Division of Research
Graduate School of Business Administration
The University of Michigan

July 1984

ADAPTING DOCUMENT RETRIEVAL SUBJECT DESCRIPTIONS TO RELEVANT USER INQUIRIES

Working Paper No. 383

Michael D. Gordon

1. Introduction

The goal of document retrieval is to help get the "right documents" into the hands of the "right inquirers". However, when inquirers attempt to retrieve documents (books, journal articles, correspondences, etc.) by identifying their subject content, document retrieval can be far from successful. This lack of success arises from the complexity of our language--imprecision, multiple meanings of a single word, more than one way of saying the same thing. No matter how carefully a document's subject is described, there will be those who will retrieve it and find it without relevance, as well as those who do not retrieve it yet would have found it relevant if they had. This paper will show that documents may be described more effectivley by altering their subject descriptions as we learn more about what good subject descriptions of the documents should be. The results of a novel type of document retrieval simulation indicate that an adaptive algorithm can adjust document subject descriptions to produce descriptions superior to those with which the document was originally indexed. A discussion of these results forms the central part of this paper.

This paper begins with a brief discussion of the goals of document retrieval followed by a look at the current lack of effectiveness in meeting them. Indeterminacy in language is implicated as being at the root of the problem, and we then see that the inquirer bears an unusual burden of possibly having to

look for information in a quite unnatural fashion. The suggestion is made that document retrieval be regarded as a two-way conversation between retrival system and inquirer. The paper, building toward the premise that information about the way inquirers actually do search by subject for documents can be incorporated into a retrieval system in order to adjust document subject descriptions, first comments on the customary notion of (relevance) feedback as well as that of automatic indexing. Finally, an adaptive algorithm is discussed along with its application to the subject description of documents. A method for simulating document retrieval is desccribed, and the results of a simulation testing the adaptive algorithm in a document retrieval context are presented. The paper concludes with some observation concerning adaptive subject description.

This paper is concerned exclusively with improving the subject descriptions of documents, as opposed to "contextual" descriptions such as document author, date of publication, etc. All mention in this paper of "document descriptons" should be taken to mean subject descriptions of documents. I also use the the phrase "subject term" (or "term") and "subject descriptor" (or "descriptor") interchangeably. A "document description" should be thought of as a set of one or more subject terms.

2. Effectiveness of document retrievcal

People requiring information to solve problems will find, out of those "documents" they examine, some which satisfy their

need and some that do not. If a document retrieval system operated with prescience, inquirers would be furnished with just those documents that best meet their needs. Although such a hope is an impossibility, and retrieval systems which rank output in order of expected usefulness to an inquirer are trying to achieve a more reasonable goal (Robertson [14]), the idea that we are trying to furnish the "right documents" to the "right inquirers" remains a convenient and useful way to think of the situation.

Unfortunately, the current state of document retrieval falls far short of meeting this goal. In a recent investigation, Blair and Maron [3] investigated the effectiveness of full text retrieval. The mechanics of full text retrieval, which allows an inquirer to obtain just those documents which use in their text a set of subject terms he or she chooses, seems to imply success in meeting the goal of bringing together the right documents and the right inquirer. To the contrary, the Blair and Maron study reported recall of less than twenty percent for a group of attornerys, actually involved in litigation, who continually resubmitted full text queries (sets of terms) until they felt satisfied that at least seventy five percent of the documents pertinent to their legal needs had been furnished.[1] (A recall figure such as 20% meant that after an attorney had made all

---

[1] Recall can be defined to be the proportion of documents relevant to a query that are actually retrieved. Two other useful measure of retrieval effectiveness are: Precision--the proportion of retrieved documents that are relevant to a query; and Fallout--the proportion of non-relevant documents that are actually retrieved by a query.

efforts at retrieval, there were still four times as many known, unretrieved documents that he/she was to judge relevant to his/her information need than there were relevant-retrieved documents.) Even those documents "vitally relevant" to their leagl preparation, which the attorneys made every effort to retrieve, were delivered fewer than fifty percent of the time.

Why should recall have been so low? The group of attorneys was intimately acquainted with the situation they were researching, as well as the vocabulary used to discuss it. But, even in this relatively constrained domain, enough linguistic indeterminacy existed so that the inquiring attorneys were unable to identify just those words which the relevant documents used. The inquirers had run up against a most formidable document retrieval problem: many words mean close to the same thing, and one word may mean many things. As a result, retrieval effectiveness is less than what we might hope for.

3. Implication for subject description

Indeterminacy also presents a severe obstacle to one attempting to index (give an official description to) a document with a set of descriptive terms (or weights): if a given subject term can mean different things to different inquirers, and different subject terms can be used to mean the same thing, how is a document to be "<u>properly</u>" described? In short, describing a document with a set of subject terms mirrors the process of searching for the document. A document indexer must be concerned

with what terms an inquirer will choose to try to locate documents. An inquirer must be concerned with what terms have been chosen to describe documents.

The similarity breaks down in one very important way, however. Dissatisifed inquirers have the opportunity to alter their queries. Document descriptions, on the other hand, remain the same, even if they do not serve to furnish the right documents to the right inquirers.

If we begin to regard document retrieval as a kind of conversation, we see that the inquirer bears an inordinate share of the responsibility for making the conversation work. To receive useful documents, the inquirer may be forced to revise continually his or her query until he or she employs one that matches the descriptions of those documents that will prove to be truly relevant. Even if there is a consensus among pertinent inquirers (those who will find a certain document relevant) that some document is best described by a certain set of terms, this will not have any influence on the description attached to that document. If the documents's description does not match the query of such an inquirer, the inquirer is put in the somewhat unusual position of having to determine an alternate description of the information he or she seeks--even though he or she is issuing a query typical of most others who would find the document useful to their needs.

In fact, document retrieval be should be regarded as a type

of conversation in which both parties make an attempt to communicate. The premise of this resesarch is that document retrieval systems can be made more effective by finding out (with feedback) which descriptors inquirers who have need for a document actually employ in looking for it, and then altering document descriptions to agree better with these sets of terms.

Regarding information retrieval as a form of conversation builds upon Maron and Kuhns' [12] suggestion that the notions of "relevance" and "about" be operationally grounded. In arguing for the profitability of basing the definition of about and relevance on conditional probabilities, Maron and Kuhns indicate the need to consider the role of the inquirer in defining just what it is a document is "about" (should be called).[2] My work goes beyond Maron and Kuhns' by 1) suggesting an adaptive means to incorporate inquirers' usage of descriptors without reliance on human indexers who must estimate conditional proabilities; and 2) producing effective document descriptions probabilistically without any assumptions of term independence or dependence. As we shall also see, the adaptive approach is

---

[2] Maron and Kuhns offer these definitions:

Doc-i is about term-j =

P (inquirer requested information on a subject designated by term-j | he/she found doc-i relevant).

Doc-i is relevant to term-j =

P (inquirer found doc-i relevant | he/she requested information on a subject designated by term-j).

economical of space in its collection of feedback data (What queries have been used by satisfied inquirers who retrieved document-x? What queries were issued by inquirers with no need for document-x who were presented with it anyway?) and is sensitive to changes in the language inquirers use to retrieve documents.

In short, the same indeterminacy that plagues inquirers hinders effective description of documents. But, we can learn about the language inquirers are using to help answer the question: How can a document be best described?

4. Retrieval "conversations"

Let us develop our analogy between document retrieval and carrying on a conversation. Consider this exchange:

I: Say, can you tell me how to get to Chrysler?

D: Which Chrysler? The car dealership?

I: No, Chrysler Stadium, for the basketball game.

D: Oh, Chrysler Arena.

I: Yeah, Chrysler Arena.

D: It's on Main about a mile south of Seventh.

I: Main...that street there?

D: No, let me tell you again, it's easy. Go right at Seventh, that street with the light, to Pauline. Go left there and you'll hit Main. You'll be a block or two north.

I: One second: Seventh to Pauline to Main. And Seventh is that next street ahead?

D: No, that's Tenth, so Seventh is 3 more blocks...at the light.

I: ...to Pauline. Is that at a light, too?

D: Yes, and you have to go either left or right there, it's a T-shaped intersection. And you go left...

I: ...left...to Main?

D: Just park on the street, if you can. Follow everyone the people you'll see walking from there.

I: Got it. Thanks.

Why did this request for direction succeed? Mostly because when either the inquirer (I) or the describer (D) realized he or she was not being understood, he or she could think up another way of saying the same thing.

This simple lesson makes just as much sense in a document retrieval context. An inquirer looking for a particular document should not expect to find it if, instead, he or she asks for it by the wrong title. Similarly, common sense indicates that if an inquirer chooses a poorly selected set of subject terms, his or her likelihood of getting a useful document is sharply reduced. In such cases, an inquirer should be able to submit a better query, just as our direction seeker (I) clarified that he wanted to attend a basketball game, not look for a new car.

The concept of relevance feedback, as described in Salton [15] by Rocchio [13], is based on the premise that queries more effective in isolating relevant documents can be constructed, with the use of feedback, from less effective queries. However, relevance feedback fails to make proper adjustments for poorly described documents. (A poorly described

document is one whose description serves to provide it too often to those who are not interested it it, and/or not often enough to those who are. If a poorly described document were to receive a better description, both of these problems would become less pronounced.)

Two attempts at vector modification are reported in Salton [15]. First, Friedman et al [6] attempt to modify the complete document vector space with the discovery of each relevant document for every query. Besides the unwieldiness of this approach, after each query the document space is reset to its state before modification. Thus, any benefit coming from modification is not transmitted to future searchers. Second, Brauen [5] reports attempts to adjust permanently document vector weights in response to user queries. Though promising, the approach appears to have some defects: One, Brauen's approach does worse when it receives information about which documents are not relevant to which queries than when it does not receive this information. Two, "control" queries, which should not necessarily be more easily retrieved after vector modification, exhibit better recall-precision performance as a result of modification than do the "test" queries toward which modification was directed. Three, Brauen claims that each document which is redescribed is relevant to possibly just one query (this being the only one which will influence the redescription of the document). The effects of inquirer indeterminacy (differences in query formulation among inquirers looking for the same document) are, therefore, not fully

explored.

By not building in a mechansim to change document descriptions, an inquirer may have to rephrase continually his/her query in an effort to retrieve relevant documents--even if producing these modified queries appears to the inquirer to be a counter-intuitive activity. This situation can be likened to our inquirer (I) looking for Chrysler. The describer, D, provides compact, presumably adequate directions, "...on Main, about a mile south of Seventh," which are insufficient for I, who clearly is not as familiar with the territory as D. The clarifications that follow lead to a string of better, more easily followed directions to help the traveler find the basketball game. But better and more easily followed for whom? Plainly, for I, the inquirer, who needed the directions in the first place. D's original directions ("...on Main, about a mile South of Seventh") were adequate from his/her own point of view. Yet the conversation never would have worked had not more detail been supplied.

Several points begin to stand out with apparent application to document retrieval. First, an inquirer (looking for either directions to a basketball game or for a reference to a document) must submit a query that "makes sense." "Chrysler" could have referred to the automobile manufacturer or the basketball arena. "Chrysler," alone, did not make enough sense. The process of relevance feedback can be viewed as constructing queries which "make sense." Second, the suitability of a

description (directions for getting to the game; set of subject terms used to describe a document) must be judged in reference to the person who must use it. "...on Main about a mile south of Seventh" is not effective for the lost person, I, even though the directions are unflawed. Similarly, a document indexed as being "about terms A, B, C" may only be about those terms to the indexer who supplied them. For others, alternate descriptions may do a far better job. Third, an information retrieval system which does not itself change its descriptions is not operating optimally. If our same dircetion giver, D, were beseiged by queries asking for "Chrysler" (ambiguous) or "Chrysler Stadium" (no such place), he or she would quickly realize that he/she was actually being asked for the location of Chrysler Arena and could offer directions. A document retrieval system, to be effective, should do the same: not read inquirers' minds, but begin to pick up patterns, just as D could do. I contend that, particularly, a document retrieval system can pick up those linguistic "patterns" inquirers use to try to retrieve a given document. That is, it can adjust its document descriptions to help get the right documents into the hands of the right inquirers. That is precisely what the results of the simulation I conducted will show. Such adjustments in desscriptions, together with relevance feedback, help make retriveal a two-way conversation.

5. Automatic indexing

The point of this research is that document descriptions

are not adequate unless they do their job (get the right documents into the hands of the right inquirers), and that we can use feedback, as we will see, to get this job done better. Let me anticipate a possible objection to what I have just said: document redescription is never necessary because of automatic indexing.

Beginning with Luhn [11], efforts to analyze text statistically have counted word frequencies within a given text; word frequencies across a population of texts; the proportion of a population of documents that employs a given term; and term co-occurrence patterns. See, for example, Sparck-Jones [17], Sparck-Jones [18], Bookstein and Swanson [4], Harter [8], and Salton and Yang [16]. The data from such studies have become the organizational points from which theories of term selection or term or query weighting have been made. So why, if we know about how words are distributed within and between documents, singly and in pairs, do we need any kind of feedback to tell us how ocuments should be described?

The reply to this anticipated objection is that we don't know that inquirers will be using words in the same patterns that document authors put them down on paper. Authors write. Inquirers are inquiring, asking: the processes are different. They give rise to different patterns of language, just as written English and spoken English are not exactly the same language. Also, the interest inquirers have in a document may be different from the document's most salient statistical features.

Inquirers may, in general, find a certain document most relevant/useful because of some novel way it makes some well known point. This novelty may not be evident from a statistical analysis of the document's syntax. But with feedback from inquirers, aimed at findout what they feel a document's importance is, (how they are asking for it), we have a better chance of finding out.

The Blair and Maron [3] study should caution us about thinking that words in text simply lie there like "handles" which enable inquirers to pick out and pick up just those documents that suit their information needs. Instead, we see from that study that it is both the way the inquirers use words, together with the way that a document has been represented (by its full text, by human-supplied descriptors, by some automatic indexing procedure) that will spell the success of failure of retrieval. Therefore, attempts to "automatically index" a document cannot succeed completely since they ignore the way inquirers make their queries.

6. Genetic adaptation of document subject descriptions

A simulation experiment (see section 7-12) was carried out to see if improved access to a document would result by redescribing it more suitably relative to queries to which we have learned already the document is, or is not, relevant. Again, the idea is that the retrieval system, acting as one party in a conversation, can--and should--make adjustment to

make the conversation work. The conversation is about where to find references to information. The adjustments the system can make are in the descriptions it supplies to documents. The shape of these adjustments should be contoured by previous conversations with previous inquirers: Given other inquirers found this document relevant, what terms did they use in inquiring for it? And, similarly, what terms were used by inquirers who retrieved the document and found it not relevant? Using information from these previous "conversations," document redescription will try to increase the similarity between the document's description and its "relevant queries," and decrease the similarity between a document's descriptions and its "non-relevant" queries. As a result, retrieval effectiveness, as measured by both recall and precision (or fallout) should be improved.

An adaptive algorithm, invented by Holland [9], was used to incorporate information about previous queries as it controlled redescription of document subject descriptions. The algorithm borrows from genetics: In the background is an "environment" with respect to which various objects prove their fitness. Through competition, (in the sense of survival of the fittest), and by introducing variability among competing objects, new "generations" of objects replace old ones, with the newer being better adapted (with respect to some measurable characteristics of the environment) than the old. In adapting document descriptions, we hope to obtain descriptions which are "more fit" with respect to the queries we have already learned about.

Leaving aside for now document retrieval, we can consider the genetic algorithm in a bit more detail in an entirely different context. Suppose we have a positive, real-valued function, f, taking values on the closed real interval [0.0, 1.0] (which can be measured to 30 binary places of precision). The genetic algorithm can help us find the optimum value of f by operating, essentially, as follows:

1) Arbitrarily, pick a set of n (here, n=10) binary fractions between 0.0 and 1.0, each with 30 places of binary precision.

2) Repeat unitl "improvment" ceases:

a) For each binary fraction, x, in the current set, calculate f(x)

b) For each binary fraction in the current set, calculate its relative fitness (see Figure 1 and the following paragraph).

c) Reproduction: discard the current set of n binary fractions, replacing it by a new set comprised of copies of the just-discarded binary fractions in numbers equal to their relative fitness.

d) Cross-over: exchange (as in genetic cross-over) parts of these newly created binary fractions with each other (see Figure 2).

While a theoretical explanation of the algorithm is available, (see Holland [9] and Bethke [1]), I will simply make a few comments about the algorithm and show its operation with respect to our example. Suppose, as Figure 1 shows, we have arbitrarily selected ten binary fractions and determined the fitness value of each--the value of f(-) with each fraction as an argument. To calculate the "relative fitness" of a given binary fraction, we divide its fitness by the average fitness of the set of 10 fractions. We thus obtain ten relative fitness measures (again see Figure 1).

	x	fitness(x) (i.e., f(x))	relative fitness(x)
1)	.1010011	4	2.0
2)	.0100110	2	1.0
.	.	.	.
.	.	.	.
10)	.1001001	2	1.0
	Avg.	2.0	1.0

Figure 1--Ten competing binary fractions

The ten competing binary fractions (shown here having 6-place binary precision) have their fitness, f(-), measured. The relative fitness of a fraction is the ratio of the fraction's fitness to the average fitness of the set of fractions.

Armed with these relative fitnesses, the original set of ten fractions is replaced by a temporary set of ten new ones. This set is comprised of two copies of the first fraction; one copy of the second; and so on, these numbers obtained from the relative fitness calculations.[3]

Clearly, such "competition," if continued, would quickly produce ten copies of the fraction which gave f a higher value than any of the nine other fractions. To explore the huge space of 30-place binary fractions in order to find the optimal value of f(-), some variability is necessary. It is introduced through cross-over. Using the temporary set of ten fractions, we form five distinct pairs. For each pair, we randlomly pick a different cross-over point, which we use in breaking (paired) fractions in two and recombining them (see Figure 2). As a result of crossing over this entire set of temporary strings in this way, we obtain a new (non-temporary) set which we will again "test" (calculate f(-) of) in another iteration of the algorithm.

What we see, with suitable classes of of functions, is that, through adaptation, the set of fractions, on average, attains greater and greater value of f(-), the value approaching the maximum value possible. The "search" through the 2**30 fractions is conducted quite rapidly and, in fact, with "implicit parallelism." That is, any fraction actualy stands as

---

[3] Relative fitness scores which are not integer values have their fractional parts treated stochastically.

reproduced strings

before cross-over	after cross-over
110 | 001101	110 | 110000
010 | 110000	010 | 001101

Given the randomly generated cross-over position 3, the components of two randomly-paired (9-place) binary strings are exchanged. Vertical bar (|) indicates cross-over position. We can consider these strings to be binary fractions having values between 0.0 and 1.0.

Figure 2--Cross-over of binary strings

an exemplar of many, many other fractions. Consider the (8 place) binary fraction .00110001. It stands as an exemplar of the schema 0******* (the set of binary fractions with left-most position 0). It also stands for the schema 001****1 (binary fractions with left-most digits 001 and right-most digit 1).[4] The speed of the genetic algorithm comes from the most fit schemas tending to arise and proliferate in the set of (the ten) current fractions. All the while, combinations of schemas interact implicitly in a competition to find the fittest of them all.

In genetic adaptatioon of document descriptions, we will let a document be described simultaneously by multiple, competing descriptions. Each description will be a binary string, easily interpreted as a set of keywords just in the way that

{computers, hardware, network} abbreviated {C, H, N}

can be considered identical to

```
   A  B  C ... H ... N ...
 < 0  0  1 ... 1 ... 1 ... >
```

7. Document retrieval simulation: basic method

Isn't there competition inherent in the conversation

[4] An m-place binary string represents 2**m schemas.

between inquirers and retrieval system? After all, many inquiries can be made for a given document; and there are many ways for a document to be described.[5]

The linguistic competition suggested by Zipf [20], the disagreement among indexers reported by Zunde and Dexter [21], as well as the poor retrieval performance documented by Blair and Maron [3], each suggest suggest, in different ways, that describing a document cannot "just be done;" instead, it should be continually adjusted until it is "done right."

Three document retrieval simulations were performed: a "recall" simulation, a "fallout" simulation, and a "recall-falout" simulation. The goal of the first was to change the description of a document so that the modified description would match better the queries it should match. Such an improvement would produce improved recall. In the fallout simulation, the goal was the opposite: modify the description of a document so that the modified description would match worse the queries it should not match, thereby reducing fallout. And the goal of the recall-fallout simulation was to do both at once: modify a document's description so that the new description matches better the queries it should match and

---

[5] One is reminded of the competition Zipf [20] describes: speakers (tending to put forth minimal effort) want to talk in vague, general terms. Listeners (wanting to understand those speakers with minimal effort) want the speakers to speak with great care and precision. Conversations (and usage of language) tend to follow a middle ground. Speakers choose words which are better chosen than if no one were listening. Listeners must put forth moderate effort to try to dig out the meaning from speech which is somewhat vague and imprecise.

matches worse the queries it should not.

The name of each of these three simulations helps explain the way the simulation was conducted. Consider the "recall" simulation. Usually, recall is defined operationaly as the proportion of documents relevant to a query that are actually retrieved. A probabilistic definition may be given, too, relating a query and relevant documents:

recall (for query q) = P (doc retrieved | that document is relevant to q)

In the recall simulation I conducted, a variation of this probabilistic definition was used. In this variation, the document in question, not the the query, was fixed:

recall (of doc-x) = P(doc-x is retrieved by some query | doc-x is relevant to that query)

Operationally, this probabilistic definition says we may measure the recall (recallability might be a better word) of a document by determining the proprotion of times it is retrieved in response to queries to which it is relevant. Measuring recall in this way gives a realistic indication of the more customary meaning of the term (Gordon [7]). The "recall simulation" measured the effectiveness of document redescription in improving recall measured in this way.

Actually, in my simulation, recall was not measured by determining the proportion of times a given document is retrieved in response to a set of "relevant queries"; instead, I used an associational measure of recall (recallability) which was very similar:

recall (of doc-x) = average pair-wise association between the description of document-x and each member of the set of queries to which doc-x is relevant

Experimentation revealed that associational matching could be adapted with better results than "binary" matching (is a document relvant to a query or not?). I used a Jaccard's score to measure the association between a document description and a query that description should match. With a document represented by a set of subject terms, X, and a query represented by a set of subject terms, Y, the Jaccards's score calculates their degree of association as

$$|X \text{ intersect } Y| \quad / \quad |X \text{ union } Y| \ .$$

This association score, by no means the only one possible, has been used to measure representational consistency among document indexers (describers) (Zunde and Dexter [21]).

In the recall study, then, a description of a document (with respect to a fixed query, q, that should retrieve that document) is said to be improved by adaptation if its Jaccard score calculated with respect to that query rises as the result of adaptation. For example, if document-x should be retrieved by query q, where query q is represented by the set of terms {A, B, D, X}, then a change of the description of document-x from {A, C, X, Z} (Jaccard score of 2/6 with q) to {A, B, X, Y, Z} (Jaccard score of 3/6 with q) is considered recall improvement with respect to query q.

The way the recall simulation was conducted was based on obtaining a set of queries, Q, to <u>all</u> of which a given document,

document-x say, could be considered relevant. Then, the average pair-wise Jaccard score between each member of Q and the description of document-x could be calculated, from which we may obtain an "average" level of association between the description of a document and the queries to which it is relevant. By changing the description of document-x through adaptation, and then again calculating the average pair-wise Jaccard score between each member of Q and the new description of document-x, we have the basis for measuring the improvement in the recall of that document (see Figure 3).

Genetic adaptation of document descriptions actually depends on a document having several descriptions in force at one time, however. If a document is originally described by the n descriptions D-0 = {d-01, d-02, ... d-0n}, and, as a result of adaptation, comes to be described by a final (different) set of n descriptions D-f = {d-f1, d-f2, ... , d-fn}, we can simply compare the average Jaccard score obtained by matching pair-wise each description in D-0 to each query in Q with the average Jaccard score obtained by matching each description in D-f to each query in Q. Just as with single document descriptions, an improvement in this average is considered desirable. (See Figure 4 where the [overall] average Jaccard score, G, is shown for a set of n descriptions in generation-g. The comparison just described would contrast G resulting from matching D-0 with the set of relevant queries with G resulting from matching D-f with the same relevant queries.)

Description of document-x

query	before adaptation	after adaptation
q-1	J-before-1	J-after-1
q-2	J-before-2	J-after-2
.	.	.
.	.	.
q-n	J-before-n	J-after-n
Avg.	$1/n\sum_j$ J-before-j	$1/n\sum_j$ J-after-j

Figure 3--Level of matching between a description of a document before and after redescription with each of a set of "relevant queries"

J-before-i is the Jaccards's score match between the original description and the i-th relevant query; J-after-i the match between the new description and the i-th relevant query. Adaptation is successful if $1/n\sum_j$ J-before-j < $1/n\sum_j$ J-after-j.

	relev_x_q1	...	relev_x_qM	Avg_Matching Score
desc_x_g1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_gN	J(gN,q1)	...	J(gN,qM)	$1/M \sum J(gN,qi)$

N descriptions of document-x in generation-g

Overall average, 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_qi and document description desc_x_gj is indicated by J(gj,qi). Row averages give "average matching scores" for each document description. 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 averge matching score greater than G relative to the same relevant queries is an improvement on the generation-g set of descriptions.

Figure 4--Matching of descriptions and relevant queries

Summarizing, document simulation depends on knowing, in advance, queries to which a document is relevant. Using these, we may calculate the recall (recallability) of a document which is described in some way. Similarly, we may compare the effectiveness of two descriptions (or two sets of descriptions) by comparing the (average) recall each provides.

8. Relevant queries

The last section has made reference to "queries to which a document is known to be relevant." And, in fact, the simulations that were conducted depended on determining the relevance of a document to a query. In this section, I will explain how such a determination was made.

Let us suppose I read document-x and then choose, from a fixed set of subject terms, those I feel apply to that document. On the other hand, suppose I am hoping to find a document (which, it will turn out, will be document-x) which will satisfy my information need. In an attempt to describe my information need I again choose a set of subject terms which serve to describe the information I require. The two different sets of subject terms I've chosen should be quite similar.[6]

The similarity between describing a document and inquiring

---

[6] Blair [2] cautions that making an inquiry and describing a document are not really identical processes since an inquirer is describing information related to a problem he/she is interested in while an indexer does not have this "problem orientation."

for it means that to obtain a set of "relevant queries"--queries to all of which a given document is known to be relevant--to use in a document retrieval simulation, it is not necessary to collect data from a library or on-line retrieval system. Instead, we may have people read a document and describe it by choosing subject terms from some fixed vocabulary. The desriptions we obtain can be regarded as either a way to index the document or a way to inquire for it by subject.

9. Experimental design

Each of the recall, fallout, and recall-fallout simulations depended on artificial "relevant queries" (and/or artificial "non-relevant queries") as sections 7 and 8 suggest. In this section, I'll explain an experiment that was conducted to collect data which yielded these artificial queries and discuss the procedure for constructing them.

A group of 77 undergrads was recruited to read rather short, non-technical articles dealing with computers.[7] The group was homogenous in that each member had taken at least one, but not more than two, programming courses.

Each subject read four articles, randomly selected from a database of eighteen articles. For each of the eighteen articles, a different set of subject terms was assembled into a

---

[7] More precisely, forty five undergraduates participated, some more than once, for a total of 77 participations.

questionnaire.[8] Each reader of a given document would weight each term in the questionnaire. (A weight of 7 indicated the reader felt the term "definitely" applied to the subject content of the document; a weight of 5, that the term applied "somewhat"; a weight of 3 that the term applied "not too much"; and a weight of 1 that the term did "not [apply] at all." Intermediate values reflected intermediate degrees of applicability.) For each of the eighteen documents in the data base, then, a group of approximatley seventeen subjects filled out a questionnaire aimed at describing that one document.

By regarding a filled-in questionnaire as being a way to make an inquiry, (approximately) seventeen relevant queries were obtained for each document, each query being a <u>set</u> of subject terms. The procedure for doing so was to take, for each questionnaire, just those terms weighted at level 7 or 6 to be a member of the set constructed from that questionnaire. Translation from a set of terms, such as

{computer , hardware, network} or {C, H, N}

to the vector (string)

```
   A  B  C  ...  H  ...  N  ...
 < 0  0  1  ...  1  ...  1  ... >
```

for use by the genetic algorithm is straightforward.

Intuitively, such a relevant query makes sense, for such a query is composed just of those terms a reader/inquirer feels

---

[8] Actually, a questionnaire contained between 16 and 39 subject <u>phrases</u> (mean 27.4), depending on the document. A subject phrase might be something like "office efficiency," "computers of the future," etc.

definitley (or almost definitley) describes the document he or she has read (is looking for).

Remember that a fallout and recall-fallout simulation were conducted, too. In the fallout simulation, the goal of adaptation was to redescribe a document to match worse (lower average Jaccard score) those queries to which the document is known not to be relevant. (The recall-fallout simulation attempts to redescribe a document to accomplish both the goals of recall adaptation and fallout adaptation at once.)

Accordingly, a set of "non-relevant" queries was needed for each document in the simulation. The procedure for obtaining such a set of non-relevant queries was to take, for each of the approximately seventeen filled-in questionnaires corresponding to that document, just that set of terms the describer weighted at a level of 5 or 4 (described the document "somewhat" or a little bit less). Taken together, these seventeen "non-relevant queries" each represent a description <u>not</u> of the document from which the questionnaire data was obtained, but of a document somewhat like it. Regarded as a query, each such non-relevant query should ideally not retrieve (match the description of) the document which inspired its description. Yet, these descriptions were purposely chosen to resemble somewhat relevant descriptions (that is, relevant queries) in order to simulate the occurrence of documents being erroneously retrieved in response to a non-relevant query.

This section has explained the technique for obtaining the raw data from which artificial "relevant" and "non-relevant" queries were constructed. Also described were the rules for converting these data into artificial queries.

10. Recall simulation: conduct and results

We have already discussed the components of the recall simulation experiment except for the selection of the original description of any document. In this section, the conduct and results of the recall simulation will be presented.

To provide the most digestible picture of this simulation, we look again at Figure 4. We see in that figure that each of the relevant-queries, relev-x-q1, ... , relev-x-qM is represented. This set of relevant queries is selected as described in the previous section. Each set is unique for a particular document. That is, the set of relevant for document-x is completely different than the set for document-y. If Figure 4 were redrawn for each generation in the recall simulation of document-x, (all simulations ran forty generations), the same set of relevant queries would be depicted in each figure.

Recall that the genetic algorithm requires that the document being described be simultaneoulsy represented by several "competing" descriptions. The n descriptions desc-x-g1, ... , desc-x-gN in Figure 4 are these n, competing descriptions in the g-th generation of the simulation. Initially

(generation 1) this set was taken to be identical to the set of relevant queries for the same article. Of course, as a result of adaptation, the set of descriptions of document-x will change from generation to generation. What is hoped is that after forty generations (when adaptation has been completed) that the overall average matching score (G in Figure 4) will be "better" (higher) than it was in generation 1.

The recall simulation was conducted separately for each of eighteen different documents. That is, for each document, a different set of relevant queries was constructed and that document was assigned its own initial set of document descriptions (generation 1 description set). The fitness (to be used by the genetic algorithm) of any generation-1 description, desc-x-1i, was calculated to be its average pair-wise Jaccard's score calculated with respect to each of the relevant queries, (its "average_matching score" in Figure 4). From these fitness scores, generation 2 descriptions of document-x were generated, replacing the original set, and these in turn were evaluated with respect to the relevant query set to produce the generation-3 description set, and so on. Simulation experiments revealed that for each of the eighteen documents separately studied, improvement seemed to have levelled off by the fortieth generatoin of such adaptation. As a result, measurement of improvement compared the original description set attached to a document with the resulting set after 40 generations of adaptation.

Since a Jaccard's score of 1.00 (100% association) means the two sets being compared are identical, whereas a Jaccards's score of 0.00 (0% association) means they are disjoint, what I hoped to see was that adaptation would boost the overall average Jaccards's score between (fixed) relevant queries and (changing) document descriptions from generation 1 to generation 40.[9] In fact, in each of the eighteen recall simulation experiments, such an elevation did occur. All eighteen documents considered together, the overall average Jaccard's match between the fixed relevant queries pertinent to a a document and the changing set of descriptions associated with that document rose from a Jaccard's score of 39.53 (generation 1) to a Jaccard's score of 48.88 (generation 40) (a relative increase of over 24%; see Table 1 and Figure 5).

But, it was important to determine whether, somehow, document redescription had had the disasterous "side effect" of causing just as great an elevation in Jaccard matching between queries that could be deemed not relevant to the document being redescribed. Said again, redescribing a document should only make it more likely the document will be retrieved by those inquirers who will judge the document relevant--not more likely it will be retrieved by all inquirers.

The non-relevant queries (see secton 9) were used as a control in determining whether elevation in Jaccard score

---

[9] For readability, Jaccard scores are multiplied by 100 to resemble percentages.

	Gen1	Gen6	Gen11	Gen16	Gen21	Gen26	Gen31	Gen36	Gen40	Abs Chng	% Rel Chng
Doc 1	36.03	39.67	41.05	41.65	43.80	44.57	45.08	45.54	46.08	10.05	27.89
Doc 12	44.45	48.16	50.50	50.99	50.90	51.12	51.84	52.01	52.06	7.61	17.12
Doc 17	42.19	45.48	47.17	49.36	51.32	52.15	52.34	53.01	53.01	10.82	25.65
Doc 18	39.36	48.53	49.89	51.45	51.73	52.30	52.34	52.45	52.46	13.10	33.28
Doc 19	41.12	44.96	46.26	46.79	47.28	48.03	49.18	49.46	49.91	8.79	21.38
Doc 21	43.01	47.66	48.31	50.04	50.14	50.60	51.55	51.75	52.12	9.11	21.18
Doc 22	33.45	35.07	38.16	38.76	39.64	40.72	41.90	43.00	43.83	10.38	31.03
Doc 23	31.81	37.75	38.88	38.88	39.00	39.38	39.70	39.83	39.83	8.02	25.21
Doc 25	54.21	60.91	63.08	64.70	64.78	64.95	64.96	64.93	65.02	10.81	19.94
Doc 27	37.92	42.68	43.84	44.38	44.70	45.32	45.53	45.59	45.98	8.06	21.26
Doc 28	28.06	30.79	32.86	33.20	34.85	35.51	36.19	36.38	36.57	8.51	30.31
Doc 30	48.15	50.56	53.37	54.77	56.14	56.35	57.30	57.43	57.94	9.79	20.33
Doc 32	47.36	51.67	55.16	56.49	57.74	58.07	58.33	58.47	58.48	11.12	23.48
Doc 33	39.95	42.63	44.83	45.05	45.94	46.08	47.30	47.38	47.52	7.57	18.95
Doc 34	36.80	41.72	41.55	41.79	42.29	43.00	44.73	45.58	45.92	9.12	24.78
Doc 35	39.83	45.94	48.15	48.42	49.58	49.80	50.06	50.28	50.52	10.69	26.84
Doc 36	31.23	34.52	36.65	38.40	39.33	39.73	40.15	40.20	40.40	9.17	29.36
Doc 7	36.66	36.68	39.27	41.22	41.33	41.60	41.90	42.34	42.28	5.62	15.33
Avg.	39.53	43.63	45.50	46.46	47.25	47.74	48.35	48.65	48.88	9.35	24.07
S.D.										1.62	4.77

This table lists the overall average matching scores between document descriptions and (fixed) relevant queries, displayed at five generation intervals.

Table 1--Recall performance

matching was constrained to relevant queries (as desired) or whether the "side effect," above, had taken place. Since the non-relevant queries were composed of subject terms that "just about described a document--but not quite," they seemed suitable for this control study. That is, a (fictitious) inquirer issuing non-relev-xi (the i-th query to which dcoument-x was posited not to be relevant) likely might receive document-x when he/she made that inquiry to the retrieval system. This occurrence would come about since the subject terms comprising non-relev-xi are those that are close (but not quite close enough) to being the strongest terms that inquirer actually did select in describing doc-x. Accordingly, it seemed that if adaptation of a document's subject would inadvertently improve the average Jaccard's score for any queries, it would be such "non-relevant" queries.

A way to measure inadvertent improvement was to calculate the average Jaccard's score obtained from comparing the non-relevant queries for a given document and the generation-1 description set for that document, and then repeating this calculation with the same set of non-relevant queries and the generation 40 (final) description set. In this way, we can see the level to which attempts to increase the average Jaccard's match between a document's description(s) and its relevant queries also increased the average Jaccard's score match for the non-relevant queries. Table 2 shows the results of these measurements. This table also shows how well the <u>intended</u> improvement (increasing the average Jaccard's score between relevant queries by adapting descriptions) compared to

Document	Change in overall average matching score from gen1 to gen40	
	relevant queries	Non-relevant queries
Doc 1	10.05	4.14
Doc 12	7.61	1.81
Doc 17	10.83	-5.10
Doc 18	13.11	8.51
Doc 19	8.79	4.75
Doc 21	9.11	0.96
Doc 22	10.38	0.08
Doc 23	8.01	-8.18
Doc 25	10.81	-3.29
Doc 27	8.06	6.87
Doc 28	8.51	11.43
Doc 30	9.79	3.05
Doc 32	11.12	2.94
Doc 33	7.56	5.41
Doc 34	9.11	3.48
Doc 35	10.69	1.91
Doc 36	9.17	1.69
Doc 7	5.62	-5.82
Avg.	9.35	1.92
S.D.	1.62	4.89

Data expressed in units of Jaccard's points.

Table 2--Increase in overall average matching for non-relevant queries versus relevant queries

inadvertent improvement (the increase in Jaccard's score between the set of non-relevant queries brought about by recall adaptation). As the table reveals, for fourteen documents out of eighteen, there was some inadvertent improvement (increase in Jaccard's score for non-relevant queries). Far more importantly, however, these figures show that seventeen times out of eighteen inteneded improvement dominated inadvertent. What this means is that it becomes easier and easier to distinguish a query that is relevant to a document from one that is not. As VanRijsbergen [19] points out in discussing term discrimination, what is desired of a subject term is an ability to distinguish better documents that will be relevant to a query from those that will not. In describing documents with complete descriptions, the objective should be the same. Genetic adaptation does precisely that by increasing the average difference in Jaccard score between queries relevant to a document and queries not relevant to it. For example, for document 1 redescription provided an additional 5.91 Jaccard point separation by which relevant and non-relevant queries can be distinguished.

To summarize, genetic adaptation does redescribe documents so that the new descriptions it produces match relevant queries better than the original descriptions. What's more, this improvement does not suffer the side effect of having arbitrary queries also match these new descriptions better to a commensurate degree.

11. Fallout simulation

A fallout simulation was conducted to see if retrieval exclusiveness (correctly failing to retrieve a document when it should not be retrieved) could be improved, just as the recall simulation tried to improve recall inclusiveness (correctly retrieving in response to relevant queries).

Like the recall simulation experiments, an adaptive experiment was conducted independently for each of eighteen different documents. For any document, what I hoped to do through adaptation was to decrease the overall average degree of association between the set of descriptions associated with the document and those queries to which that document could be judged non-relevant. The non-relevant queries (used in the recall simulation as a control) were used in in fallout simulation as the "environment" against which to adapt (see Figure 6). In that figure, note that there are two differences in comparison to the recall experiment. First, as just mentioned, the non-relevant queries remain constant from generation to generation, each generation providing a measure of "fitness" for the current generation of document descriptions. Second, what this fitness should be measuring is the degree to which a document's descriptions are unlike the non-relevant queries for that document. That is, the more fit the description, the lower its average (Jaccard's) association with the set of a non-relevant queries. To make the genetic algorithm work, since it reproduces descriptions with higher fitness

	<--Non-relevant queries-------> non-rel_x_q1 ... non-rel_x_qM			Avg Matching Score
desc_x_g1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_gN	J(gN,q1)	...	J(gN,qM)	$1/M \sum_i J(gN,qi)$

N descriptions
of document-x
in generation-g

Grand average, G, =

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

The matching procedure is the same as for relevant queries and document descriptions. J(gj,qi) denotes the Jaccard's match between non-relevant query non-rel_x_qi and document description desc_x_gj. The (fallout) fitness of desc-x-gi = G + (G - Avg Matching Score(desc-x-gi))

Figure 6--Matching of descriptions and non-relevant queries

scores, something had to be done to make lower association translate into higher fitness. What I did was simple: calculate the overall average association between all descriptions (in generation-g) and all non-relevant queries, calling this average G. Then, calculating the average Jaccard's score between desc-x-gi (the i-th description in generation g) and all non-relevant queries to be X, take the fitness of desc-x-gi to be G + (G -X). In words, give each document description in generation-g a fitness value which is exactly the same amount higher than G as its average Jaccards's score (here, what we called X) is lower than G. Or: invert all calculated average Jaccard's scores around G.

In simulating fallout improvement, a document was originally described exactly as it was described in the recall experiment: by using its "relevant queries" as its generation-1 set of descriptions. As a result of adaptation, the set of documents attached to a description changed from the initial generation, generation 1, to the final generation, generation 40. What was expected was that the average level of Jaccard's association (G) would drop from generation-1 to generation-40, the final generation in the simulation. This, in fact, occurred quite dramatically. But, only by seeing if such a desired drop in association could be achieved along with the kind of increase in association obtained in the recall experiment could the claim of complete adaptive "success" be made. For this reason, discussion of fallout improvment is deferred to the next section.

13. Recall-fallout simulation

In the recall simulation it was demonstrated that redescription could be performed that would make document descriptions match better the queries those descriptions should match. The fallout simulation showed that adaptation could also be performed to reduce the level of association between document descriptions and queries to which the associated document is not relevant. Both of these ends are desirable, but somewhat contradictory in their aims. Recall improvement generally worsens fallout. Fallout improvment generally worsens recall. So, the finding of the recall-fallout simulation experiments that both recall and fallout improvement can be achieved simultaneously through genetic adaptation is impressive. In this section, I describe the recall-fallout simulation experiments and their results.

The format of the recall-fallout simulation is actually nothing more than the recall simulation and fallout simulation conducted simultaneouly (see Figure 7). Notice in that figure that each one of the competing descriptions of document-x in force during generation-g is compared both with the (fixed) set of relevant queries to see how well it matches them and also with the (fixed) set of non-relevant queries to see how dissimilar it is to those. As a result, the fitness of one of these descriptions, say desc-x-gi, the i-th description of document-x in generation g, is taken to be a weighted sum of the "recall fitness" of desc-x-gi and the "fallout fitness" of

	relev_x_q1	...	relev_x_qM	Avg_Matching Score
desc_x_g1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_gN	J(gN,q1)	...	J(gN,qM)	$1/M \sum_i J(gN,qi)$

N descriptions of document-x in generation-g

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

****

	non-rel_x_q1	...	non-rel_x_qM	Avg Matching Score
desc_x_g1	J(g1,q1)	...	J(g1,qM)	$1/M \sum_i J(g1,qi)$
.	.		.	.
.	.		.	.
.	.		.	.
desc_x_gN	J(gN,q1)	...	J(gN,qM)	$1/M \sum_i J(gN,qi)$

N descriptions of document-x in generation-g

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

Each document description 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'.

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

Figure 7--Matching of descriptions with relevant and non-relevant queries

desc-x-gi. Said again: if a description's recall fitnes (fitness with respect to the relevant query set) is r, and its fallout fitness (fitness with respect the non-relevant query set) is f, take the description's fitness in this simulation to be r + wt*f. The weight, wt, a non-negative real number, was used in order to balance the relative strength of recall and fallout adaptation. Where recall adaptation produced average recall improvement to rise by about 20% (as measured by relative improvement in Jaccard's score), fallout adaptation tended to reduce association by 80% or more. As a consequence, some experimentation showed a weight of about 0.5 would bring recall and fallout in balance, mitigating the side effect of fallout adaptation which would have made simultaneous recall improvement nearly impossible.

Like the two simulations before it, the recall fallout simulation really consisted of eighteen independent experiments, each conducted on a different document. The method of selecting the initial set of document descriptions for any document was the same used in both the recall and the fallout simulation. The way in which relevant and non-relevant sets of queries were obtained also remained the same. Each of these eighteen experiments, like those in the recall and fallout simulation, ran for forty generations.

The results of these experiments is shown in Table 3. Importantly, notice that fifteen times out of eighteen, recall and fallout improvement (rise for recall, decline for fallout)

	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 12	44.45	50.59	13.81	17.83	7.69	-56.87
Doc 17	42.19	52.53	24.51	17.12	11.05	-35.36
Doc 18	39.36	50.58	28.51	21.08	25.87	+22.72
Doc 19	41.12	47.33	15.10	18.83	17.58	- 6.64
Doc 21	43.01	52.45	21.95	18.00	16.04	-10.89
Doc 22	33.45	40.09	19.85	18.11	13.87	-23.41
Doc 23	31.81	39.98	25.68	12.92	4.28	-66.87
Doc 25	54.21	64.43	18.85	13.72	8.33	-39.29
Doc 27	37.92	46.65	23.02	17.65	13.25	-24.93
Doc 28	28.06	30.23	7.73	19.34	14.52	-24.92
Doc 30	48.15	57.72	19.88	16.88	18.45	+9.30
Doc 32	47.36	57.09	20.54	16.69	16.81	+0.72
Doc 33	39.95	44.29	10.86	20.29	13.75	-32.23
Doc 34	36.80	43.95	19.43	18.25	16.16	-11.45
Doc 35	39.83	47.64	19.61	17.88	13.03	-27.13
Doc 36	31.23	37.99	21.65	14.75	8.53	-42.17
Doc 7	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 a Jaccard's score of 36.03 (before adaptation) to a Jaccard's score of 42.86 (18.96% improvement). The same document redescription resulted in the average match between doc 1's non-relevant queries and document descriptions dropping from a Jaccard's score of 20.07 to a Jaccard's score of 14.47 (a 27.90% improvement).

Table 3--Recall-fallout improvement

was achieved. Arguing statistically, such dual improvement is significant at a level < .005[10] See also Figure 8.

Translating to real world terms, the success in improving both recall and fallout through adaptation means that a document can be redescribed more similarly to those queries in response to which it should be retrieved, and, at the same time, less similarly to those queries in response to which it should not be retrieved. That is, more often will interested inquirers retrieve a given document and less often will it be retrieved by those not interested in it.

14. Summary and conclusions

A document is well described if its description makes it likely that the document will be furnished to the right inquirers. This paper has argued that communication about the way inquirers ask for documents can help a retrieval system better describe those documents. Operationally, using an associational measure as a maching functon, we say one description of a document is better than another in promoting recall if it matches better the queries to which the document it is describing is relevant. Similarly, a better description, in the sense of fallout, bears less (statistical) association to

---

[10] We use a sign test with p=1/2 and q=1/2. That is, we hypothesize that, by chance, recall and fallout will both improve 50% of the time documents are redescribed (p=1/2); and 50% of the time this dual improvement will not occur (q=1/2). In fact, there were fifteen "successes" (occurrences of dual improvment) in eighteen trials.

the queries to which the document it describes is not relevant than does a worse description.

A recall simulation, a fallout simulation, and a recall-fallout simulation were conducted to support the argument that communication from inquirers can improve document descriptions. Each of these simulation was actually a collection of eighteen individually run simulation experiments. A single recall simulation experiment attempted to redescribe a document to make it more likely to be retrieved in response to relevant queries. A single fallout simulation experiment adjusted the way a document was described to try to make it less likely likley to be retrieved in response to a non-relevant query. In a recall-fallout simulation experiment, both of these goals were attempted at at once.

A document's description was actually taken to be a set of subject descriptions, each one a set of subject terms. Such multiple descriptions were used to allow the operation of an adaptive algorithm that could improve the way a document was described. When it came time to match a query to a document, a consensus, of sorts, was established: what was the average Jaccard's score match between the given query and each member of of the set of descriptions currently used to describe the document? In real-life retrieval, such a consensual decision would determine which documents would be furnished to a query.

The simulation results indicated conclusively that we can

adapt document descriptions via feedback to make them do their job better. In the recall-fallout simulation, fifteen times out of eighteen a document's description was altered in such a way so that both the document was more likely to be furnished to inquirers would would find it relevant and also less likely to be furnished to those who would not. In the remaining three cases, there was a mixed effect: the document's redescription caused it to be more likely to be furnished in response to relevant and non-relevant queries alike. But, even here, the improvement in recall exceeded the worsening in fallout (on average, for these three documents, 10.17 rise in Jaccard score for relevat queries vs. 2.16 rise in Jaccard score for non-relevant queries). That is, improved association to relevant queries was stronger than it was with non-relevant queries.

Again, for emphasis: the critical observation to make about this study is that document redescription can be achieved to produce better document descriptions. In fact, as I have argued, ignoring the way inquirers ask for documents short circuits the reciprocal communication between system and inquirer on which retrieval ought to be based.

The adaptive mechanism I have described produces document descriptions that, based on past experience, should help get the "right documents" into the hands of the "right inquirers." Some other research indexing methods, less empirically based, build up document descriptions from assumptions of statistical independence of subject terms which have this flavor: Inquirers

will pick term X in describing a document with probability 0.8 and will pick term Y in describing it with probability 0.7; therefore, an inquirer will describe the document with both X and Y with probability 0.56 (0.8 * 0.7). The indexing method I have proposed does not hinge on this (potentialy dangerous) assumption. Instead, entire document descriptions evolve, with whatever term dependencies that occur through usage being accurately reflected in the document's description.

Certain details of the study I conducted are more or less arbitrary, and discussion about them should point the way to further study. Certainly, other matching functions besides a Jaccard's score could be used to reflect association, a Cosine measure (Salton [15]) to name just one. And the adaptive machinery I have descibed in this paper can be applied just as easily to a retrieval model base on other types of matching, such as Salton's cosine. In fact, exploring the space of retrieval models to see which ones are best suited for adaptation, and to determine which retrieval models are most appropriate for which type of inquirer (see Lancaster [10]) for a discussion of differences among inquirers) is an important area of study (Gordon [7]). Similarly, translating improvement in Jaccard's score associational matching to improvement in recall/precision using some other retrieval model has not been investigated and warrants study.

The genetic algorithm is also only a candidate adaptvie algorithm. But studies such as those reported by Zunde and

Dexter [21] show there is enormous variation in the way people describe (or, ask for) information. Similarly, the questionnaire data I obtained from subjects who described the same document showed that there is rarely complete agreement or disagreement about the applicability of a subject term to a document. In my research, the questionnaire data about subject term applicability obtained from document readers were "collapsed" so that, if in choosing weights from 1 to 7, a reader said either that a term applied "definitely" (7) or a bit less (6) to the document he/she read, he/she was in the "yes" category for that term and that document. Otherwise, he/she was in the "no" category. All told, even after collapsing, the (approximatley) seventeen readers who read any doccument agreed completely on a subject term's applicabilitty (by all being in the "yes" category for a subect term, or by all being in the "no" category) for fewer than one term in ten. A powerful algorithm, like the genetic algorithm, is quite sensical, then, in its ability to "explore" a document description space which explodes combinatorially with inquirer indeterminacy. In fact, a comparison of the genetic algorithm against more deterministic indexing algorithms employing the same feedback data used by the genetic algorithm reveals the superiority of the former (Gordon [7]). And, through its multiple descriptions of documents, the genetic algorithm economically summarizes past user satisfaction. That is, the current set of descriptions attached to a document are, in a sense, a "state" at which the system has arrived as a result of interactions of past inquirers and previous document descriptions. As a state, a document's

descriptions is really a "shorthand," or "code," of past inquiry, meaning that other, likely less compact, histories of past satisfaction need not be collected.

The simulation technique used in this research can be improved by choosing more selectively relevant queries, non-relevant queries, as well as documents for study. With better selection, we may be able to represent more faithfully some collection of documents and simulate an inquiring community more like those who will actually be attempting to retrieve certain documents.

Concern that multiple descriptions are too uneconomical of storage or processing time are of secondary importance to the need that exists to improve document retrieval by subject. Also, accelerating processing speeds, as well as "sparse-array" like storage techniques mitigate both of these concerns. Still, improvement in physical database design should be studied if documents are to be multiply described.

Another reason for employing adaptation of descriptions is actually to improve efficiency. If I make a query and retrieve several different documents that I find relevant, then the description of each of those documents will tend to become more like the query I made. As a result, a "clustering based on use" should emerge, with different documents all relevant to closely related information needs being described similarly to each other. Like traditional clustering schemes, searching (matching)

can then be constrained to a fraction of the database. But unlike clustering which ignores the inquiries used to retrieve a document, the type of "clustering based on use" that I'm suggesting will develop ensures that documents which are often rightly retrieved together will be clustered together. This is not necessarily the case with clustering which lumps together documents according to the words in their title or text, or those which share terms selected by human indexers, for then there is no guarantee that these documents will ever need to be used (retrieved) together.

In summary, emphasizing that inquirers should have a primary role in describing documents leads us toward building systems to allow communication from inquirers to information retrieval systems in an atempt to improve document descriptions. That this should, and can, be done is the major idea in this paper.

REFERENCES

1) Bethke, Albert Donnally,
"Genetic algorithms as function optimizers,"
Ph.D. thesis, University of Michigan,
Ann Arbor, Michigan, 1981

2) Blair, David,
"Pragmatic aspects of inquiry,"
Ph.D. thesis, University of California, Berkeley,
Berkeley, California, 1981

3) Blair, David, and Maron, M.E.,
"An evaluation of retrieval effectiveness for
a full-text document retrieval system,"
Unpublished manuscript, 1982

4) Bookstein, A., and Swanson, D. R.,
"Probabilistic models for automatic indexing,"
Journal of the American Society for Information Science,
25 , 1974

5) Braun, T. L.,
"Document vector modification,"
chapter 24 in Salton

6) Friedman, S. R., Maceyak, J. A., and Weiss, S. F.,
"A relevance feedback system based on document transformations,"
chapter 23 in Salton

7) Gordon, Michael
"Adaptive subject description in document retrieval,"
Ph.D. thesis, University of Michigan,
Ann Arbor, Michigan, 1984

8) 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,"
Journal of the American Society for Information Science, 26,
1975

9) Holland, John
Adaptation in Natural and Artificial Systems
University of Michigan Press,
Ann Arbor, Michigan, 1975

10) Lancaster, F. W.,
Evaluations of the Medlars demand search service
U. S. Department of Health, Education, and Welfare
from Report GER 12760
"Measures of effectiveness and criteria for evaluation of
a document processing system,"
Rome Air Development Center, Jan., 1968

11) Luhn, H.P.,
"The automatic creation of literature abstracts,"
IBM Journal of Research and Development,
as reported in VanRijsbergen

12) Maron, M. E., and Kuhns, J. L.,
"On relevance, probabilistic indexing, and information retrieval,"
JACM, 3, 1960

13) Rocchio, J. J., Jr.,
"Relevance feedback in information retrieval,"
chapter 14 in Salton

14) Robertson, Stephen E.,
"The probability ranking principle in information retrieval,"
Journal of Documentation, 33, 1977

15) Salton, G.,
The SMART Retrieval System--Experiments in Automatic Document Processing,
Prentice-Hall
Englewood Cliffs, N. J., 1971

16) Salton, Gerald, and Yang, C. S.,
"On the specification of term values in automatic indexing,"
Journal of Documentation 29, 1973

17) Sparck Jones, Karen,
"A statistical interpretation of term speccificity and
its application in retrieval,"
Journal of Documentation, 28, 1972

18) Sparck Jones, Karen,
"Index term weighting,"
Information Storage and Retrieval,"
vol 9, 619-633, 1973

19) VanRijsbergen, C. J.,
Information Retrieval,
Butterworth and Co., Ltd., London, 1979

20) Zipf, George Kingsley,
Human Behavior and the Principle of the Least Effort,
Addison-Wesley, Cambridge, Mass., 1949

21) Zunde, Pranas, and Dexter, Margaret, "Indexing consistency and quality," American Documentation, July, 1969