIEVAL

Keith H. Stirling
School of Library and Information Sciences
Brigham Young University
Provo, Utah 84602

Introduction

A document retrieval system should rank documents in order of their usefulness or satisfaction to the users. This principle was first explicated in the classic paper by Maron and Kuhns (1). Additional considerations concerning document ranking have been suggested by other researchers (2,3). Particular attention will be given here to the ranking algorithm appropriate for those presenting the same request, but having different information needs. The research on which this report is based identifies limitations associated with sequencing rules that use a probability ranking technique (4). Three basic and somewhat interdependent limitations will be discussed.

Limitation #1

Ignoring the variations in information needs among users increases the amount of nonrelevant material encountered.

Although the Bayesian decision techniques use relevance feedback for estimating search-term weights, differences in information needs among users have generally been ignored. Instead, all user satisfaction data have usually been summarized into single relevance judgments. For an example, see (5).

This study compared the use of summarized with individualized relevance feedback as applied to ranking procedures. Results indicated that a ranking algorithm using summarized feedback can perform with only a ten percent increase in the nonrelevant material encountered when at least a 3:1 ratio occurs between types of information needs and number of retrieved documents. However, as the ratio decreases to 1:1, users must deal with an average increase of twenty-seven percent of nonrelevant material with no increase in the number of relevant documents acquired. If seven out of ten users judge a given document as relevant, summarizing the relevance judgments into one probability statistic does not differentiate among the $\binom{10}{7} = 120$ different ways 7/10 may be represented. Ranking the retrieved documents by summary statistics, generally, results in an increased amount of nonrelevant material encountered. See Figure 1.

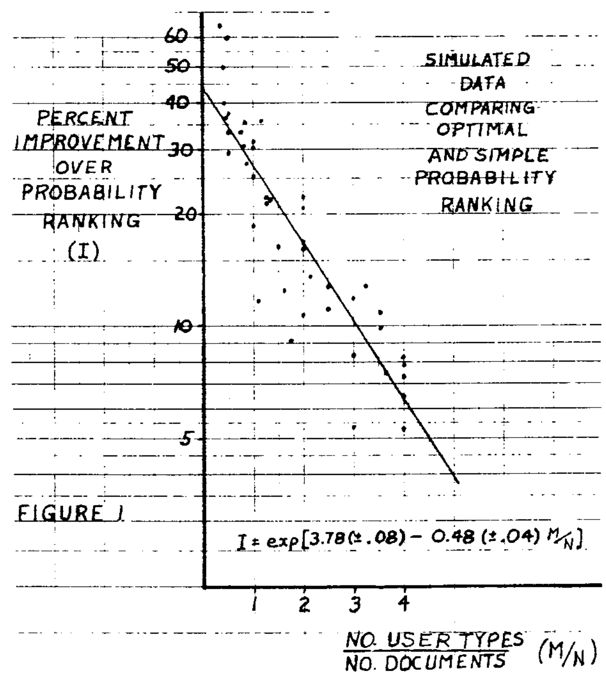


FIGURE 1

Limitation #2

Dual objectives are at work in document ranking algorithms. Ignoring this fact decreases performance.

An optimal ranking criterion for document retrieval systems has two objectives. The first is to maximize the probability that a user will obtain relevant material, and the second is to minimize the amount of nonrelevant material the user will encounter. Because of the conflicting information needs among the users, attaining the one objective does not automatically achieve the other. In fact, ranking the retrieved set of documents by decreasing probability of usefulness to the users will most likely increase the amount of nonrelevant material the users will encounter. This would also be true when using cosine methods or intersection/union ranking schemes. The only time that the simple probability ranking techniques satisfy both objectives is when

(1) no more than one relevant document is required to satisfy the information needs of any one user, or

(2) no more than two documents are retrieved for the request.

See Figure 2 for a comparison between optimal and simple probability ranking.

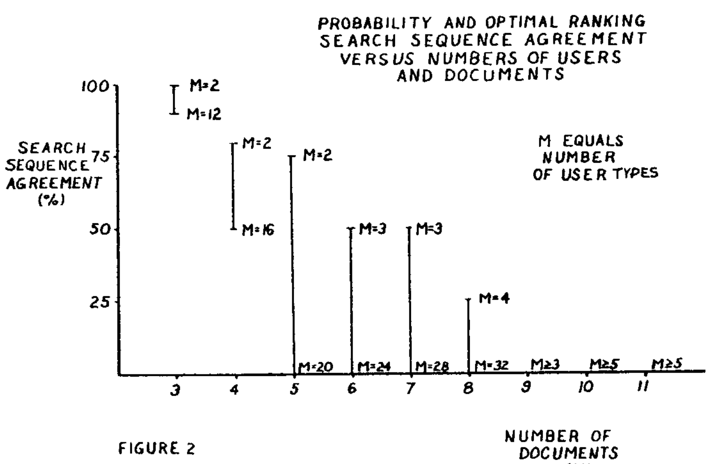


FIGURE 2

Limitation #3

<u>While all ranking algorithms imply the existence of a "stop search" point, most, if not all, ranking rules have ignored it. However difficult identifying it may be, the failure to do so degrades performance.</u>

When retrieval system users read the entire retrieved set of documents, any benefit due to ranking is negated. Therefore, each user type must have some sort of threshold level or utility value at which they stop considering any more retrieved documents. If the need for this "stop search" point is ignored, one of the objectives of document ranking, minimizing the number of nonrelevant documents encountered by the users, is lost. On the other hand, if the "stop search" requirement is included in the algorithm, some of the nonrelevant material will be witheld from the users' view.

To illustrate the "stop search" concept consider the following example. Nine different types of users representing different information needs have each submitted the same request to the document retrieval system. Five documents are retrieved. Assume that each type of user considers two of these documents relevant. The "stop search" point is reached when the two relevant documents have been encountered.

The documents in Figure A are ranked by decreasing probability of relevance, while the documents in Figure B are optimally ranked. Relevant documents are indicated by "x". The nonrelevant documents encountered prior to "stop search" points have been shaded.

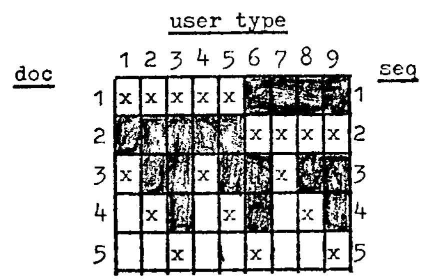


Figure A

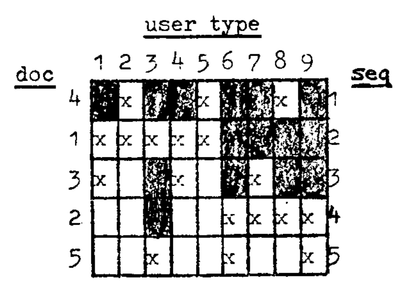


Figure B

Figure A has eighteen nonrelevant documents encountered prior to the users' "stop search" points, while Figure B has fifteen. The optimal ranking procedure used in Figure B provided a seventeen percent decrease in the number of nonrelevant documents encountered.

The optimal document ranking algorithm used here is computationally complex and is similar to the classic traveling salesman problem which requires exponential time or space for its solution. However, in the traveling salesman problem the "costs" encountered by a salesman in a given city are known before the tour. In the optimal document ranking algorithm the costs associated with a nonrelevant document either occur or do not occur depending upon whether the nonrelevant document is encountered by the user. Because of the complexity of the optimal algorithm, criteria are needed to determine when its use is cost effective.

To determine cost effectiveness, the number of "cost" computations required by the probability ranking procedure is proportional to $N(M+\log_2 N)$, while the number of "cost" computations required by the optimal procedure is proportional to $NM(2^{N-1}-1)$. On the other hand, the differential benefit in using the optimal ranking procedure in comparison with the probability ranking procedure may be approximated by the regression equation $A = \exp[-2.94 -0.04(M) +0.283(N)]$. The latter expression predicts the number of nonrelevant documents each user type will avoid on the average due to the optimal procedure as compared with the probability ranking procedure.

Combining the expression representing differential benefit and differential cost, a performance measure, P, is defined as

$$P = C\begin{bmatrix}\text{nonrelevant}\\ \text{documents}\\ \text{avoided}\end{bmatrix} - \alpha\begin{bmatrix}\text{number}\\ \text{cost}\\ \text{computations}\end{bmatrix}$$

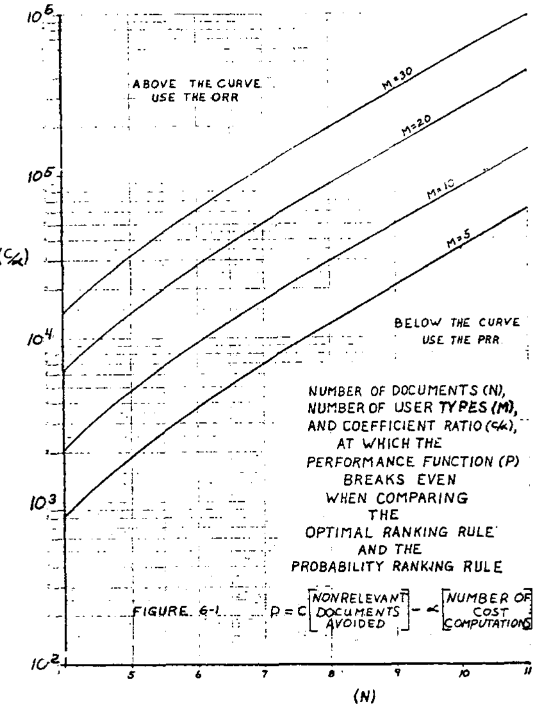
where C and α are constants. Figure 3 displays the breakeven point between the optimal and probability ranking procedures. For example, if six documents(N) were retrieved for ten types of users (M) then the value of C/α must be greater than 10^4 for the optimal procedure to be cost effective.

For the notion of optimal ranking to be considered, some assumptions were necessary.

(1) User feedback, regarding their relevance judgments, has been captured by the retrieval system.

(2) A steady state environment exists in which the system's data base, the type of user population, and the type of information requests are relatively stable.

(3) The relevance estimates from past users



can be used to rank documents for future users who may submit similar information requests. In addition, it should be noted that the optimal ranking procedure presented here is not restricted to binary relevance judgments. Any numerical evaluation is permissible.

If there is a real concern about maximizing the probability that users will obtain relevant material while minimizing the amount of nonrelevant material they encounter, then knowledge of differences in both information needs and "stop search" points must be considered. Using these criteria, this study provides a standard for comparing document ranking procedures.

References

1. Maron, M.E. and J.L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. JACM, 7, 1960, 216-44.

2. Robertson, S.E. The Probability Ranking Principle in IR. Journal of Documentation, 33:4, 1977,294-304.

3. Kraft, D.H. Stopping Rules and Their Effect on Expected Search Length. Information Processing & Management, 15, 1979, 47-58.

4. Stirling, K.H. The Effect of Document Ranking on Retrieval System Performance: A Search for an Optimal Ranking Rule. (Unpublished dissertation) Berkeley, Ca.: University of California, School of Library and Information Studies, 1977.

5. Robertson, S.E. and K. Sparck Jones. Relevance Weighting of Search Terms. JASIS,27:3,1976,129-146.