PERFORMANCE MEASUREMENT IN A FUZZY RETRIEVAL ENVIRONMENT

Duncan A. Buell and Donald H. Kraft
Department of Computer Science
Louisiana State University
Baton Rouge, Louisiana 70803

ABSTRACT

We shall consider retrieval performance measures for generalized (non-Boolean) queries and indexing functions. The meanings of recall and precision in such a generalized system will be discussed. Finally, we shall explore the meaning and difficulty of using such measures to compare Boolean and non-Boolean retrieval systems.

1. INTRODUCTION

Measuring the extent to which a computerized document retrieval system fulfills the goals set for the system is a complex problem that involves everything from initial goal specification to the actual underlying computer software. An average user will view the system as a "black box." The user makes requests; the system responds. Numerous factors will thus affect the evaluation of the system by such a user. These include such varied aspects as physical ease of use, the user's ability to understand how to formulate requests, the coverage of the desired topic by the collection (and the coverage of new or old material at the user's appropriate level), and even the user's own knowledge of the topic in which he is interested. We emphasize that we are investigating only a narrow aspect of retrieval system evaluation. We consider not the "human engineering" required to provide the average user with the information he desires, but the establishment of quantitative standards by which to measure the ability of the mathematics and logic of the retrieval decision mechanism to select for retrieval in response to a request the same set of documents which would have been selected by a human expert unaided by the automated system. [9, 16, 18] Among these standards are recall and precision and associated measures and various measures of "value" returned in comparison to the search length. These measurements are well-defined for systems with Boolean indexing and standard Boolean query-to-document matching functions.

2. BOOLEAN SYSTEMS AND THEIR PERFORMANCE MEASUREMENT

We begin by examining Boolean systems and measures of their performance. A set D of documents exists for retrieval. The documents are indexed by a set of terms I using a membership function (MF) f:D x I --> {0,1}, with

$$f(d,t) = 1 \text{ if document } d \text{ is "about" } t,$$

$$f(d,t) = 0 \text{ if document } d \text{ is "not about" } t.$$

A user of the system specifies his request as a query to the system, which must then compute for each document a retrieval status value (RSV) e that measures whether or not the document is "about" the query. In a strictly Boolean system, the values of e are either 0 or 1, and the RSV can be thought of as a generalized MF. That is, those documents with an RSV of 1 are "about" the query, and those with RSV 0 are "not about" the query. Those documents deemed by the system to be "about" the query are then selected for retrieval.

The human expert may be the user, but this is not necessary. The model of system performance should be capable of isolating the system components from each other so that each component's contribution to overall system performance can be determined. Since we are interested in determining how well non-Boolean mechanisms measure relevance, only this factor should be considered. Other factors that could affect a naive user's view of the system include the language of the document, date of publication, readability, physical accessibility of the document, the level of knowledge required to comprehend the document, and whether or not the user has already seen the document. Some but certainly not all of these factors can be incorporated into the query. Thus, an expert is needed to determine the relevance of a given document to a given query without regard to these external factors.

Recall and precision, the most commonly used measurements of performance, are defined as follows. For any query, the retrieval system separates D into complementary sets RT and D-RT of documents retrieved and not retrieved, respectively. A human expert, however, might separate D into sets RL and D-RL of documents which are or are not actually relevant to the query. Recall, precision, and related measures are measures of the extent to which RT matches RL. Specifically, recall is defined as the quotient

$$\frac{\text{number of documents both relevant and retrieved}}{\text{number of relevant documents.}} \quad (1)$$

Precision, on the other hand, is the quotient

$$\frac{\text{number of documents both relevant and retrieved}}{\text{number of retrieved documents.}} \quad (2)$$

In set notation these are, using $|A|$ to denote the cardinality of a set A,

$$\text{recall} = \frac{|RT \cap RL|}{|RL|} \quad (3)$$

$$\text{precision} = \frac{|RT \cap RL|}{|RT|}.$$

To set the stage for later discussion, we convert the cardinalities into MF computations. Remembering that in a Boolean system the MF of the intersection of two sets is the (elementwise) minimum of the two MF's, we can write, using fT and fL to denote the MF's of RT and RL, respectively:

$$\text{recall} = \frac{\sum \text{Min}\,(\,fL(d,q)\,,\,fT(d,q)\,)}{\sum fL(d,q)} \quad (4)$$

$$\text{precision} = \frac{\sum \text{Min}\,(\,fL(d,q)\,,\,fT(d,q)\,)}{\sum fT(d,q)}$$

(The summations extend over all elements d of D.)

In addition to recall and precision, there are other measures, including normalized recall and precision, fallout, and generality, all of which are functions of the four cardinalities $|RT \cap RL|$, $|D\text{-}RT \cap RL|$, $|RT \cap D\text{-}RL|$, and $|D\text{-}RT \cap D\text{-}RL|$. [9, 16, 18] The crucial requirement for the definition of these measures is that relevance and retrieval are both Boolean functions.

3. GENERALIZED SYSTEMS

The previous description has been of a totally Boolean system, in which both the MF f and the RSV e are Boolean. It is possible to generalize either or both of these to be non-Boolean. A difference exists between the basic nature of the various methods of generalization. [2, 3] On the one hand, mechanisms for computing RSV's have been proposed [1, 4, 5, 12, 13, 14, 19] which attempt to preserve the analogy as closely as possible with Boolean set theory. Waller and Kraft have used the term "separability" to describe a mechanism which allows for RSV computation as a parallel process to forming subsets of D which are "about" parts of the query. Even in the work of several of the above authors in applying fuzzy subset theory to Boolean-like queries that have weights or thresholds on the queries, an attempt has been to preserve separability to as great an extent as is possible.

On the other hand, Salton and others have proposed and used for some years mechanisms which treat both the documents and the queries as vectors in a linear space. For example, Salton has done extensive work using a "cosine coefficient." Indexing the m terms of I as {t(j)}, one can then view each document d as a vector in an m-dimensional vector space; the vector has a 1 in the j-th coordinate iff f(d,t(j))=1. Queries

are similarly represented, so that the RSV is the cosine of the angle between the document vector d* and the query vector q*:

$$\frac{d^* \cdot q^*}{((d^* \cdot d^*) \quad (q^* \cdot q^*))^{**1/2}}$$

where · indicates the dot product of the vectors. [16] This sort of RSV is a measurement of "how close" a given document is to the query in that space. Some of these can be put in the form of topological metrics (distance functions); the cosine coefficient cannot. Further, some of these, the cosine coefficient included, do not have the same properties when the MF values are allowed to assume values other than 0 and 1. These differences in interpretation can indeed affect the use of the RSV's in computations measuring performance.

One problem which immediately arises in measuring performance is that, if the RSV's are no longer simply 0 and 1, then a new interpretation must be made of "about" and of "retrieved." The first problem is resolved if a set-theoretic, indeed fuzzy set-theoretic, interpretation can be placed on all numerical values involved. [7, 20] In a system in which RSV's are not simply 0 and 1, however, it is no longer the case that one would simply retrieve a subset of the documents. The user might instead be given information on a ranked list of documents, for example, and asked to specify a threshold above which to actually retrieve. Or, following the ideas of Cooper [6], the user might be given the ranked list and allowed to retrieve one document at a time until he decided that he had seen enough . There are several possibilities; the problem remains the same--the set RT is definable not by the RSV, but by system convention or, worse yet (from the point of view of predictability for use in numerical measurement), by user whim.

By a generalized retrieval system, then, we shall mean a system in which either

1. the indexing function is Boolean, the queries resemble Boolean queries, but the RSV's are not Boolean (this would include retrieval mechanisms such as the cosine coefficient); or

2. the indexing function is fuzzy, queries resemble Boolean queries, and RSV computations follow normal fuzzy subset rules (this would be a simply fuzzy-subset system, as described by Sachs, Tahani, and others [10, 11, 15, 17]); or

3. the indexing function is fuzzy, the queries have weights or thresholds attached to terms and/or subexpressions, and RSV computation is not necessarily simply a fuzzy-subset MF computation (this would include systems such as those suggested by Bookstein [1], Buell and Kraft [4, 5], Radecki [12, 13, 14], and others).

Such generalized retrieval systems would normally not retrieve a document subset simply from RSV computation; the RSV computation would provide a ranking on the documents. Actual retrieval would take place only after more information (retrieval thresholds, etc.) had entered the system.

4. PERFORMANCE MEASUREMENT FOR GENERALIZED SYSTEMS.

We first take up the question of recall, precision, and related measures. As has been pointed out, the crucial factor in computing these measures is that the sets of "retrieved" and of "relevant" documents are Boolean, so that cardinalities can be computed. In a fuzzy system (as in 2 or 3 above), these concepts no longer define Boolean sets. One can, however, look upon the RSV e as a fuzzy MF defining membership in the set which is "about" the query. Further, it will be necessary to have a human expert assign fuzzy membership functions that measure the extent to which a document is "relevant" to the query. This latter MF, which we shall call r(d) (or simply r), is the "true" fuzzy MF for document d. With e and r, then, we have two rankings on the documents D. The purpose of performance evaluation is to measure how closely e approximates r. With these two MF's, then, we can define the generalized recall and precision measures in a manner analogous to (4):

$$\text{recall} = \frac{\sum \text{Min} (e(d) , r(d))}{\sum r(d)} \qquad (5)$$

$$\text{precision} = \frac{\sum \text{Min} (e(d) , r(d))}{\sum e(d)}$$

(The summations extend over all elements d of D.)

In like manner, then, having defined fallout and generality, respectively, as

$$\frac{\text{number of documents retrieved and not relevant}}{\text{number of documents not relevant}}$$

and

$$\frac{\text{number of relevant documents}}{\text{number of documents}}$$

one can define their fuzzy analogues

$$\text{fallout} = \frac{\sum \text{Min}\,(\,e(d)\,,\,1-r(d)\,)}{\sum\,(1-r(d))} \qquad (6)$$

$$\text{generality} = \frac{\sum r(d)}{\sum 1}$$

We point out that some care must be taken in these definitions. The usual contingency table (as defined, for example, in Salton [16]) has Boolean cardinalities a, b, c, and d for the sets RL $\cap$ RT, RT $\cap$ D-RL, D-RT $\cap$ RL, and D-RT $\cap$ D-RL, respectively. Recall, for example, can then be defined as a/(a+c). If one were to use this definition of recall to generalize, the fuzzy "analog" would be

$$\text{recall} = \frac{\sum \text{Min}(\,e(d)\,,\,r(d))}{\sum \text{Max}(\text{Min}(e(d),r(d)),\ \text{Min}\,(1-e(d),\ r(d)))}$$

This function, however, does not represent the proportion of relevant documents retrieved to the total number of relevant documents.

Having defined the fuzzy analogs of recall and related measures, it remains to be seen that a justification exists for defining them in this way. This is gained from the natural generalization from Boolean to fuzzy sets. If recall and its related measures are defined as the quotient of "number of documents with property X" and "number of documents with property Y", and if both the RSV and the expert's ratings r(d) can be interpreted as fuzzy-subset membership functions, then the above definitions are the precise fuzzy analogs of the Boolean definitions.

In attempting to compare the use of these Boolean and fuzzy performance measures, we must consider the philosophy of fuzzy indexing, as it has a significant contribution to the values which will be assumed by these fuzzy measures. If a Boolean indexing function is used, then only those documents which are "about" a given term are given an index value of 1. There are two philosophies in performing such an indexing task. On the one hand, one could adopt a narrow view and only assign values of 1 if the document were truly "about" the term. The alternative view is a broad one, that even some small "aboutness" warrants the assigning of an index of 1 to the document. This dichotomy indeed reflects the traditional view of recall and precision as being negatively correlated. If the narrow indexing view is chosen, then certainly precision should be high, though recall may be low. If the opposite view is taken, then recall would be high and precision low.

In comparing a given set of documents indexed by both Boolean and by fuzzy functions, the dichotomy between Boolean indexings affects a comparison of the fuzzy and Boolean indexings. If the Boolean indexing was narrow, then documents which had a Boolean index of 0 would be likely to be given small, but non-zero, fuzzy index values. The fuzzifying of the indexing would then involve changing the index values for documents with Boolean indexes of both 0 and 1. If the Boolean indexing was broad, however, then the fuzzy indexing would change primarily those index values which had previously been 1. A previous index of 1, assigned to avoid missing relevant documents rather than to truly specify documents "about" a term, would now become a small, positive, fuzzy index value.

We point out that the subjective nature of the Boolean indexing is present even if the indexing is actually done automatically. An automatic indexing system, perhaps using term frequencies, probably would produce fuzzy values rather than Boolean ones [16] . It is only at the last stage of processing that a threshold would be applied to convert all indexes above some threshold to 1, all those below to 0. The difference between narrow and broad indexing is then a difference between a large threshold, perhaps somewhere around .75 or .8, and a small one, perhaps .1.

Having noted the effect that indexing philosophy can have on the index values assigned, it remains to be seen what effect this then has on the comparison between Boolean and fuzzy recall-like measures. If the Boolean indexing is narrow, one would expect that relatively few documents would be retrieved; that is, the set RT would be small. By converting to a fuzzy indexing function, the numerator in recall would increase. This would be true because presumably those documents with Boolean index of 1 would have fuzzy indexes near 1; the decrease in the summation would be more than offset by the addition of terms with smaller fuzzy index values. On the other hand, if the Boolean indexing were broad, so that the conversion to fuzzy indexes led essentially to a decrease in the positive index values, then the numerator in recall could be expected to decrease. If the relevance function were Boolean, then one would see recall values increase in changing from narrow Boolean indexing to fuzzy indexing, and decrease in changing from broad Boolean indexing to fuzzy indexing.

To illustrate the above remarks, consider the case of five documents indexed by a single term t, with fuzzy MF's as defined below:

doc	d1	d2	d3	d4	d5
MF	1.0	.75	.5	.25	0

Had these documents been indexed by a narrow Boolean MF, the values would probably have been

doc	d1	d2	d3	d4	d5
MF	1	1	0	0	0

Had the Boolean indexing been broad, however, the values would have been

doc	d1	d2	d3	d4	d5
MF	1	1	1	1	0

The difference, then, in the numerical interpretation of "about" varies from 2.0 total in the narrow case to 2.5 in the fuzzy case (the decrease from 1 to .75 from document d2 is more than offset by adding in the new non-zero values) to 4.0 in the broad case (in which several 1's become smaller positive values). The disparate sizes of numbers imply that substantial problems could arise in attempting to directly compare numerical performance measures.

One further complication can arise in trying to compare recall and precision measures if the generalized system is similar to those proposed by the authors [4, 5] or by Bookstein [1]. In this case, relevance weights or thresholds are assigned to each term in the query by the user, and the final RSV is a function of the fuzzy index values, the weights or thresholds, and the method for processing the "Boolean" query. The problem arises in that it will no longer be possible to interpret the RSV as an absolute fuzzy MF. The RSV's will express appropriate relative membership to other RSV's computed by the same system, but the actual numerical values will not be comparable on an absolute scale to other fuzzy MF's.

This last complication may lead to a substantial rethinking of the use of recall-like measures in performance evaluation of generalized retrieval systems. Indeed, such systems actually produce more information about documents by providing a ranked list of documents instead of simply a set of retrieved documents. If it is similarly possible to conveniently obtain the expert's ranking r, then the performance evaluation should reflect the greater quantity of information to be evaluated. This would imply the use of rank-order measures comparing either the actual numerical values of relevance (if the RSV's can be used as absolute measures) or simply the relative rankings (if the RSV's are only relative measures).

Salton [16] suggests a normalized recall measure

$$R_{norm} = 1 - \frac{\sum r(d) - \sum i}{n(N-n)}$$

and a normalized precision

$$P_{norm} = 1 - \frac{\sum \log r(d) - \sum \log i}{\log \frac{N!}{(N-n)!n!}}$$

where r(d) is the rank of document d based on e values in descending order. These assume a user searches until all documents in RT have been found and that all relevant documents are equal in value to the user. One could generate similar measures based on the ranks of the documents based on e versus r. However one does generate such measures, the key issue is the comparison of the two ranked lists.

5. CONCLUSIONS

We have seen that the usual performance evaluation measures of recall, precision, fallout, and generality have analogues in retrieval environments in which decisions about "retrieval" and "relevance" are no longer Boolean. Although problems do exist in interpreting the numerical values which will be obtained from these measures and in comparing those values to those obtained from Boolean retrieval systems, the problems can be overcome by taking into account the nature of the indexing function. Finally, we have raised the question as to whether rank-order comparison measures might not be more appropriate for evaluating those systems whose natural output consists of rank orderings of documents.

REFERENCES

1. Bookstein, A., "Fuzzy Requests: An Approach to Weighted Boolean Searches," Journal of the American Society for Information Science, v. 31, 1980, pp. 240-247.

2. Buell, D., "An Analysis of Some Fuzzy Subset Applications to Information Retrieval Systems," Fuzzy Sets and Systems. (to appear)

3. Buell,D., "A General Model of Query Processing in Information Retrieval Retrieval Systems," Information Processing and Management. (to appear)

4. Buell, D., and D. H. Kraft, "A Model for a Weighted Retrieval System," Journal of the American Society for Information Science, v. 32. (to appear)

5. Buell, D., and D. H. Kraft, "Threshold Values and Boolean Retrieval Systems," Information Processing and Management. (to appear)

6. Cooper, W. S., "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," American Documentation, v. 19, 1968, pp. 30-41.

7. Kaufmann, A., Introduction to the Theory of Fuzzy Subsets. Academic Press: New York, 1975.

8. Kraft, D. H. and A. Bookstein, "Evaluation of Information Retrieval Systems: a Decision Theory Approach," Journal of the American Society for Information Science, v. 29, 1978, pp. 1-40.

9. Lancaster, F. W. Information Retrieval Systems, 2nd. ed. Wiley: New York, 1979.

10. Negoita, C. "On the Application of the Fuzzy Sets Separation Theorem for Automatic Classification in Information Retrieval Systems," Information Science, vol. 5, 1973, pp. 279-286.

11. Negoita, C. and P. Flondor, "On Fuzziness in Information Retrieval," International Journal on Man-Machine Studies, v. 8, 1976, pp. 711-716.

12. Radecki, T., "Mathematical Model of Information Retrieval Systems Based on the Concept of Fuzzy Thesaurus," Information Processing and Management, v. 12, 1976, pp. 313-318.

13. Radecki, T., "Mathematical Model of Time Effective Information Retrieval Systems Based on the Theory of Fuzzy Sets," Information Processing and Management, v. 13, 1977, pp. 109-116.

14. Radecki, T., "Fuzzy Set Theoretical Approach to Document Retrieval," Information Processing and Management, v. 15, 1979, pp. 247-259.

15. Sachs, W. M., "An Approach to Associative Retrieval Through the Theory of Fuzzy Sets," Journal of the American Society for Information Science, v. 27, 1976, pp. 85-87.

16. Salton, G., Dynamic Information and Library Processing. Prentice-Hall: Englewood Cliffs, NJ, 1975.

17. Tahani, V., "A Fuzzy Model of Document Retrieval Systems," Information Processing and Management, v. 12 1976, pp. 177-187.

18. van Rijsbergen, C. J., Information Retrieval, 2nd. ed. Butterworths: London, 1979.

19. Waller, W. G. and D. H. Kraft, "A Mathematical Model of a Weighted Boolean Retrieval System," Information Processing and Management, v. 15, 1979, pp. 235-245.

20. Zadeh, L., "Fuzzy Sets," Information and Control, v. 8, 1965, pp. 338-353.