III. A New Evaluation Measure

J. Joiner and L. Werner

Abstract

The problems of evaluation and the needed criteria of evaluation measures in the SMART system of information retrieval are reviewed and discussed. Performance characteristics of a good evaluation measure are examined. The suggested measure $Pr_{H,N}(P<R/n)$, (the probability under the hypergeometric distribution, that the precision could be strictly less than that precision attained, where R = number relevant in the sample drawn, N = total number in collection and n = size of sample drawn) is introduced and tested against the various criteria needed for a good evaluation measure. A statistical test of significance is explained.

1. Introduction

Among the principal obstacles to the evaluation of information retrieval methods are the following:

1) Interpolation between points of recall results in errors which are unsatisfactory in one way or another, depending upon the type of interpolation used.

2) A recall-precision curve sometimes gyrates wildly and the averaging of many curves over queries has questionable reliability.

3) The statement "method A is better than method B" often depends upon the value one is measuring. A unique value measuring both recall and precision would be best.

4) Queries with different numbers of relevant documents do not receive different amounts of credit, although just by random

chance it is easier to get a relevant document for a query with 30 relevant than for one with 4 relevant.

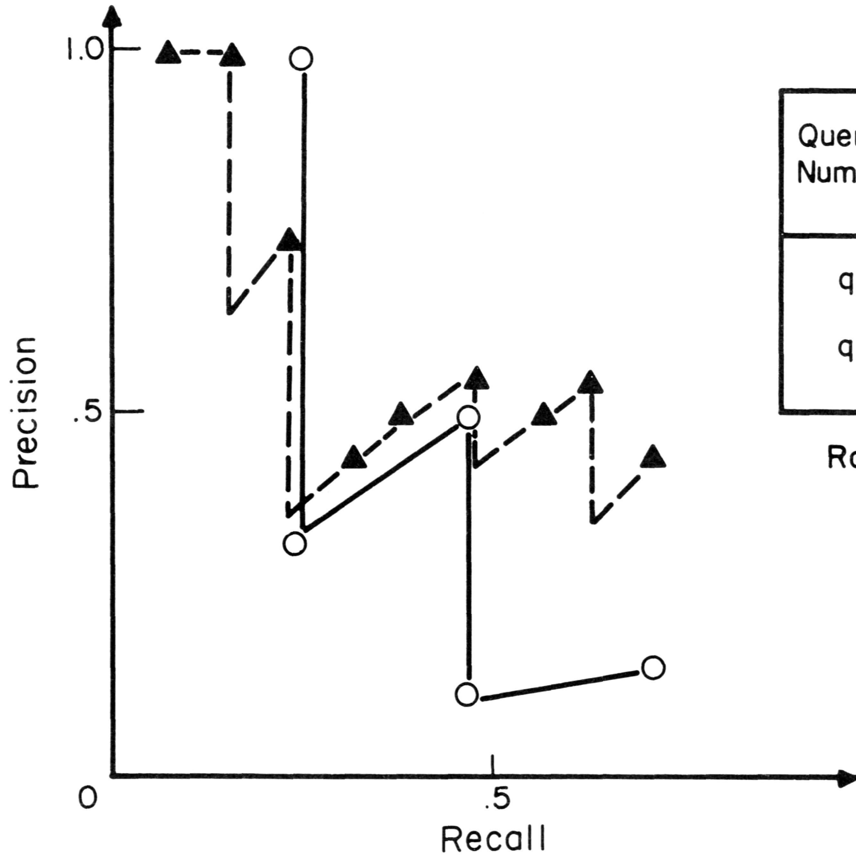
5) It has not been determined how to handle the evaluation of feedback methods in relation to the relevant documents retrieved before feedback.

This report will attempt to discuss and propose solutions to these specific problems.

2. Problems of Evaluation

One important aspect of information retrieval is obtaining a value for a given method which is a true measure of the method's effectiveness over many queries. On a recall-precision graph, the points of recall where one measures this effectiveness are .1, .2, . . . , .9, 1.0. However, the only points available for a query with n relevant documents are 1/n, 2/n, . . . , n-1/n, n/n. Obviously, for queries with different numbers of relevant documents, one may expect that none of the query's points will coincide with the points .1, .2, . . . , .9, 1.0. But presently, by interpolation of some kine, the precision values are found for each query at these points. Fig. 1 shows one such method. There can be no real justification for any method of interpolation used, for it is impossible to estimate a discrete function at a nonexistent point. Therefore, what is needed to solve this problem is a new base index for the graph that would involve no interpolation. An index that would, over many queries with different numbers of relevant documents, have only common points for all queries.

The averaging of points of recall-precision where interpolation must occur tends further to distort the measure of effectiveness. But even over



Query Number	Total No. of Relevant Documents	Ranks of Relevant Documents
q_1	4	1, 4, 20
q	12	1, 2, 4, 9, 11, 14, 15, 20

Ranks of Relevant in Top 20

Recall – Precision Graph for two Queries

Fig. 1

points that coincide with equal recall, different values can be obtained by different methods of averaging. Even at these common points, the values being averaged are somewhat in doubt. No correlations are made in the precision values for the generality number (G = number of Relevant/total number in collection), which reflects how easy it would be, under random conditions alone, to select relevant documents. A good performance measure should control this randomness factor. Control should be in the sense that when the generality ratio is decreased in a way which preserves the observed performance level, the effect of the generality ratio on a performance measure could be observed. The measure proposed is known to reflect the generality number under equal performance but a method of splitting a collection into two collections suggested by R. Williamson has not been tested.

3. Criteria for a Good Evaluation Measure

A good performance measure should fulfill the following criteria:

1) Recall values measure the effectiveness of a method by comparing the number of relevant documents retrieved to to total number of relevant documents, while precision measures this performance by comparing the number of relevant retrieved documents to the total number of retrieved. These intuitively seem to be the best measures of performance available. Their biggest drawback is that they are two unique values not one. A good measure should reflect both.

2) The generality number, as stated before, reflects the degree of effect that pure random chance selection will have on the method of retrieving relevant documents for a certain query. With this controlled queries can be com-

pared on a more common basis.

3) In theory (at the least), the measure should appeal to the user and tester and the values obtained should have a logical range. A range from 0 to 1 best suits a measure of performance and effectiveness. Any system which is effective at all should have values of the measure closer to 1 than to 0.

4. The Probability Measure

A very large urn is filled with 200 documents. For query q_o there are 20 documents that are relevant and 180 that are not. If, at random, 20 documents are drawn from that urn without replacement, the probability that less than 3 relevant documents are chosen completely at random is

$$P_{H,200}(R<3) = P_{H,200}(R = 2) + P_{H,200}(R = 1) + P_{H,200}(R = 0)$$

$$= \frac{\binom{20}{2}\binom{180}{18}}{\binom{200}{20}} + \frac{\binom{20}{1}\binom{180}{19}}{\binom{200}{20}} + \frac{\binom{20}{0}\binom{180}{20}}{\binom{200}{20}} \quad .$$

This is equivalent to finding the probability by random chance that the precision is less than 3/20 for $P_{H,200}(R<3) = P_{H,200}(R/n<3/n) = P_{H,200}$ $(P<3/20)$. The higher this probability is, the less likely it would be that the precision achieved was obtained by chance. This measure could be evaluated at any point n (equals the number of documents retrieved) that might be wanted for investigation.

As precision increases from m/n to $(m + 1)/n$, this value goes from $P_{H,200}(P<m/n)$ to $P_{H,200}(P<(m + 1)/n)$ which is equal to $P_{H,200}(R<m)$ and $P_{H,200}(R\ m + 1)$ where

$$P_{H,200}(R<m) = P_{H,200}(R = m-1) + P_{H,200}(R = m-2) + \cdots + P_{H,200}(R = 0)$$

and

$$P_{H,200}(R<m+1) = P_{H,200}(R = m) + P_{H,200}(R = m-1) + \ldots + P_{H,200}(R = 0).$$

When $P_{H,200}(R<m)$ is subtracted from $P_{H,200}(R<m+1)$ the answer is always positive since one more single hypergeometric probability is added to $P_{H,200}(R<m+1)$. Since $P_{H,200}(R<m) < P_{H,200}(R<m+1)$ is equivalent to stating that $P_{H,200}(R/n<P<m/n = \hat{p}_1)$ $P_{H,200}(P<m+1/n = \hat{p}_2)$, and $\hat{p}_1<\hat{p}_2$, then as precision increases the performance measure increases.

This same argument holds for recall because

$$P_{H,200}(r<m) = P_{H,200}(r/R<m/R) = P_{H,200}\ (recall<m/R) .$$

As the recall increases from m/r to $m+1/r$ more probability is added to the measure and it therefore increases.

The probability itself incorporates the generality number and it will be shown by example how this generality affects the measure. All three of the criteria which are most needed by a unique performance measure are therefore combined in this value. The theoretical range, 0-1, of this measure is also appealing to testing procedures and analyzing of results. Some measures for arbitrarily chosen results are shown in Table 1.

The use of this measure for feedback is the same as without feedback except that when the ranks of the relevant documents retrieved in the first pass are frozen the measure adjusts for this by use of a new generality number. Suppose for a single query and two methods

number of documents = 200

number of relevant documents = 12

Ranks of Relevant Documents:
1, 2, 3, 10, 11, 14, 15, 20, 40, 50, 69, 78.

Number Relevant	Number Drawn	Measure
1	1	0.94000
2	2	0.99668
3	3	0.99983
3	4	0.99935
3	5	0.99844
3	6	0.99698
3	7	0.99490
3	8	0.99212
3	9	0.98859
4	10	0.99868
5	11	0.99988
5	12	0.99980
5	13	0.99968
6	14	0.99997
7	15	0.99999
7	16	0.99999
7	17	0.99999
7	18	0.99999
7	19	0.99998
8	20	0.99998
8	21	0.99998
8	22	0.99997
8	23	0.99997
8	24	0.99996
8	25	0.99995
8	26	0.99994
8	27	0.99992
8	28	0.99990
8	29	0.99988
8	30	0.99985

Performance Results for up to
30 Retrieved Documents

Table 1

Suppose that in the first 10 documents Method I produces 4 relevant and Method II 2 relevant. Then evaluation starting with this information on a feedback pass would evaluate the measure as

Method I Conditions	Method II Conditions
$n = 190$	$n = 190$
number relevant = 8	number relevant = 10

Performance would thereafter reflect exactly the same measures as if conditions for Methods I and II were starting conditions.

5. Tests

One method of comparing two or more methods over the same set of queries in the same document collection would be to average the measure over the number of documents retrieved. This procedure would give one number for each method and the highest such number could be stated to represent the best method.

The difficulty with this method is that there is no way to know the statistical properties of this average and therefore slight differences in the average of method i vs. method j cannot be proven significant. With a fixed set of queries and a fixed collection there is no randomness involved anyway.

Randomness can be introduced into the problem by claiming that the queries are a sample drawn from a set of queries and that the test results show that at any point n a population of queries divides into a multinomial distribution where method i has probability p_i of being the most successful. This procedure is discussed in May's thesis. Table 2 shows the suggested partition of queries and methods over n, the number of documents retrieved.

Table 2 also shows a fictitious set of results. There is no hope of being correct in a decision if in reality the methods are exactly alike, so

Three Methods M_1 M_2 M_3 Five Queries Q_1 Q_2 Q_3 Q_4 Q_5

At point $n_1 = 1$, 1 document retrieved

Value 1 given to method which has highest value. In case of a tie at some point, choose one of the tied methods by chance.

Example:

	M_1	M_2	M_3
Q_1	0	1	0
Q_2	1	0	0
Q_3	0	0	1
Q_4	0	0	1
Q_5	0	0	1

For each $n_i = i$, i documents retrieved, sum over queries for each method

Example:

	M_1	M_2	M_3	
$N_1 = 1$	1	1	3	5
$N_2 = 2$	1	2	2	5
$N_3 = 3$	0	2	3	5
.				
.				
.				
$N_{20} = 20$	1	3	1	5
:				
$N_{200} = 200$	2	1	2	5

Again, sum, over n_i this time, for total for methods.

Total:

M_1	M_2	M_3	
147	320	533	1000

Estimate:

$$\hat{P}_1 \text{ for } M_1 = \frac{147}{1000} = .147$$

$$\hat{P}_2 \text{ for } M_2 = \frac{320}{1000} = .320$$

$$\hat{P}_3 \text{ for } M_3 = \frac{533}{1000} = .533$$

Sample Calculation

Table 2

one can only state the "probability" of being correct in choosing method 3 in the example given if the ratio p_3/p_2 $(=\theta)$ is actually greater than some θ specified by the experimenter.

For the example given assuming there is a multinomial distribution (which is unlikely) and further that $p_3/p_2 = 1.5$, then the probability that the choice of method 3 is best is over .98, using Bechhofer's procedures. It should be stressed that this is not to claim a valid statistical test but only to give some idea of the possible confidence one could have in choosing the largest p_i as representing the best method.

Bibliography

Bechhofer, R. E., Elmaghraby, S., and Morse, N., "A Single-sample multiple Decision Procedure for Selecting the Multinomial Event which has the Highest Probability", Annals of Mathematical Statistics, Vol. 30, No. 1, March 1959.

Cooper, W. S., "Expected Search Length — A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems", American Documentation, January 1968.

Goffman, W., and Newill, V. A., "A Methodology for Test and Evaluation of Information Retrieval Systems", Information Storage and Retrieval, Vol. 3, 1966.

Hodges, J. L., and Lehmann, E. L., Basic Concepts of Statistics, Holden-Day, San Francisco, 1964.

Lesk, M. E., "SIG — The Significance Programs for Testing the Evaluation Output", Report No. ISR-12 to the National Science Foundation, Section II, Cornell University, Department of Computer Science, 1967.

May, C., "Evaluation of Search Methods in an Information Retrieval System", an unpublished thesis for Master's of Arts degree, June 1968.

Salton, G., and Lesk, M.E., "Computer Evaluation of Indexing and Text Processing", Report No. ISR-12 to the National Science Foundation, Section III, Cornell University, Department of Computer Science, 1967.

Williamson, R. E., "A Proposal to Ascertain the Relationship between the Generality Ratio and Performance Measure", unpublished paper.