Division of Research
Graduate School of Business Administration
University of Michigan

November 1985

PROBABILISTIC AND GENETIC: ALGORITHMS IN DOCUMENT RETRIEVAL

Working Paper No. 450

Michael Gordon
University of Michigan

ABSTRACT

Document retrieval involves inquirers issuing computer-processable queries in order to receive references to relevant documents. Commonly, operational document retrieval systems fail to furnish many relevant document references while also furnishing many that are not relevant. Probabilistic models of document retrieval address this lack by providing a theoretical basis for deciding which documents to retrieve.

In implementation, however, probabilistic document retrieval is more problematic since it relies on unsatisfactory statistical assumptions and is based on questionable means of estimation. This paper proposes using a novel, adaptive approach to document retrieval which is inspired by the difficulties in implementing probabilistic models. The approach associates a competing set of descriptions with a document and employs a probabilistic, "genetic" algorithm which alters this set according to the queries used and relevance judgments made in actual retrieval. A simulation experiment, discussed in full, indicates the effectiveness of this approach in screening relevant from non-relevant documents. There are no serious difficulties in implementing this approach, including those that arise in probabilistic models.

1. Introduction

Document retrieval involves inquirers issuing queries in order to receive references to relevant documents. Commonly, the record stored in a bibliographic database to describe a document consists of a set of subject terms or "keywords" (sometimes with associated weights). In this research, documents receive multiple descriptions in an attempt to resolve problems arising from different inquirers seeking the same document in dissimilar ways. A probabilistic algorithm is also used to adjust these descriptions and provide a better means of getting documents to just those inquirers who will find them useful. The algorithm is free of assumptions of subject term independence that weaken most probabilistic models.

The paper begins by discussing the problem of representing documents. Particular attention is paid to probabilistic models, primarily to see how the limitations in implementing these theoretically-based models can inspire other forms of document retrieval. Document redescription will then be considered, and a probabilistic approach to to redescription relying on multiple descriptions of documents will be described. The effectiveness of that approach will be reported and then discussed.

2. Representation

The number of documents stored in computer-accessible fashion is quickly growing due to convenient, inexpensive word-processing and subscription document retrieval services housing large numbers of document references. As a result, both demand and need for effective document retrieval systems are on the rise.

Document retrieval is primarily a problem of representation: representing documents by storing some form of description of them in a database; and representing inquirers' information needs with computer processable queries. If forming adequate representations of both documents and users' needs were completely understood, there would be no need for further research in this field. Inquirers would easily find just those documents useful to them. Instead, providing adequate document representations and adequately expressing an information need so as to have success in retrieving textual information are actually quite difficult. Zunde and Dexter have documented the inconsistency among trained experts in trying to describe identical documents [20]. Blair points out that, if trained experts disagree so considerably in representing documents, one cannot hope for inquirers to be any more consistent in trying to retrieve them [1]. And, as many people are personally aware, the results of performing a computerized literature search can often be disappointing. Recently, a report of less than satisfactory full-text retrieval effectiveness has appeared [2].

Because of the difficulty of representing documents adequately, various theories have been advanced describing how documents might be best described. For instance, by accounting for the discrimination value of a term, it is argued that certain terms will make it easier to distinguish relevant from non-relevant documents while others will make it more difficult [16]. Recently, models formally incorporating probability have been described. Cooper and Maron give a utility-theoretic argument for deciding whether a term be used to describe a document: use the term if and only if retrieval satisfaction will be better with the term supplied than without [6]. Harter claims the number of times a term occurs in a document is a theoretically justified indicator of the term's suitability in describing a document [10]. Bookstein and Swanson explore the use of Harter's model as a decision-theoretic tool for indexing documents [3].

Indeed, models based on probability and decision theory are appealing because of their theoretical nature. Two variants of probabilistic models will now be discussed. A discussion will point out certain difficulties in implementing these models that, if overcome, might lead to enhanced retrieval performance.

In the Maron and Kuhns probabilistic model, each point in the sample space is a triple indicating: the query an inquirer would use on some occasion, the name of a document that might potentially be furnished for the query, and the relevance evaluation the inquirer would give concerning that document [13]. Thus, the events of primary importance in the model are of the form:

$P(\text{document}_i \text{ is relevant} \mid \text{particular query } Y)$,

Y being a non-empty set of query terms, $\{y_1, y_m, \ldots, y_n\}$ an inquirer might use to look for documents. In words, what is suggested is that, at different times, different relevance assessments of the same document will be made for identically constructed queries; and what is to be predicted is the probability that any given document will be relevant to such a query in the future. Further, the model prescribes a method for indexing documents.

Whereas the model just described relies on past inquiring behavior in estimating probabilities that inquirers will find documents relevant, the second probabilistic model relies on the keywords (i.e., subject terms) used to describe a document to estimate the same probabilities [14]. In this model, sample points indicate: the subject terms with which a document is described and the relevance judgment for that document with respect to an implicit, undescribed query. The events of importance have this form:

$P(\text{document}_i \text{ is relevant} \mid \text{document}_i\text{'s description is } X)$, where X is a set of binary keywords. We refer to this model as the Robertson and Sparck Jones model for the work they did in deriving query weights [14], noting historical antecedents to this approach in Bookstein and Swanson [3]. Van Rijsbergen more recently further articulated the Robertson and Sparck Jones model [18]. Croft has extended the model to the case where keywords are assigned probabilistically to documents rather than deterministically [7].

Under both the Maron and Kuhns and the Robertson and Sparck Jones models, Bayes' theorem is invoked to calculate probabilities. For instance, the latter ultimately ranks documents according to

$P(X \mid Rel) / P(X \mid {\sim}Rel)$ or, identically, by the ranking

$P(X \mid Rel) / P(X)$.

(X, Rel, and ~Rel indicate the events that a document is described by the set of subject terms, X; a document is relevant; and a document is non-relevant, respectively.)

Whereas the attractiveness of probabilistic models comes from their theoretical grounding, their implementation presents certain difficulties. First, in computing the probability that a document is described with the set of subject terms, X, the assumption is usually made that these terms are **distributed independently** within both the set of relevant documents and the set of non-relevant documents. (An independence assumption is also used in the Maron and Kuhns model.) The assumption buys mathematical tractability, but as Van Rijsbergen, among others, has pointed out, the assumption is discrepant with the fact that it is precisely subject term combinations (statistical dependencies among subject terms) that indicate the content of documents [19]. Therefore, even though experiments report that probabilistic models based on this independence assumption can improve retrieval performance [8], efforts have been made to devise simple dependency models less restricted by the independence assumption [4] [19]. However, such models require costly, computationally inefficient calculations, and also neglect dependencies among more than two terms [17].

A second problem that arises in probabilistic models is **estimation.** The Maron and Kuhns models relies on human estimations of the probability of using a given search terms in retrieving and finding a document relevant. Thus, it is hampered by the same problems in its implementation as manual retrieval models: inconsistent and possibly unreliable human judgments.

The Robertson and Sparck Jones model, on the other hand, makes empirically-based estimates of the distribution of terms within the relevant and non-relevant sets of documents. Given a small number of retrieved documents (usually ten or twenty) and an inquirer's relevance assessment of them, frequency data are used to calculate both p_i's (probability that term_i is used in indexing a relevant document) and q_i's (same probability for a non-relevant document). The model relies on this same pattern of subject term distributions existing in the collection as a whole so that relevant and non-relevant documents may be distinguished.

Such estimations of subject term distributions may be misleading, however. First, what should be sought is a population parameter for p_i's and q_i's. If documents are well described by subject terms, documents which are likely to be found useful together will be similarly described and retrieved together by inquirers searching for information. However, individual differences and inconsistencies among inquirers suggest it will never be possible to divide a document collection (here is one set of documents that will be useful together; here is another; etc.) in the same way for all inquirers. At best, then, subject terms can establish sets of similarly described documents which, collectively, are likely to be useful to inquirers "as a whole." Therefore, if there is a pattern described by subject terms, it is more likely to serve to demarcate relevant and non-relevant documents for the population of inquirers in general than for any individual making a query. Yet, in implementation, the Robertson and Sparck Jones model demarcates relevant and non-relevant documents based on the assessment of a single inquirer. To the extent that this separation is an artifact instead of indicating some pattern existing in the document database at large, it will not lead to useful prediction for retrieval. In addition, the point estimates we obtain for parameters such as p_i and q_i may be quite unreliable due to the small sample sizes on which they are based. In document retrieval, as Van Rijsbergen points out, "it should now be apparent that the problem of estimating binomial parameters from small samples has no obvious best solution [19]."

A probabilistic model unifying the Maron and Kuhns and the Robertson and Sparck Jones models has been proposed, too [15]. The model suggests the distribution of query terms together with document subject terms should provide evidence for retrieval. In this model, X is a set of binary random variables, x_i, each of which indicates, for arbitrary document_x, either that it is described by term_i (i.e., $x_i(\text{document}_x) = 1$)) or is not ($x_i(\text{document}_x) = 0$). Y is a set of binary random variables, y_i, which indicate which terms have been employed in a particular query. Probabilities of the form

$$P(\text{Rel} \mid X, Y)$$

are to be calculated in making a retrieval decision about a document. In practice, by Bayes' rule, this requires knowing the joint distribution of X union Y.

In a life-sized document retrieval system, document descriptions and queries can each be composed from vocabularies containing thousands of terms.

Even in principle, such a joint distribution of subject and query terms is only available at **exorbitant cost** by analyzing all possible queries in conjunction with any possible document description. Similarly, a Bahadur-Lazarfeld expansion of a joint distribution of 1,000 terms exhibiting third-order dependencies requires the estimation of over 166,000,000 parameters [17]. Thus, any feasible implementation of this model would again rely on less than realistic independence assumptions.

To summarize then, if a (probabilistic) retrieval model is to provide satisfactory document retrieval we should hope to see the following:

1. the model should not rely on independence assumptions;
2. feedback data should not be based on single inquiries (or a small set of inquiries);
3. the computational cost of the model should be acceptable.

In the sequel, a means of retrieval which satisfies these objectives will be described.

3. Document redescription

Since describing documents well is so difficult, one way to improve document descriptions is to perform the description process repeatedly. Each repetition would attempt to improve the description attached to a document. In essence, document redescription is an attempt to determine from past inquiries how a document should have been described so that its description can be modified and made more satisfactory to future inquirers. Such an approach is based on the assumption that there will be identifiable regularities in a large enough set of inquirers' requests for a given document.

Other attempts to modify a document's description has been reported. Brauen adjusts document term weights to improve retrieval effectiveness [5]. Successful to some degree, his approach does worse when it receives feedback concerning both successful and unsuccessful searches than when it receives only the former, and "control" queries exhibit better recall-precision performance as a result of document modification than the "test" queries toward which document redescription is directed. More recently, Furnas has described an adaptive indexing system that learns alternate terms to use in identifying various sources of information (i.e., "documents") [9]. In essence, once a sought document is identified, the system uses all the (single-term) queries that searcher has used unsuccessfully in trying to locate the document to update frequency counts relating it to these terms. Furnas' approach has only been applied to very small databases (255 documents in the system most extensively studied) and only permits analysis for single-term queries. Clearly, customary bibliographic retrieval is of a quite different character requiring correspondingly different approaches.

4. Genetic adaptation of documents

We have seen that representing documents so that they may be effectively retrieved is difficult. Further, we have noted that an attempt can be made to make document description an iterative process, but that research results have not conclusively established the viability of this approach. In this section, we will see a means of adaptive document redescription that provides improved retrievability of relevant documents. Some initial remarks describe the philosophy underlying the approach before it is described. In following sections, results of simulation experiments documenting the success of the approach are reported and comments on using the approach to represent documents are made.

As was pointed out, inquirers are bound to disagree about the proper description of a document (that is, they'll each be requesting it somewhat differently), so the best indexing results will arise from describing a document to best represent it for the group of inquirers who will find it useful (rather than for any particular individual). As a result, a novel approach to indexing used here involves simultaneously supplying alternative descriptions to the same document and then deriving better descriptions based on feedback indicating which of the alternative descriptions best describes the document. The retrieval model, (its use of feedback momentarily ignored), is this: a document is described by several complete descriptions (for instance, several sets of keywords, or several sets of keyword weights); an inquirer issues a query; and each description of the document is matched against the query as if the document were described with only a single description. The average of these separate matching scores serves as the basis for retrieving or not retrieving a document. (Other functions of these separate matching scores are possible, too, possibly the maximum or median.)

As mentioned, the retrieval model makes use of feedback to change the way a document is represented, rather than keeping the representation static. This is done with a "genetic" (or adaptive) algorithm which is of demonstrated success in many domains [11] and currently is being studied in artificial intelligence research aimed at promoting learning [12]. Consider a set of objects, each of which is performing an identical task, and assume each object can be represented by a string of symbols. The genetic algorithm operates on such a set of representations, replacing it with another set, then another, and so on. The replacement attempts to produce new sets of objects (more precisely, object representations) in each succeeding "generation" which, on the whole, are more fit (perform the designated task better) than their predecessors.

The algorithm, applied to the task of document redescription, iterates this two-step process:

```
Repeat
    1) For any particular document, measure the "worth"
       (i.e., "performance" or "fitness") of each of its
       (fixed number of) descriptions. That is, determine
       how well each description serves in providing the
       document to just the right inquirers.
    2) Replace the set of descriptions currently associated
       with that document:
        a) Throw away its current set of descriptions
        b) Establish a new set of descriptions out of the
           set just discarded, using more "parts" of
           descriptions which had higher worth. Each of
           these descriptions will likely be different from
           all descriptions in the just discarded set.
Until some criterion is obtained.
```

In other words, what is occurring is a process that attempts to mimic genetics, promoting a population of descriptions built up of parts ("genes") of its fittest members. The first step in the process seeks to determine which descriptions are best doing their job (getting a document to just those inquirers who will find it relevant); the second step exploits the information gained in the first but also introduces variety. Together, the two steps seek regularities among the best descriptions, promote descriptions exhibiting such regularities in succeeding generations, and try out completely novel descriptions in an effort to improve upon the descriptions already tested.

Even after adaptation, a document retains a set of descriptions. If deemed desirable, these could then be replaced by a single description derived from this set.

With this background, the results of performing adaptive redescription are now described, and the details of the algorithm are left to the Appendix.

5. Results

The effectiveness of applying the genetic algorithm to document redescription was tested experimentally. The basic paradigm assumes the same set of queries (with identical relevance judgments) is repeatedly issued to the retrieval system. The algorithm uses only knowledge indicating the queries to which a document is relevant and the queries to which it is not in adapting document descriptions. Thus, to adapt the way a document is described, it was necessary to collect a set of "relevant queries" (queries to which a document is judged relevant) and "non-relevant queries" for each document studied. The judgments providing these query sets were made by undergraduate college students. In all, a "relevant query set" and "non-relevant query set" were obtained for each of eighteen different documents. On average, there were seventeen descriptions per document and a like number of both relevant and also non-relevant queries. (Each was a set of unweighted terms without any Boolean connective.) Using these, a series of eighteen separate simulations was conducted:

For any given document, an initial set of descriptions was needed to begin the simulation. The same college students used to make relevance judgments about documents provided these initial descriptions. In fact, the initial set of document descriptions and the set of relevant queries for that document were identical, the assumption being that the query one puts to find a relevant document and the description one would provide for that document ought to be the same. Each of the eighteen documents studied was treated independently, meaning it had its own set of initial descriptions and its own set of relevant and non-relevant queries.

A snapshot of generation_g of the simulation for hypothetical document_x can be seen in Figure 1. Several things should be noticed: One, there are N generation_g descriptions of this document. (Each of these is a set of subject terms.) In generation_1, these were the descriptions originally supplied. Subsequently, they will have been modified. Two, there is an associated set of M relevant queries for this document. Three, a Jaccard's score matching function is used to compute the similarity of every description of document_x to each of its relevant queries. (The Jaccard's score association between two sets X and Y is #(X intersect Y) / #(X union Y), #(S) meaning the cardinality of set S. In this case X and Y are the set of terms used to describe a document and the set of terms used in posing a "relevant" query for that document, respectively. The Jaccard's score is a common measure of association in document retrieval [18], and use of other association measures would not influence the operation of the simulation or the expected results.) Four, notice that the Average Matching Score, i.e., "worth," of each of the descriptions currently in force is indicated. It is this "worth" that is exploited by the genetic algorithm in producing descriptions in generation_{g+1}. Finally, notice that the overall level of association (the Overall Average, G_g) between descriptions in use during the current generation, g, and the set of relevant queries is indicated. It is this statistic that indicates how well the current generation of descriptions is performing its job.

If genetic adaptation were to succeed in improving document descriptions, then, on average, the level of association between a document and its relevant queries should be higher in generation 40, upon completion of the simulation, than it was in generation 1 using the original set of descriptions. That is, G_{40} should exceed G_1. Such improvement did, in fact, occur for each of the eighteen documents studied (averaging approximately 25% improvement in Jaccard score). Further, although a redescribed document exhibited some increase in similarity to its non-relevant queries (used as experimental control), in seventeen of eighteen cases this was less than the increase relative to relevant queries, which, on average, was nearly five times as great. See Figure 2 and Table 1.

In a second series of simulations, the adaptive procedure was put to a more severe test: Redescription was attempted to raise the document's average level of association to its "relevant queries" and, at the same time, to reduce the document's average level of association to its "non-relevant queries," (which were selected on the basis of their similarity to relevant queries). In other words, the attempt was made to redescribe a document so that it would more likely be retrieved by those who would find it useful and less likely retrieved by those who would not. Again, each of the eighteen documents was adaptively redescribed independently of all the others. As before, these simulations were run for forty generations.

Figure 3 provides an in-progress snapshot of the redescription of hypothetical document_x in conformance with the goals above. Notice that a document has one set of descriptions associated with it, but that these will be matched against both relevant queries and non-relevant queries. G_g, as before, indicates how well, on average, the prevailing descriptions of the document are at matching relevant queries. G'_g makes that same indication for non-relevant queries. For adaptation to succeed, G_g should increase from its generation_1 level while G'_g should fall: in practical terms, this would mean that documents are more likely to be retrieved by relevant queries (since they are now described more similarly to these queries) and less likely to be retrieved by non-relevant queries (due to decreased similarity). Should G_g and G'_g both rise, adaptation still might be deemed successful if the increase in the former is greater than the latter. The advantage, in that case, is that there is now a greater difference between the expected level of association between a document and a relevant query and the expected level of association between that same document and a non-relevant query. Thus, when presented with a query of uncertain relevance, it becomes easier to tell whether or not to retrieve the document.

In conduct, G_g did increase as a result of adaptation for each of the eighteen documents studied. In fifteen cases out of the eighteen, G'_g dropped (signifying that, absolutely, non-relevant queries matched worse a document when it was redescribed to match its relevant queries). In the remaining three cases, G'_g rose slightly, but less than the corresponding rise in G_g for the same document (signifying that, relative to the increased association between relevant queries and adapted document descriptions, an improvement in filtering non-relevant queries was made). See Figure 4 and Table 2.

6. Discussion

Earlier, we suggested that document retrieval can be improved if: 1) the underlying model is not reliant on independence assumptions; 2) the feedback pertaining to relevance assessments that is elicited is aggregated across a sufficiently large group of inquirers with similar information needs; and 3) these two criteria are attainable at reasonable computational cost. Success in meeting each of these criteria is now examined for the model of genetic adaptation described.

For any document in the simulation, the distribution of query terms over its set of relevant queries is easily tabulated. Let the proportion of relevant queries (for a given document) using term_i be p_i. Then, one procedure for indexing a document employing a set of M descriptions suggests that this set employ subject term_i for p_i*M of its descriptions and not employ term_i in the remaining $(1-p_i)$*M cases. Under the assumption that the distribution of any term_i is independent of any term_j, (j <> i), and by making use of complete knowledge of all relevant queries used to retrieve a given document, we would obtain identical distributions of subject terms in both the set of descriptions used to describe a document and the queries employed in an attempt to retrieve it. Such a "theoretically derived" set of document descriptions might be considered near optimal on the assumption that maximal similarity between a set of document descriptions and its relevant queries will arise when every term is distributed identically in both sets.

The actual effectiveness of describing a document with such a "theoretically derived" set of descriptions was compared to the effectiveness of genetically adapting descriptions for the same documents. The results pointed out the superiority of adaptation: for each of the documents studied, the adapted descriptions more effectively matched relevant queries than did the "theoretical" descriptions, the improvement averaging approximately 25 percent (as measured by Jaccard's score).

The genetic algorithm is responsible for this improvement. By its action, combinations of index terms which best serve to describe a document, rather than just individual terms, proliferate from generation to generation. More technically, in promoting competition among a set of objects represented as strings of symbols, and then introducing variability after reproducing these strings in proportion to their effectiveness, there will be increasing representation of the fittest "schemata" over time. (A schema names a hyperplane, or set. For instance, the binary schema 01#001# stands for the set of four 7-place strings {0100010, 0100011, 0110010, 0110011}, "#" standing for "instantiate in any valid way possible.") In the case of document descriptions, fit schemata are those subject term combinations that best describe documents. Thus, descriptions of documents are built up out of index term combinations quite differently, and more effectively, than if index terms were supplied independently to documents. This is consonant with Holland's commentary on the proof of the genetic algorithm: "In effect, useful linkages are preserved and non-linearities (epistases) are explored ... giving a performance ... which is not simply the sum of their [for us, subject terms'] individual performances." And, with a suitable number of descriptions attached to a document, by the central limit theorem, sets of subject terms "with the higher average fitness quickly predominate [11]." Or, again, the genetic algorithm produces document descriptions which surpass in performance those that could be generated from identical information using assumptions of statistical independence. The operation of the genetic algorithm differs in this respect from other document retrieval feedback techniques (used to alter document descriptions or queries) which modify the weight of any given subject or query term independently of all others.

A second desideratum of a document retrieval system is that it collect feedback with which to base retrieval on knowledge gained about "inquirers as a whole," rather than on the basis of an individual query or inquirer. This criterion, which accounts for individual difference among inquirers, underlies the operation of the system described, as the current descriptions associated with a document are derived as a result of relevance assessments of the document issued by all past inquirers who were furnished it.

In being independence-free and describing documents to meet the needs of a population of inquirers, the model has theoretical value. In addition, the computational cost of the model does not outweigh its advantages. In supplying a document with several (say n=15) descriptions rather than one, we suffer a linear (i.e., fifteen-fold) increase in both the storage required to store descriptions as well as the time required to match queries to document descriptions. Although the increase in storage is inevitable, one expects that efficient means of compressing descriptions and decreasing storage costs could mitigate this disadvantage. (It is possible, too, to multiply describe only the most actively sought documents in the collection or to replace the set of descriptions with which a document is represented by a single, "consensual" description once adaptation is complete.) Importantly, notice that by incurring this increased storage cost, we obtain a distribution of subject terms within the description set of a document which is more effective than that "theoretically derived" using an independence assumption together with complete information about relevant queries. In short, we are accomplishing what, in probabilistic models, is so difficult: obtaining effective "probability" estimates which can be used to improve the retrieval of documents.

The matching costs, potentially more damaging, are more easily dealt with in practice. With clustered files of document descriptions, or by means of file index construction, the documents deemed potentially relevant to a query can be restricted to a greatly reduced fraction of the entire database. Thus, although we have a linear increase in the number of matches to be made, constraining the search to a small subset of the database considerably lessens this concern. (It is possible, too, to match a query against a few, not all, of the descriptions of the set of documents initially suspected to be relevant. Then, using these matching scores, the retrieval system can continue to match against all descriptions of just those documents indicating greatest likelihood of relevance.) Research into improving the time and space complexity of genetic adaptation of document descriptions should, of course, precede implementation of a system based on this model.

7. Conclusion

The difficulty in describing documents well has been indicated. Probabilistic models of document retrieval have been discussed to indicate the difficulty in achieving theoretical soundness along with effective implementation. Although less theoretically guided than strict probabilistic retrieval models, an adaptive approach has been described which overcomes the problems probabilistic models suffer: 1) implementations based on independence assumptions; and 2) probability estimations either of uncertain reliability or value or attained only at prohibitive cost. (We note, too, that even rigorous probabilistic models are, in practice, heuristic to a considerable extent. For instance, the Robertson Sparck Jones model formally calculates probabilities of relevance without regard to queries at all! In practice, though, queries are used heuristically to restrict computations.) The effectiveness of the adaptive approach has been documented.

In an operational retrieval system, an initial set of descriptions for a document could be obtained by means of competing automatic indexing procedures. Alternatively, models suggesting the probable effectiveness of employing given subject terms to documents could be used to stochastically generate an initial set of descriptions.

The described simulation experiments model a document being repeatedly requested by the same set of queries. In actuality, the way a document should be described is likely to change over time as it is requested differently. Being a model that redescribes documents rather than leaving their descriptions fixed, the adaptive model discussed automatically accommodates such changes. Since the model employs multiple descriptions of any document, one of these could include just those query terms from a recent query to which the document was, or should have been, judged relevant. In this way, entirely new subject terms can be incorporated in describing a document. In the model presented, what is occurring is that those terms already supplied are being more effectively distributed across the various descriptions of the document.

APPENDIX

The genetic algorithm repeats the two-step process already outlined in an attempt to provide increasingly effective document descriptions over time:

```
Repeat
  1) Measure the performance of competing document
     descriptions
  2) Replace the set of descriptions
Until some criterion is attained.
```

Figure 1, in the text, helps explain the details of the algorithm as used in this study. Each of the generation_g descriptions of document_x shown is really a binary document vector. For example, we might have:

$$\begin{array}{lcccccccccc} & & T_1 & T_2 & T_3 & T_4 & . & . & T_k & \\ \text{desc\_x\_g}_1 = & < & 1 & 1 & 1 & 0 & . & . & 0 & > \end{array}$$

where each of T_1 through T_k is a subject term (or phrase) that is either being employed in describing a document (1) or is not (0).

Both of the steps in the algorithm above are now more completely explained. For a proof of the effectiveness of this class of algorithms under various conditions, see Holland [11].

1) Measure performance of competing descriptions

The Average Matching Score for each description is indicated in the right most column of Figure 1. This measures how well each competing description "performs" (matches, on average, the M relevant queries for this document). We call this a description's "fitness."

2) Replacement of the set of descriptions

a) **Relative Fitness:**

Calculate, for each description, $desc\_x\_g_i$,

Relative_Fitness ($desc\_x\_g_i$) (1 <= 1 <= N).

Relative_Fitness ($desc\_x\_g_i$) =

Avg Matching Score($desc\_x\_g_i$) / F

where

$$F = (1/N) * \sum_{m=1}^{N} \text{Avg Matching Score } (desc\_x\_gm)$$

b) **Reproduction:**

Create Relative_Fitness ($desc\_x\_g_m$) copies of $desc\_x\_g_m$ (1 <= m <= N)

Treat fractional relative fitnesses stochastically.

Discard $generation_g$ descriptions.

c) **Cross-over:**

Randomly partition this newly created set of N descriptions into floor (N/2) pairs (plus a single remaining description if N is odd).

For each pair, j, pick a random cross-over point, p_j, 1 <= p_j <= k - 1 (k the length of the vector). Form the $generation_{g+1}$ set of document descriptions as follows (set initially empty):

Add to set:

initial (desc-pair j_1) + final (desc-pair j_2)

initial (desc-pair j_2) + final (desc-pair j_1)

where

desc-pair j_1 and desc-pair j_2 are the pair of document descriptions in the j-th pair;

initial (desc-pair j_t) = first p_j positions in vector (desc-pair j_t (t = 1,2)

final (desc-pair j_t) = last $(k - p_j)$ positions in vector desc-pair j_t (t = 1,2)

+ = string concatenation.

For odd N, remove a randomly chosen description from the set just generated. Pair it with the as yet unpaired description. Apply cross over to this additional pair and place this newly created pair into set.

For instance, if the following two subject descriptions (below, left) comprise the j-th pair, and p_j is randomly selected to be 3, we would see

Before crossover							After crossover						
T_1	T_2	T_3	T_4	.	.	T_k	T_1	T_2	T_3	T_4	.	.	T_k
< 1	1	1	0	.	.	1 >	< 1	1	1	0	.	.	0 >
< 0	1	1	0	.	.	0 >	< 0	1	1	0	.	.	1 >

The new set of document_x descriptions would replace those in Figure 1, and the entire adaptive process would be repeated.

Note: Figure 3 presents a slightly more complicated situation, differing in its calculation of relative fitness. There, the fitness of any description depends on both its "recall" fitness (similarity to relevant queries) and its "fallout" fitness (dissimilarity to non-relevant queries).

That is, in Figure 3, the **fitness** of a document description, say desc_x_g_i, would be equal to:

(1) Avg Recall Matching Score (desc_x_g_i) +

$$wt * (G'_g - [\text{Avg Fallout Matching Score } (\text{desc\_x\_}g_i) - G'_g])$$

Three observations pertain to this formula:

1) The first addend, Average Recall Matching Score (desc_x_g_i), reflects the description's similarity to relevant queries.

2) The second addend reflects the description's dissimilarity to non-relevant queries. The term the G'_g - [Avg Fallout Matching Score (desc_x_g_i) - G'_g] is exactly the same magnitude above G'_g as Avg Fallout Matching Score (desc-x-g_i) is below it. This "inversion" is necessary so that descriptions good at matching relevant queries and descriptions good at not matching non-relevant queries both contribute in an "above average" fashion to the overall fitness the description. (That is, descriptions which are quite dissimilar to non-relevant queries should contribute "fallout fitness" values greater than G'_g.)

The **relative fitness** of desc_x_g_i was calculated to be:

$$\text{fitness } (\text{desc\_x\_}g_i)/(1/N \sum_{j=1}^{N} \text{fitness } (\text{desc\_x\_}g_j))$$

3) The weight, wt, in expression (1) was employed to balance the differing effects of a description's Avg Recall Matching Score (recall fitness) and "inverted" Avg Fallout Matching Score (fallout fitness) on its overall relative fitness. Some experimentation indicated a weight of 0.50 was appropriate to cause G_g to rise and G'_g to fall in succeeding generations.

	relev_x_q$_1$	...	relev_x_q$_M$	Avg Matching Score
desc_x_g$_1$	J(g1,q1)	...	J(g1,qM)	1/M $\sum_i$J(g1,qi)
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_g$_N$	J(gN,q1)	...	J(gN,qM)	1/M $\sum_i$J(gN,qi)

N descriptions
of document$_x$
in generation$_g$

Overall Average, G_g, =

$$\frac{1}{M * N} \sum_k \sum_i J(gk,qi)$$

Each of document$_x$'s M relevant queries is matched against each of the document descriptions in force in generation$_g$. The match between relevant query relev_x_q$_i$ and document description desc_x_g$_j$ is indicated by J(gj,qi). Row averages give "Average Matching Scores" for each document description. G_g, the grand average, gives the overall average matching score for the document descriptions in force in the current generation, g.

A set of descriptions of document$_x$ which produces an Overall Average matching score greater than G_g relative to the same relevant queries is an improvement on the generation$_g$ set of descriptions.

Figure 1--Matching of descriptions and relevant queries

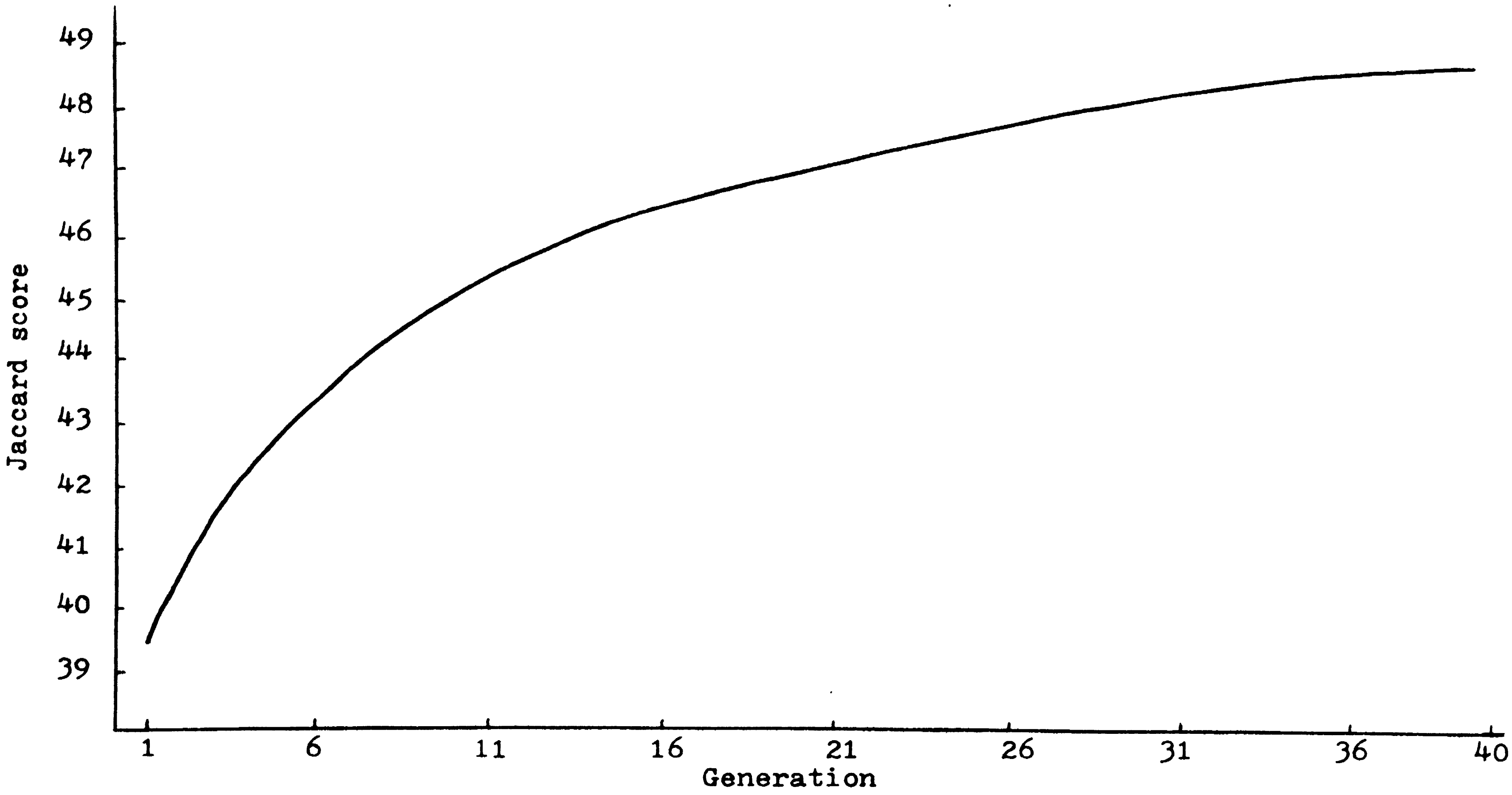

Figure 2 --Recall improvement--all documents combined

	relev_x_q_1	...	relev_x_q_M	Avg Recall Matching Score
desc_x_g_1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_g_N	J(gN,q1)	...	J(gN,qM)	$1/M \sum_i J(gN,qi)$

N descriptions
of document$_x$
in generation$_g$

Grand average, G_g, = $\frac{1}{M * N} \sum_k \sum_i J(gk,qi)$

`************************************************************`

	non-rel_x_q_1	...	non-rel_x_q_M	Avg Fallout Matching Score
desc_x_g_1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_g_N	J(gN,q1)	...	J(gN,qM)	$1/M \sum_i J(gN,qi)$

N descriptions
of document$_x$
in generation$_g$

Grand average, G'_g, = $\frac{1}{M * N} \sum_k \sum_i J(gk,qi)$

Each document description set is matched with each relevant query and also with each non-relevant query. For each document description, an average recall matching score is calculated with respect to the relevant query set (row averages above the starred line), and an average fallout matching score is calculated with respect to the non-relevant query set (row averages below the starred line) and then "inverted" around G'_g.

Note that, above the dotted line, J(gi,qj) indicates the Jaccard match between description desc_x_g_i and relevant query rel-x-q_j, whereas below the line it indicates the Jaccard match between the same description and non-relevant query non-rel_x_q_j. G_g and G'_g are calculated with respect to the pertinent queries.

Figure 3--Matching of descriptions with relevant and non-relevant queries

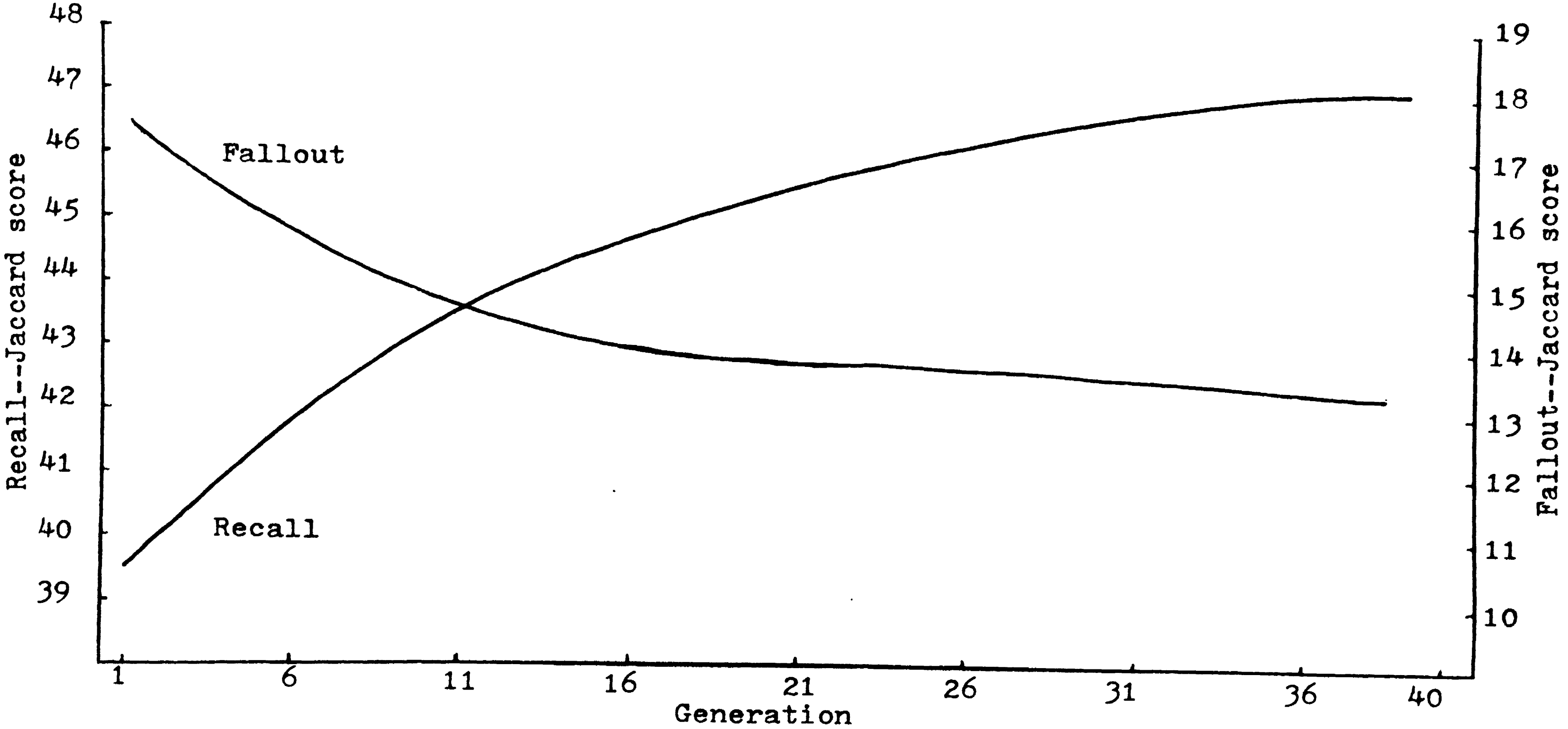

Figure 4 --Recall-fallout improvement--all documents combined

Document	Change in Overall Average Matching from gen_1 to gen_{40}	
	relevant queries	non-relevant queries
Doc 1	10.05	4.14
Doc 2	7.61	1.81
Doc 3	10.83	-5.10
Doc 4	13.11	8.51
Doc 5	8.79	4.75
Doc 6	9.11	0.96
Doc 7	10.38	0.08
Doc 8	8.01	-8.18
Doc 9	10.81	-3.29
Doc 10	8.06	6.87
Doc 11	8.51	11.43
Doc 12	9.79	3.05
Doc 13	11.12	2.94
Doc 14	7.56	5.41
Doc 15	9.11	3.48
Doc 16	10.69	1.91
Doc 17	9.17	1.69
Doc 18	5.62	-5.82
Avg.	9.35	1.92
S.D.	1.62	4.89

Data expressed in units of Jaccard's score.

The pair of table entries in a row (like 10.05 and 4.14 in row 1) indicate intentional and inadvertent improvement, respectively. That is, after Document-1 was redescribed for 40 generations, the Overall Average Matching score relative to its relevant queries was intentionally elevated by 10.05 Jaccard points; similarly, the same redescription inadvertently increased document-1's overall average matching score 4.14 points relative to a set of non-relevant queries.

H_o: For any document (table row), the greater change in Overall Average Matching is equally likely to occur with respect to relevant queries (intentional change) or with respect to non-relevant queries (inadvertent change).

Reject H_o, $p < .0001$, sign test;

Conclude: adaptation promotes greater recall improvement with relevant than non-relevant queries.

Table 1--Increase in overall average matching for non-relevant queries versus relevant queries

	RECALL			FALLOUT		
	Gen 1	Gen40	%Chng	Gen 1	Gen40	%Chng
Doc 1	36.03	42.86	18.96	20.07	14.47	-27.90
Doc 2	44.45	50.59	13.81	17.83	7.69	-56.87
Doc 3	42.19	52.53	24.51	17.12	11.05	-35.36
Doc 4	39.36	50.58	28.51	21.08	25.87	+22.72
Doc 5	41.12	47.33	15.10	18.83	17.58	- 6.64
Doc 6	43.01	52.45	21.95	18.00	16.04	-10.89
Doc 7	33.45	40.09	19.85	18.11	13.87	-23.41
Doc 8	31.81	39.98	25.68	12.92	4.28	-66.87
Doc 9	54.21	64.43	18.85	13.72	8.33	-39.29
Doc 10	37.92	46.65	23.02	17.65	13.25	-24.93
Doc 11	28.06	30.23	7.73	19.34	14.52	-24.92
Doc 12	48.15	57.72	19.88	16.88	18.45	+9.30
Doc 13	47.36	57.09	20.54	16.69	16.81	+0.72
Doc 14	39.95	44.29	10.86	20.29	13.75	-32.23
Doc 15	36.80	43.95	19.43	18.25	16.16	-11.45
Doc 16	39.83	47.64	19.61	17.88	13.03	-27.13
Doc 17	31.23	37.99	21.65	14.75	8.53	-42.17
Doc 18	36.66	41.68	13.69	16.35	8.31	-49.17
Average	39.53	47.12	19.09	17.54	13.44	-24.81

This table indicates the initial (pre-adaptation) level of association between a document and its relevant queries and its non-relevant queries, as well as final (post-adaptation) levels of the same measures. For doc 1, for example, we see that document redescription caused the average Jaccard's match between relevant queries and document descriptions to rise from 36.03 (before adaptation) to 42.86 (an 18.96% improvement). The same document redescription resulted in the average match between doc 1's non-relevant queries and document descriptions dropping from 20.07 to 14.47 (a 27.90% improvement).

H_o: For any document (table row), it is equally likely that either a) recall and fallout will both be improved because of adaptation or b) one or both of recall or fallout will not be not be improved.

Reject H_o, $p < .01$, sign test.

Conclude: adaptation simultaneously improves both recall and fallout.

Table 2--Recall-fallout improvement

ACKNOWLEDGEMENTS

The author thanks David Blair for his input during the course of this research. He and Manfred Kochen, both of the Computer and Information Systems area, The University of Michigan, are also thanked for their comments on various versions of this paper.

REFERENCES

1. Blair, D.C. Indeterminacy in the subject access to documents. *Info. Proc. and Mgt*., forthcoming.

2. Blair, D.C. and Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Comm. Assoc. Comp. Mach*., Vol. 28, No. 3, (1985), pp. 289-299.

3. Bookstein, A. and Swanson, D.R. A decision theoretic foundation for indexing. *J. Amer. Soc. Info. Sci*., Vol. 26, No. 1, (1975), pp. 45-50.

4. Bookstein, A. and Kraft, D. Operations research applied to document indexing and retrieval decision. *J. Assoc. Comp. Mach*., Vol. 24, (1977), pp. 410-427.

5. Brauen, T. Document vector modification. In *The SMART Retrieval System-Experiments in Automatic Document Processing*. G. Salton, ed. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

6. Cooper, W.S. and Maron, M.E. Foundations of probabilistic and utility-theoretic indexing. *J. Assoc. Comp. Mach*., Vol. 25, (1978), pp. 67-80.

7. Croft, W.B. Document representation in probabilistic models of information retrieval. *J. Amer. Soc. Info. Sci*., Vol. 32, (1981), pp. 451-457.

8. Croft, W.B. and Harper, D.J. Using probabilistic models of documentation without relevance information. *J. Documentation*, Vol. 35, (1979), pp. 285-295.

9. Furnas, G.W. Experience with an adaptive indexing scheme. In *Proc. of ACM SIGCHI Conf. on Human Factors in Computing Systems*, (San Francisco, CA, Apr. 14-18, 1985). ACM, New York, 1985, pp. 131-135.

10. Harter, S.P. A probabilistic approach to automatic keyword indexing. Part 1: On the distribution of specialty words in a technical literature. Part 2: An algorithm for probabilistic indexing. *J. Amer. Soc. Info. Sci*., Vol. 26, (1975), pp. 197-206, 280-289.

11. Holland, J. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, 1975.

12. Holland, J. Escaping brittleness. In *Proc. of the International Machine Learning Workshop*, (Monticello, IL, 1983).

13. Maron, M.E. and Kuhns, J.L. On relevance, probabilistic indexing, and information retrieval. *J. Assoc. Comp. Mach*., Vol. 7, No. 3, (1960), pp. 216-244.

14. Robertson, S.E. and Sparck Jones, K. Relevance weighting of search terms. *J. Amer. Soc. Info. Sci*., Vol. 27, (1976), pp. 129-146.

15. Robertson, S.E., Maron, M.E. and Cooper, W.S. Probability of relevance: a unification of two competing models for document retrieval. Info. Tech.: Rsch. Dvpt., Vol. 1, (1982), pp. 1-21.

16. Salton, G., Yang, C.S. and Yu, C.T. A theory of term importance in automatic text analysis. J. Amer. Soc. Info. Sci., Vol. 26, No. 1, (1975), pp. 33-44.

17. Tague, J., McClellan, C. and Nelson, M. The hyperterm model of a bibliographic database. Canadian J. Info. Sci., Vol. 9, (1984), pp. 37-58.

18. Van Rijsbergen, C.J. Information Retrieval, second edition. Butterworths, London, 1979.

19. Van Rijsbergen, C.J. A theoretical basis for the use of co-occurrence data in information retrieval. J. Documentation, Vol. 33, No. 2, (1977), pp. 106-119.

20. Zunde, P. and Dexter, M.E. Indexing consistency and quality. Amer. Documentation, (1969), pp. 259-267.