XXI. EVALUATION VIEWPOINTS IN DOCUMENT RETRIEVAL

J. Rocchio

Summary

The fundamental notions involved in measuring the performance of a document retrieval system are examined, and it is concluded that there are two different viewpoints which may be taken depending on one's basis assumptions. The rationale for each of these is examined and the differences between them is discussed. Some illustrations are given with respect to the effect of the evaluation viewpoint on precision vs. recall measures.

1. Introduction

A document retrieval system operating on a retrieval request produces a partition of a source document collection into two disjoint subsets, the retrieved subset and the nonretrieved subset. The basis for judging the performance of the system is a comparision between this partition and the one which the originator of the query would produce if he were to examine every source document. In most cases such a direct comparision is not feasible and thus various artifacts (such as sampling, use of contrived queries, etc.) are used to approximate the ideal case. However in all instances the underlying basis for system evaluation remains the same.

In some retrieval systems a retrieval operation may produce a ranking or even a metric ordering of source documents with respect to the input query. In these cases questions of "degree of relevance" can be avoided by inducing a partition from the ordering using a cutoff criterion, or by requiring that each relevant document (one which the user would retrieve) be ranked above every member of the nonrelevant set. In this latter case then the basis for evaluation is a comparison between the ordering induced by the system and the ordering induced by the user rather than between the respective partitions.

In either case, if we assume that the set of documents relevant to each input query is both known and fixed, any performance evaluation is a measure of (a) the indexing transformation which produces the images manipulated by the system; (b) the query image (since we assume that the query structure for a fixed relevant set may vary), and (c) the search or matching function which operates on these images to produce an effective partition of the source document collection.

While the above outlines the basic mechanics of retrieval system evaluation, it neglects an important aspect of evaluation which has often been overlooked, namely the evaluation viewpoint. Under the above assumptions about the system model there still remains an important choice to be made by the evaluator. Since the notion of relevance is so complex and so dependent on personal factors, any system evaluation must clearly be statistically based. Thus, in effect, any evaluation of a document retrieval system must in fact be an estimate of the degree to which the systems ability to detect relevance matches that of humans.

The validity of such an estimate based on the system performance with respect to some finite sample set of queries depends on how representative the sample set is of the real life environment of the system. It is not however this question to which we address ourselves but rather to a related one, namely whether the evaluation viewpoint considers retrieval queries as atomic, or whether the relation between a relevant document and the set of nonrelevant documents (implicitly defined by each query for all members of the document set relevant to that query) is considered atomic. These alternative evaluation viewpoints called macro and micro evaluation respectively are both tenable, and in fact have been compared one against the other without the comparers' appreciating the distinction.

2. Macro Evaluation

Macro evaluation is a query oriented viewpoint, and as such any performance index formulated on this basis would be query distributed or averaged on a per-query basis. The justification for considering queries as atomic is simply that this corresponds to the viewpoint of the system user. The user interacts with the retrieval system via his retrieval request, and would thus judge the usefulness of the system on the basis of its performance with respect to his request. The sample mean of an evaluation parameter obtained by averaging over the total number of queries represents an estimate of the worth of the system to the average user. An underlying assumption of this approach is of course that the distribution of results over the sample set of queries allows a meaningful average performance per-query to be produced.

If the performance distribution is too wide, one might try to categorize requests in some fashion so as to produce a meaningful performance estimate for each category. This of course would still be done on a query average basis.

3. Micro Evaluation

Micro evaluation stems from a document oriented viewpoint, in which one assumes that the determining element of system behavior is the relation between a relevant document and the set of nonrelevant documents. With this viewpoint each query provides a set of samples of this relation, and thus any performance index is document distributed or averaged on a per document basis. This approach makes it necessary to justify that the set of samples provided by a single request are statistically independent and reflective of the relevance relation which pertains in general. If this is not true, the micro evaluation viewpoint will weight any performance index so derived heavily towards the systems behavior for those requests having a large number of relevant documents, thus distorting the statistical validity of the estimate. It is not however, the primary purpose of this discussion to establish which of these evaluation orientations has more merit, but rather to point out the differences so that the implications of each are clearly understood.

4. Example

As an illustration of the effect of the evaluation viewpoint on performance indices we consider now the precision-recall curve of Cleverdon which has been perhaps the most widely used evaluation tool for document retrieval systems. As Cleverdon has conceived and used this measure it is a micro evaluation performance index.

Define recall as the number of relevant documents retrieved divided by the total number of relevant documents, and precision as the number of relevant documents retrieved divided by the total number of documents retrieved. Clearly these parameters of a retrieval operation assume that the matching function of the system induces a dichotomus partition of the source document set. Each ratio is indicative of a different aspect of how this partition induced by the system compares with that induced from the users' relevance judgements. The recall ratio measures the inclusiveness of the retrieved subset with respect to the relevant set, and the precision measures the exclusiveness of the retrieved set with respect to the nonrelevant set. Cleverdon's approach was to plot these ratios, one versus the other with a parameter of the matching function (cutoff criterion) providing different points.

The evaluation of a single request by this means is clearly unambiguous. However, the evaluation of system behavior with respect to a sample set of N queries can be produced from either the micro or the macro viewpoint.

Assume that for a particular cutoff (which specifies a partition of the source collection), the total number of retrieved documents for the ith query is t_i, and the number both relevant and retrieved is r_i. Further let the total number of documents which are relevant to request i be n_i.

Under these conditions a single point on the precision-recall curve is defined according to the micro viewpoint as:

$$\text{micro recall} = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} n_i} , \qquad \text{micro precision} = \frac{\sum_{i=1}^{N} r_i}{\sum_{i=1}^{N} t_i} .$$

Thus the micro recall is the number of relevant documents retrieved per relevant document, and micro precision is the number of relevant documents retrieved per retrieved document. Unless there is only small variation of both n and t over all sample requests these parameter do not reflect the behavior of an average request.

From a query oriented point of view the above conditions define N rather than one point in the precision-recall plane. This set of points can be represented by a single average point defined by:

$$\text{macro recall} = \frac{1}{N}\sum_{i=1}^{N} \frac{r_i}{n_i} ; \qquad \text{macro precision} = \frac{1}{N}\sum_{i=1}^{N} \frac{r_i}{t_i} .$$

Thus the macro recall is just the average recall over the query sample, while the macro precision is the corresponding average precision.

It is interesting to note that if the number of relevant documents per query is constant or the number of retrieved documents per query is constant then the micro and macro recall or precision are respectively identical. This of course is expected since in this event the micro statistic merely weights the performance of each query equally and is therefore identical to the macro statistic.

A numerical example may serve to illustrate the differences in these approaches more concretly. Consider a set of two queries for which the following table describes the hypothetical system behavior.

Query		Cutoff 1		Cutoff 2	
	n	t	r	t	r
1	10	3	2	20	6
2	3	3	2	60	2

Hypothetical Retrieval Results for
A Sample of two Queries

TABLE 1

For the above retrieval results the following recall-precision figures indicate the micro and macro parameter values.

	Cutoff 1		Cutoff 2	
	micro	macro	micro	macro
recall	.31	.43	.62	.63
precision	.66	.66	.1	.17

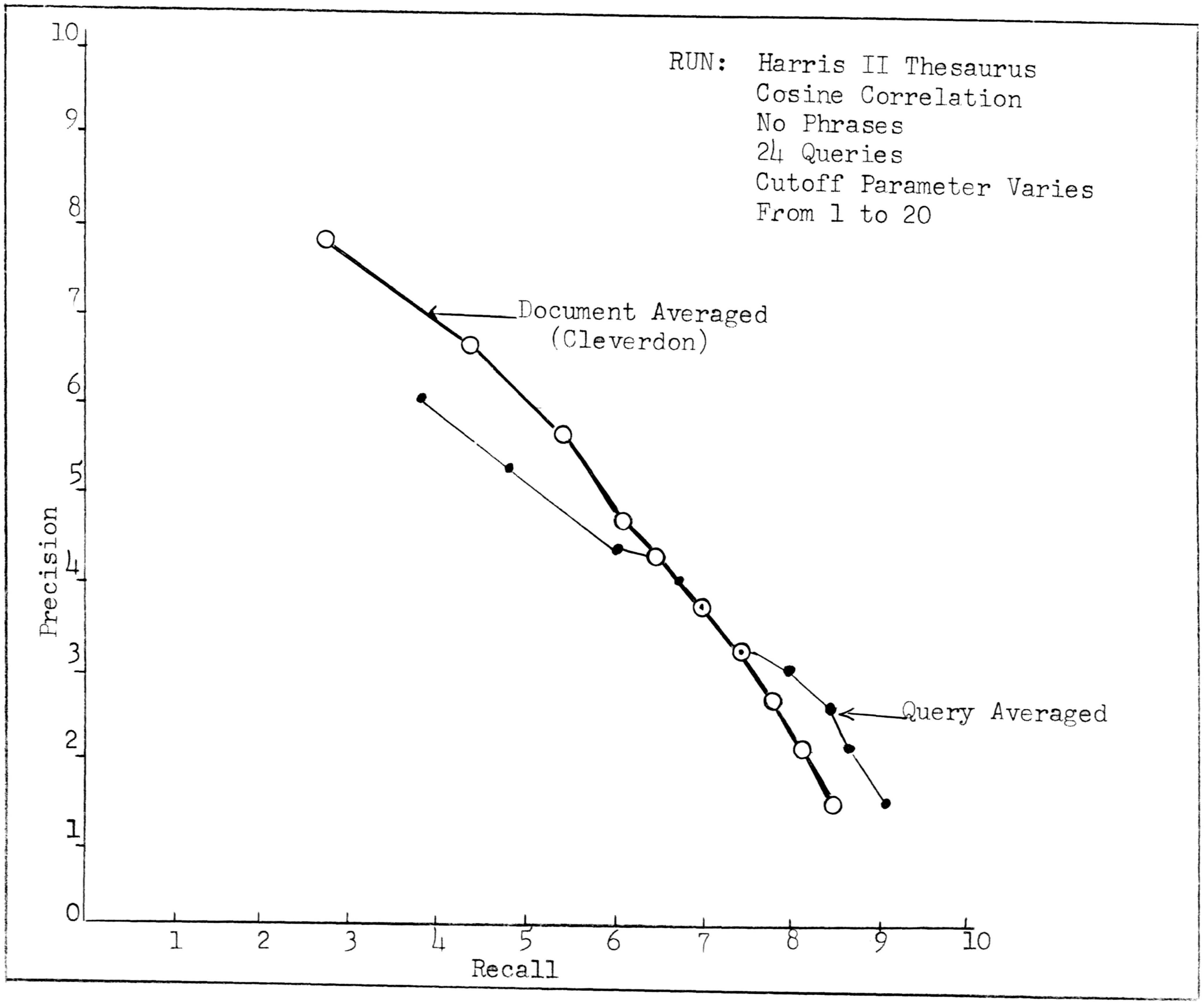
Comparison of Micro and Macro Precision-Recall Points

TABLE 2

From Table 2 it can be noted that at cutoff 1 micro recall is significantly lower which reflects the fact that the request having the large number of relevant documents (1) has much lower recall than request (2). Similarly at cutoff 2 micro precision is lower since in this case the request having more retrieved documents (2) has lower precision.

A comparision between query-averaged and document-averaged precision vs. recall curves for a sample set of 24 queries is shown in Fig. 1. The results were obtained from experiments run on the SMART system.[*] Since the matching function in SMART produces a ranked output, cutoff points were chosen by the following method: let the user examine the retrieved list in order until he encounters n consecutive nonrelevant documents; this defines a cutoff point; by varying n, a series of points on the precision vs. recall curve can be produced.

[*] Information Storage and Retrieval, Report No. ISR-8 to the National Science Foundation, Computation Laboratory of Harvard University (December 1964).



Precision vs. Recall for 24 SMART system

Figure 1

Note that although this method of specifying a cutoff is not ideal, since it ignores any metric information contained in the retrieved ordering, it might correspond quite closely to the way a user would react in practice to the system.

5. Conclusion

The only generalization which can be made with respect to the effect of these alternative viewpoints on a precision vs. recall curve is that recall results derived from the micro viewpoint will be more representative of the system behavior on queries having a large number of relevant documents, and that micro precision will be more representative of the system behavior on queries having a large number of retrieved documents. Without evidence that this in fact is not biasing the performance estimate for the retrieval system it would seem that the macro viewpoint has more merit.