III. PERFORMANCE INDICES FOR DOCUMENT RETRIEVAL SYSTEMS

Joseph Rocchio

Summary

The SMART system is designed to provide a flexible test bed for evaluating a variety of potentially useful methods of automatic content analysis in information retrieval. SMART is primarily a document retrieval system; however, many of the analysis techniques employed are also applicable to other phases of information retrieval, including fact retrieval and question-answering systems. This section introduces a generalized model of the system and derives several evaluation indices which will be used to determine the influence of the various system parameters.

1. The Model

A document retrieval system employing fully automatic indexing may be characterized by the following elements:

(a) a set of reference documents in the natural language (D);

(b) a set of retrieval requests in the natural language (Q);

(c) an index language L;

(d) a transformation T from the natural language to the index language which operates on members of the set D;

(e) a transformation T' (possibly the same as T) which maps retrieval requests to the index language L; and

(f) a search function S whose domain is the cardinal product of elements of the set Q with those of the set D, and whose range is such that S induces at least a partial order on the set D.

In the SMART system the index language L may be considered to be a property space. The transformations T and T' may take several forms. For example, natural language word stems may be mapped one-to-one into elements of L or, alternatively, T may be a composite transformation resulting from a many-to-many mapping of word stems to elements of L (thesaurus transformation) followed by mappings from L to L (hierarchy). In addition, the resulting image of a document $d_i \in D$ under a given transformation T can be either a binary property vector in L or a numeric vector. The search function S used in SMART may also be controlled, but is basically characterized by a correlation process involving a request image and the set of reference document images.

A retrieval operation in terms of this model consists in applying the search function S to the cardinal product $T(D) \times T(q)$, $q \in Q$. One may consider the ordering induced on D from the range of S to be the result of the retrieval operation, or one may introduce a decision function C whose domain is the range of S and whose range is the positive integers. Such a decision function partitions D into disjoint subsets, normally consisting of a retrieved subset and its complement with respect to D. In the general case C may, however, introduce a multi-level classification.

To evaluate the effectiveness of a retrieval operation, it is necessary to introduce a subjective element. Let us assign to each request

q, a subset D_q of D which is the set of reference documents "relevant" to q. The specification of this subset for a given q may in general be an ill-defined process. In an operational framework, D_q is that subset of R which the originator of the request q would choose if he were given the opportunity of making an exhaustive search on D. Alternatively, one may consider that corresponding to each request q, an ordering of D is defined which reflects the "degree of relevance" of each document in R to the request. In this case one may still identify a relevant subset D_q by considering those members of D for which the degree of relevance exceeds a given threshold.

Assume for the present that for each q in Q a subset D_q of relevant documents is known or, alternatively, a partial ordering of D exists which reflects the degree of relevance. The object of the document retrieval system is to produce a subset D'_q (or induce a partial order on D) which is identical to D_q (that is, equivalent to the partial order determined by the degree of relevance). Evaluation of a retrieval system then requires a determination of how each of the system elements affects the degree to which this objective is met for all members of Q.

The most commonly used performance indices of document retrieval systems are the recall and relevance ratios introduced by Cleverdon in connection with the ASLIB-Cranfield project.[1] These measures are based on having an objective set D_q of relevant documents for each query presented to the system. After a retrieval operation which produces a two-way classification of D into a retrieved subset D_a and its complement with respect to D, the following parameters are obtained:

(a) $n(D_q)$ = total number of relevant documents;

(b) $n(D_a)$ = total number of retrieved documents; and

(c) $n(D_q \cap D_a)$ = total number of relevant retrieved documents.

Using these parameters one can define:

$$\text{recall} = \frac{n(D_q \cap D_a)}{n(D_q)}$$

and

$$\text{relevance (also sometimes called "precision")} = \frac{n(D_q \cap D_a)}{n(D_a)}.$$

Clearly, recall as defined by Cleverdon is a measure of the inclusiveness of the set D_a with respect to the set D_q, while relevance is a measure of the exclusiveness of the set D_a with respect to the complement of D_q. It should be noted that the joint behavior of these parameters is required to judge performance intuitively, i.e., a recall of 1 or a precision of 1 alone does not imply satisfactory performance; however, if both recall and precision are 1, then $D_a = D_q$.

It must also be noted in connection with these parameters that the decision function C of the model is required to specify the retrieved subset D_a. In many respects this is undesirable because an additional variable is then introduced into the system. In fact, when evaluating the various content analysis techniques, including the structure of the index language L and of the transformations T and T', it is desirable to introduce as few extraneous constraints as possible. This suggests that one should deal

directly with the search function S. Another justification for so doing is the fact that the decision function is usually determined subjectively, in the sense that in practice the needs of a particular user dictate its characteristics.

In practice, the decision function C can easily be eliminated, because the search function S can be used to induce a partial ordering on D directly; a user could then request that the results of the retrieval operation be presented to him in this induced order. If this were done, the user could examine any desired subset of this ordered set, specifying, in effect, the "retrieved subset" a posteriori by the number of documents examined.

In view of these considerations a set of performance indices has been developed which may be applied directly to the ordering induced on D by a retrieval search S.

2. Evaluation Indices

Under the assumption that the ordering induced on the set of reference documents by the search process S is the principal result of a retrieval operation and that a set of relevant documents D_q is available corresponding to each request q, the objective of a retrieval operation may be recast in the following form: a retrieval operation with respect to a request q is expected to produce an ordering on the reference collection D, such that every member of the set D_q is ranked above all members of the complement of D_q with respect to $D(\bar{D}_q)$.

Note that in this formulation no emphasis is placed on any relative order among the members of the set D_q of relevant documents. While such an ordering might in theory seem desirable, the determination of an unordered set D_q is difficult enough by itself, so that imposition of an additional ordering criterion may be impractical. A partial order within D_q may, however, have some significance and, in fact, has been employed in some of the ASLIB-Cranfield experiments to specify degrees of relevance. These in turn led to the definition of different subsets D_q, but not to the specification of retrieval order with respect to relevance order.

Given the previously stated definition of the objective of a retrieval operation, two functions of the ordering induced on D may be defined which are related to the recall and relevance (precision) of Cleverdon. Consider an ordering induced on D by S such that a one-to-one mapping exists from D to the dense set of integers from 1 to n(D); increasing rank order in the set of integers then reflects decreasing connection between the request image and document image.

In this case, define:

$$r^*(i) = \begin{cases} \frac{i}{n_o} \text{ for } 1 \leq i \leq n_o \\ 1 \text{ for } n_o \leq i \leq N \end{cases}$$

and

$$p^*(i) = \begin{cases} 1 \text{ for } 1 \leq i \leq n_o \\ \frac{n_o}{i} \text{ for } n_o \leq i \leq N \end{cases},$$

where

$n_o = n(D_q)$, i.e., the number of relevant documents to the query under consideration;

$N = n(D)$, the number of documents in the reference collection; and

i = the rank index induced on D.

The function $r(i)$ is viewed as the number of relevant documents having rank order less than or equal to i divided by the total number of relevant documents. Thus, it is Cleverdon's recall as a function of the order induced on D by a retrieval operation. Clearly, $r^*(i)$ is the recall function which pertains when the retrieval operation produces an ideal ordering on D. Similarly, $p(i)$ is the number of relevant documents having rank order less than or equal to i divided by i, with $p^*(i)$ defined for the case when all members of D_q have a rank index less than every member of $\bar{D}_q$.

Hence for each query q, $r^*_q(i)$ and $p^*_q(i)$ define a desired (or objective) recall function and a desired precision function.

Since it has been assumed that S induces only an ordering on D, as opposed to a metric, these functions are strictly defined only for discrete values of the rank index i. As it is intended to extend these functions to a continuous independent variable, that is, to define a function $r^*(x)$ equivalent to $r^*(i)$, a possible anomaly is noted. This arises from the fact that it is possible, within the framework of the system, for S to produce a mapping from elements of $\{q \times D\}$ to the real line. This, in fact, occurs

when S is a correlation process which correlates a query image with the set of document images viewed as vectors in some abstract space. The process of inducing an ordering from this mapping and then treating this ordering as function of a continuous real variable gives the impression of coming full circle. In fact, there is clearly a loss of information involved since relative distance between the images of d_i and d_j is not preserved by this process. The justification for making this transformation from the domain of S to an ordering index lies in the assumption that the order so derived has significance of and by itself.

The extension then to functions of a real variable is accomplished by defining two functions $r^*(x)$ and $p^*(x)$ such that:

$$\left.\begin{aligned} r^*(x) &= r^*(i) \\ p^*(x) &= p^*(i) \end{aligned}\right\} \text{ for } x = i,\ i = 1,2,\ldots,N;$$

and further that:

$$r^*(x) = \begin{cases} \dfrac{j}{n_o} \text{ for } j \leq x \leq j+1 \\ \qquad j \text{ integral and less than } n_o; \\ 1 \text{ for } x > n_o; \end{cases}$$

and

$$p^*(x) = \begin{cases} \dfrac{j}{x} \text{ for } j \leq x < j+1 \\ \qquad j \text{ integral and less than } n_o; \\ \dfrac{n_o}{x} \text{ for } x > n_o . \end{cases}$$

At this point, recall and precision functions may be defined for the results of a retrieval operation with respect to a particular query. In particular, let the ranks of each member of the set of relevant documents D_q resulting from applying S to $\{q \times D\}$ be specified as:

$$O(i) \quad \text{for } i = 1,2,\ldots,n_o,$$

where $O(i+1) > O(i)$.

In this case:

$$r_q(x) = \begin{cases} 0 \text{ for } 1 \leq x \leq O(1); \\ \frac{i}{n_o} \text{ for } O(i) \leq x < O(i+1); \\ 1 \text{ for } x \geq O(n_o); \end{cases}$$

and

$$p_q(x) = \begin{cases} 0 \text{ for } 1 \leq x \leq O(1); \\ \frac{i}{x} \text{ for } O(i) \leq x < O(i+1); \\ \frac{n_o}{x} \text{ for } x \geq O(n_o). \end{cases}$$

At this point, a recall error and a precision error may be defined by considering:

$$\text{recall error} = \int_{x=1}^{N} (r^*(x) - r_q(x))dx;$$

and

$$\text{precision error} = \int_{x=1}^{N} (p^*(x) - p_q(x))dx.$$

Since $r^*(x)$ is an upper bound to $r(x)$ and, similarly, for $p^*(x)$ and $p(x)$, these errors are always greater or equal to zero.

To compute these integrals we introduce the unit step function $U_{-1}(x)$ defined by:

$$U_{-1}(x) = \begin{cases} 1 & \text{for } x \geq 0; \\ 0 & \text{for } x < 0; \end{cases}$$

and note that:

$$\int_{\infty}^{b} U_{-1}(x)dx = b .$$

Now $r^*(x)$ can be expressed as:

$$r^*(x) = \frac{1}{n_o}\left[U_{-1}(x-1) + U_{-1}(x-2) + \cdots + U_{-1}(x-n_o)\right],$$

and

$$r(x) = \frac{1}{n_o}\left[U_{-1}(x-0(1)) + U_{-1}(x-0(2)) + \cdots + U_{-1}(x-0(n_o)\right].$$

Therefore,

$$\int_{1}^{N} (r^*(x) - r(x))dx = \frac{1}{n_o}\sum_{i=1}^{n_o} \int_{1}^{N} \left[U_{-1}(x-i) - U_{-1}(x-0(i)\right]dx$$

$$= \frac{1}{n_o}\sum_{i=1}^{n_o} \left[0(i) - i\right]$$

$$= \frac{1}{n_o}\sum_{i=1}^{n_o} 0(i) - \frac{1}{n_o}\sum_{i=1}^{n_o} i$$

or

$$\text{recall error} = \bar{O} - \frac{n_o + 1}{2};$$

i.e., the integral of the difference between the recall function for perfect retrieval and the recall function which results for an actual retrieval is just the difference between the average rank $\bar{O}$ induced on the members of the set of relevant documents D_q by the retrieval operation, and the mean of the ranks which a perfect retrieval would induce.

To normalize this parameter to the range 0 - 1, consider the case for which the rank of every member of the set D_q is greater than every member of $\bar{D}_q$. This is clearly the case of maximum error; hence:

$$\text{max recall error} = \frac{1}{n_o} \sum_{i=1}^{n_o} N - (i - 1) - \frac{n_o + 1}{2}$$

$$= \frac{1}{n_o}\left[\frac{n_o}{2}(N + N - n_o + 1)\right] - \frac{n_o + 1}{2}$$

$$\text{max recall error} = N - n_o .$$

Therefore,

$$\frac{\bar{O} - \frac{(n_o + 1)}{2}}{N - n_o}$$

is a normalized index of over-all recall error. As this index is measuring recall error, it is desirable to reverse it. Hence:

$$1 - \left[\frac{\bar{O} - \left(\frac{n_o + 1}{2}\right)}{N - n_o}\right]$$

will be the index of recall performance.

The precision error can be expressed in terms of the unit step function as follows:

$$p^*(x) = \frac{1}{x}\left[U_{-1}(x-1) + U_{-1}(x-2) + \cdots + U_{-1}(x - n_o)\right]$$

and

$$p(x) = \frac{1}{x}\left[U_{-1}(x - O(1)) + U_{-1}(x - O(2)) + \cdots + U_{-1}(x - O(n_o))\right].$$

Now,

$$\int_{-\infty}^{b} U_{-1}(x-a)\,\frac{dx}{x} = \int_{a}^{b} \frac{dx}{x} \ln b - \ln a.$$

Therefore,

$$\text{precision error} = \int_{1}^{N} (p^*(x) - p(x))dx$$

$$= \sum_{i=1}^{n_o} \int_{1}^{N} \frac{dx}{x}\left[(U_{-1}(x-i) - U_{-1}(x - O(i))\right]$$

$$= \sum_{i=1}^{n_o} \ln O(i) - \ln i$$

$$= \sum_{i=1}^{n_o} \ln O(i) - \sum_{i=1}^{n_o} \ln i ;$$

or,

$$\text{precision error} = \ln \prod_{i=1}^{n_o} O(i) - \ln n_o !$$

Again, by the same consideration, this index may be normalized to lie in the range 0 - 1 by dividing by the maximum precision error. This error must be:

$$\text{max precision error} = \ln \prod_{i=1}^{n_o} N - i + 1 - \ln n_o !$$

$$= \ln \frac{N!}{N - n_o!} - \ln n_o !$$

$$= \ln \left(\frac{N}{n_o} \right) .$$

The normalized index of precision error is, therefore:

$$pe = \frac{\ln \prod_{i=1}^{n_o} O(i) - \ln n_o !}{\ln \left(\frac{N}{n_o} \right)} .$$

Again, since this is an index of error, an index of performance is obtained by considering 1 - pe, i.e.,

$$1 - \frac{\ln \prod_{i=1}^{n_o} O(i) - \ln n_o!}{\ln \left(\frac{N}{n_o}\right)}$$

will be the index of over-all precision performance.

Since both these indices reflect over-all performance, a value of 1 for either implies a value of 1 for the other, in opposition to the conventional recall and relevance ratios. The difference between these two over-all measures lies in the weighting given to the relative position of the relevant documents in the retrieved rank list. The recall index weights rank order uniformly, since it is sensitive to each relevant document. The precision index, however, weights initial ranks more strongly, since it is sensitive to having a high percentage of relevant documents in the initial part of the retrieved list.

The recall and precision indices derived here depend on the assumption that the ordering induced on D by S was a full order, i.e., that it could be represented by a one-to-one mapping from D to the integers from 1 to n(D). As this may not be true in general, since a partial order rather than a full order may result, a method for defining document rank in this case is required.

The most natural way of treating documents which are equivalent under a partial order induced by S is to give each member of the equivalent set the average of the ranks which would apply to the set members if they

were differentiable. Hence, if S induces the partial order $d_1 > d_2 > \{d_3, d_4, d_5\} > d_6$; on a set $D = \{d_1, d_2, d_3, d_4, d_5, d_6\}$ and $D_q = \{d_1, d_5, d_6\}$, the rank assigned to d_1 would be 1 and to d_5 would be 4, and to d_6 would be 6.

In addition, these performance indices may be extended to the case in which there is a partial ordering of the set of relevant documents D_q. In this case, the objective of a retrieval operation would need to be redefined to take account of this partial ordering. Assume that a set of relevant documents D_q for a query q is defined, and that, further, a partial ordering on D_q is specified which reflects degree of relevance, i.e.,

$$D_{q_1} > D_{q_2} > \cdots > D_{q_k},$$

where $D_{q_i} \in D_q$ and $>$ implies "more relevant than." In this case, one may define the objective of a retrieval operation as follows: a retrieval operation with respect to a query q and a partially ordered set of relevant documents D_q is expected to produce an ordering on the reference collection D, such that every member of the set D_{q_i} is ranked higher than $D_{q_{i+1}}$, and that all members of D_q are ranked higher than $\bar{D}_q$.

Corresponding to this definition, expressions for $r^*(x)$, $p^*(x)$, $r_q(x)$, and $p_q(x)$ could be defined in a manner analogous to those defined above. The development of the indices for this case is more cumbersome than the case previously considered and will not be presented here. The only significant difference which arises is due to the fact that a relevant document d_i in subset D_{q_i} may have lower retrieval rank than a document d_j

in subset D_{q_j}, where the partial ordering on D_q is such that $D_{q_j} > D_{q_i}$. This necessitates considering only the positive differences between the retrieval ranks of the relevant documents and the corresponding ideal retrieval ranks. To illustrate, consider a case where

$$D_q = \{d_1, d_2, d_3, d_4, d_5, d_6\}$$

and

$$D_{q_1} = \{d_1, d_2, d_3\} > D_{q_2} = \{d_4, d_5, d_6\};$$

let the retrieval order be $d_1, d_5, d_3, d_4, d_2, d_6, \ldots$. Then

$$\sum_{i=1}^{n(D_{q_1})} O_1(i) - i + \sum_{i=n(D_{q_1})+1}^{i=n(D_{q_1})+n(D_{q_2})} O(i) - i = (1+5+3-1+2+3) + (4+2+6) - (4+5+6) = 3 + (-3) = 0,$$

even though there is clearly a departure from ideal retrieval. By considering only positive differences, the result would be a retrieval error of 3. The same observations apply to the precision index.

3. Experimental Use

These indices have been used to evaluate the results of a variety of experiments conducted with the SMART system. As one might expect from the formulation, the range of the recall index is rather limited; i.e., a random retrieval would produce a recall index of .5, hence one would suspect

observed results to be near 1.0. In fact, the observed range is from about .9 - 1.0 with the average probably close to .97. The precision index, however, has a reasonable range for the requests examined to date, and typically varies from .6 - 1.0. In practice, then, one is forced to expand the scale of the recall index, so that a range 0 - 1 is no longer maintained. For the results obtained to date a scale expansion of 5, introduced so as to maintain an upper value of 1.0, produces a range for the recall index similar to that of the precision index. The scaled recall index has, therefore, been defined as:

$$1.0 - 5(1.0 - x),$$

where x is the normalized recall index.

Two related performance indices may be derived from the two which have been considered. These are useful in the case where a particular query is subjected to a set of retrieval operations which are to be compared.

It may be remembered that the recall error was found as:

$$\text{recall error} = \bar{O} - \frac{n_o + 1}{2}.$$

Since $\bar{O}_{max} = \frac{n_o + 1}{2}$, an index with a maximum value of 1 may be defined as:

$$\text{rank recall} = \frac{\frac{n_o + 1}{2}}{\bar{O}}.$$

A similar observation for the case of the derived precision error produces a precision index:

$$\log \text{precision} = \frac{\log n_o !}{\log \prod_{i=1}^{n_o} O(i)}.$$

The advantage of these indices lies in the fact that they are simpler and, therefore, easier to compute than the normalized indices, and that the rank recall takes on a wider range for the results which have been observed. The disadvantages of both these measures is their dependence on n_o, the number of relevant documents for the query in question. This dependence makes it impossible to average these indices over a set of queries and thus their usefulness is limited.

REFERENCES

1. Cleverdon, C. W., "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," ASLIB-Cranfield Research Report, Cranfield (October 1962).