3

Retrieval effectiveness

Cornelis J. van Rijsbergen

3.1 Introduction

Information storage and retrieval systems have been with us for many years now. Attempts to evaluate or measure their performance have been going on almost as long. This is not an entirely unrelated development since in designing and building any new system the question of its desirability, quality, value and benefit should arise naturally. In evaluating information storage and retrieval systems, those that deal with the retrieval of references to documents, much of the effort has gone into measuring variables based on the **relevance** of documents to the question put to the system. This aspect of evaluation is clearly only one part of the overall evaluation of any retrieval system. These relevance-based variables are chosen to reflect in some way what has now become known as the **retrieval effectiveness**: the ability of the system to retrieve relevant documents while at the same time suppressing the retrieval of non-relevant documents. The most well known pair of variables jointly measuring retrieval effectiveness are precision and recall, precision being the proportion of the retrieved documents that are relevant, and recall being the proportion of the relevant documents that have been retrieved. Singly, each variable (or parameter as it is sometimes called) measures some aspect of retrieval effectiveness; jointly they measure retrieval effectiveness completely.

The measurement of precision and recall, or of any other similar pair of variables, is different in many respects from the measurement of variables in, say, the physical sciences. Each variable is based on the availability of data about the relevance of particular documents to a query. Although one can make a case for an objective notion of relevance, many researchers believe that relevance is entirely subjective, that is, given the same query but put by different users, different documents will be judged relevant. In this respect relevance behaves more in the way an observable behaves in quantum physics, since its measured value is not determined except in probability. The distribution of values associated with an observable will follow a certain probabilistic law determined by the state of the system. Unfortunately in information retrieval a similar probabilistic law for relevance does not exist. Hypotheses about the user population could be formulated to establish such a law, but its usefulness would be doubtful.

This analogy between relevance and an observable in quantum mechanics (QM) should not be taken too seriously, its use is mainly to highlight the inherent uncertainty associated with relevance. One interesting aspect of the analogy is, however, that a further similarity between the formalism for information retrieval and QM becomes apparent when one considers the well established trade-off between precision and recall. This is similar to the Heisenberg uncertainty principle in physics, where for example momentum and position cannot be measured simultaneously to any desired level of accuracy: increasing the accuracy for one leads to a necessary decrease in accuracy for the other. Similarly in attempting to increase precision we always find a decrease in recall. In fact under some mathematical models in information retrieval the trade-off is a necessary one, and not simply observed empirically.

It must be emphasized that measuring retrieval effectiveness is a form of derived, as opposed to fundamental, measurement. The fundamental 'quantity' involved is relevance; once this has been established we can attempt to measure retrieval effectiveness. Because of the difficulties involved in establishing a theory of relevance, relevance-based measures have typically been used in an experimental (artificial as opposed to operational) context, that is, in one where the relevance of a document has been decided in advance. Given such an experimental set-up, it would appear that retrieval strategies can be evaluated for their effectiveness in terms of, say, precision and recall without any difficulty. Unfortunately life is not that simple, and difficulties arise with the form of measurement at different levels. A few are as follows:

(1) Sampling level?
(2) One or more variables?
(3) How to normalize for relevance feedback information?
(4) Effect of ranking on form of measurement?
(5) Effect of interpolation, interpretation?
(6) Effect of averaging technique?

Each of these technical problems, except the first which has already been covered by Robertson, will be touched upon in the rest of this chapter.

3.2 Theoretical foundations

The problems of measurement in information retrieval differ from those encountered in the physical sciences in one important aspect. In the physical sciences there is usually an empirical ordering of the quantities we wish to measure. For example, we can establish empirically by means of a scale which masses are equal, and which are greater or less than others. Such a situation does not hold in information retrieval. There is no empirical ordering for retrieval effectiveness and therefore any measure of retrieval effectiveness will by necessity be artificial.

The basic variables underlying any measure of retrieval effectiveness are usually precision and recall, or some other equivalent pair. The conventional way to define these is in terms of ratios; however, more recently it has proved fruitful to define them as probabilities. Recall is defined as the probability

that any given document (for the population of potential documents) that is input to the system and relevant to question Q will be retrieved in response to Q. Similarly precision is defined as the probability that a retrieved document will be relevant. Notice that these definitions refer to individual documents, so for these to make sense in terms of ratios I need to restate them in terms of sets. If A is the set of relevant documents and B is the set of retrieved documents then recall is $P(B/A)$ (read, probability of B given A) and precision is $P(A/B)$, these can then be estimated simply by the traditional ratios. Now let us look at the meaning of the probabilities more closely. Take the probability of retrieval given relevance $P(B/A)$, or more precisely the probability of a document belonging to the retrieved set, given that it belongs to the set of relevant documents. If A and B are two simple primitive predicates, this conditional probability is easy to interpret, however B is not simple. If we are to estimate the probability $P(B/A)$ then we can only do this reliably through a *random* process, but any retrieval strategy is not a random process and thus technically cannot be used to estimate $P(B/A)$.

One way of dealing with the definition of $P(B/A)$ is to describe set B in terms of a random variable. One assumes that attributes of documents can be measured or ordered in some way in relation to a query, and that the values of these attributes can be mapped into a single variable (like co-ordination level or cosine correlation) having a well defined distribution. If one further assumes that the distributions of this variable on the relevant and non-relevant sets of documents are different, then one particular value of this variable (the cut-off) can be used to discriminate between relevant and non-relevant documents. Now the probabilities of a document belonging to B conditional on A is well defined, it is the probability that, say, the matching function for relevant documents exceeds some threshold. It is not difficult to see how $P(B/\bar{A})$, fallout, can be defined in a similar manner.

Both $P(B/A)$ and $P(B/\bar{A})$ can be defined as the expected value of some other variable. When defined in this way we refer to them as **expected recall** and **expected fallout**. For this we assume that each document has associated with it a *unique* description $\boldsymbol{x}$. The uniqueness is not necessary but it simplifies the discussion. Robertson[1] in his paper on the probability ranking principle defined them as follows:

$$P(B/A) = \sum_{x \in B} P(\boldsymbol{x}/A)$$

$$P(B/\bar{A}) = \sum_{x \in B} P(\boldsymbol{x}/\bar{A})$$

In words, if one can associate a probability with each document description given that the underlying document is relevant (or non-relevant), then expected recall (or expected fallout) is simply defined as the sum of the probabilities over the retrieved set. If these probabilities were continuous, then the sum would represent the area under the curve $P(\boldsymbol{x}/A)$ or $P(\boldsymbol{x}/\bar{A})$ supported by the set B. This is of course how Swets[2] originally defined these probabilities.

A slightly different and perhaps more instructive way of defining *expected recall* and *expected precision* (rather than fallout, although this can be defined too) is in terms of the expected number of relevant documents in a set. For this we need to define $P(A/\boldsymbol{x})$, or, in words, the probability that a document

is relevant given a particular description. This probability can be derived from $P(x/A)$ through Bayes' Theorem (see van Rijsbergen[3], p. 115). The expected number of relevant documents can now be simply defined as

$$\sum_{x \in B} P(A/x)$$

From this we get:

$$\text{Expected Recall} \quad \sum_{x \in B} \frac{P(A/x)}{|A|}$$

$$\text{Expected Precision} \quad \sum_{x \in B} \frac{P(A/x)}{|B|}$$

Defining $P(B/A)$ and $P(A/B)$ in terms of expected recall and expected precision has many advantages, for one, it does not raise the same problems of interpreting the probabilities. Another major advantage is that the trade-off between expected recall and precision can be shown to hold almost immediately for certain forms of retrieval[4].

We return now to the problem of constructing a measure of retrieval effectiveness. In some ways this is a secondary problem, particularly if one admits that any such measure will be a function of at least two variables such as precision and recall. Nevertheless it may well be possible to construct a sensible measure of retrieval effectiveness independent of the traditional parameters, but this would only be worth doing if it simultaneously led to a different theory of information retrieval. One of the present major advantages of measuring retrieval effectiveness in terms of recall and precision (or fallout) is that we are able to state categorically that if retrieval is done in a certain way it will be optimal in terms of its effectiveness measured by recall and precision. It is conceivable that optimality in terms of precision and recall does not result in optimality with respect to some measure of effectiveness, although to date most sensible composite measures are optimized as well.

3.3 Optimal retrieval

One of the more interesting things that has happened in information retrieval research in recent years is that theoretical work on evaluation and on retrieval strategies have fitted together. Of course much earlier Swets tried to do this and his work did cause a flurry of papers, but their impact on further theoretical and experimental work was not felt until much later.

Probably the single most important result in which the definition of retrieval effectiveness and retrieval strategy interact is the **probability ranking principle**. This principle emerged in the work of Robertson and Sparck Jones[5], and Cooper[6]. It reads as follows:

> 'If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the request as submitted by the user, where the probabilities are estimated as accurately as possible on the basis of content derivable

> data made available to the systems for this purpose, then the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.'

The original formulation of this principle was in terms of the probability of 'usefulness' instead of 'relevance'. To date most implementations of the principle have worked with the probability of relevance. It is not too difficult to formulate a relationship between usefulness and relevance so that optimality in terms of relevance implies optimality in terms of usefulness. The crucial point is though that the estimation of the probability of relevance is based on content derivable data whereas an estimate of probability of usefulness can only be based on data not concerned with the content, e.g. language of document, date of publication. A detailed discussion of this distinction can be found in Harper's thesis[4].

As I mentioned at the start of this section, retrieval effectiveness and retrieval strategy have been fitted together to guarantee optimal retrieval effectiveness for certain strategies. The easiest way to see this is to use the definition of precision and recall in terms of expected number of relevant documents in a set. If the probability of relevance of a document represented by x is given by $P(\text{relevance}/x)$ or $P(A/x)$ then the probability ranking principle tells us that to achieve optimal retrieval we should rank the documents in decreasing order of this probability. Now the retrieved set B, defined by setting some cut-off on the ranking, will contain those documents with the greatest values of $P(A/x)$. Therefore compared with any other set of documents of the same size as B, the sum,

$$\sum_{x \in B} P(A/x),$$

will be a maximum, or in words, the expected number of relevant documents in B will maximized. This is true for any set B defined by a cut-off on the ranking. Since expected precision and recall are defined by dividing the expected number of relevant documents in B by the size of B and A respectively, expected precision and recall will be maximized at any cut-off by ranking the documents in order of their probability of relevance. The interplay of the measures of retrieval effectiveness and the definition of the retrieval strategy is quite clear. In fact ranking documents in this way ensures the optimization of a host of effectiveness measures expressed in terms of precision and recall. For example, any linear combination of precision and recall will be maximized as well.

It is important to realize that in formulating this principle very little has been said about the structure of the description x associated with a document. To estimate the probability of relevance for a particular document some assumptions will have to be made about the form of x. A common assumption is that x is a binary vector representing the absence or presence of index terms. Also, assumptions about the statistical dependence or independence of the occurrence of index terms can then be made to help in the estimation of $P(A/x)$. Briefly, this estimation is usually implemented through Bayes' rule,

$$P(A/x) = \frac{P(x/A)P(A)}{P(x)}$$

where we seek a sample of relevant documents to be able to estimate $P(\boldsymbol{x}/A)$; $P(A)$ is the same for each $\boldsymbol{x}$, and $P(\boldsymbol{x})$ is given by the system data.

It is possible to formulate an approach to optimal retrieval without using ranking. For this we need some elementary decision theory. The basis of it is that certain costs are associated with the decision that the retrieval system can make. If we assume that for each document the system can take one of two actions; a_1: retrieve, a_2: not retrieve, and that each document is in one of two states, either w_1: relevant, or w_2: non-relevant. Then we can associate with each (action, state) pair a cost l_{ij}. A table shows the association:

	w_1	w_2
a_1	l_{11}	l_{12}
a_2	l_{21}	l_{22}

With each action a we can associate an expected cost, viz.:

$$R(a_i/\boldsymbol{x}) = \sum_{j=1}^{2} l_{ij} P(w_j/\boldsymbol{x})$$

where $P(w_j/\boldsymbol{x})$ is the probability that the document will be in state w_j. Intuitively it would seem reasonable to perform that action which has the smallest expected cost associated with it. In fact, such a strategy is optimal in the following sense. If the decision rule is $a(\boldsymbol{x})$, i.e. $a(\boldsymbol{x})$ takes the value a_1 or a_2 for each $\boldsymbol{x}$, then the **overall risk** R is defined as

$$R = \sum_{\boldsymbol{x}} R(a(\boldsymbol{x})/\boldsymbol{x}) P(\boldsymbol{x})$$

This function R can be minimized for each $\boldsymbol{x}$ by choosing the smaller of $R(a_1/\boldsymbol{x})$, $R(a_2/\boldsymbol{x})$. Therefore the retrieval rule will read:

If $R(a_1/\boldsymbol{x}) < R(a_2/\boldsymbol{x})$ *then* retrieve
else do not retrieve*

So optimality here means minimizing the overall risk function.

It is interesting to analyse this retrieval rule in a little more detail. Writing out the expected cost functions in full we get

$$R(a_1/\boldsymbol{x}) = l_{11} P(w_1/\boldsymbol{x}) + l_{12} P(w_2/\boldsymbol{x})$$
$$R(a_2/\boldsymbol{x}) = l_{21} P(w_1/\boldsymbol{x}) + l_{22} P(w_2/\boldsymbol{x})$$

If we define a reasonable cost function we would set $l_{11} = l_{22} = 0$, thus reducing the comparison,

$$R(a_1/\boldsymbol{x}) < R(a_2/\boldsymbol{x})$$

to

$$l_{12} P(w_2/\boldsymbol{x}) < l_{21} P(w_1/\boldsymbol{x})$$

or equivalently

$$\frac{P(w_1/\boldsymbol{x})}{P(w_2/\boldsymbol{x})} > \frac{l_{12}}{l_{21}} \; (=\alpha)$$

* As is usual, equality is treated by deciding randomly.

Since $P(w_2/x) = 1 - P(w_1/x)$, we can rewrite this inequality as

$$P(w_1/x) > \frac{\alpha}{1+\alpha} (= \beta)$$

This rule is similar to the one specified by the probability ranking principle, the difference being that now we have an explicitly defined cut-off in terms of a cost function. If we choose $l_{12} = l_{21}$, i.e. $\alpha = 1$ our retrieval rule becomes:

If $P(w_1/x) > \frac{1}{2}$ *then* retrieve
else do not retrieve

Or in other words, if the probability of relevance is greater than the probability of non-relevance for a document we should retrieve that document. We vary the importance we attach to the cost of a **false drop** in comparison with the cost of a **recall failure** by changing the cut-off. For example, we may decide that

$$\frac{l_{12}}{l_{21}} = \frac{1}{2},$$

which means that the failure to retrieve a relevant document is twice as costly as the retrieval of a non-relevant document. For this the cut-off β is set to $\frac{1}{3}$. By ranking we avoid having to specify the cut-off in advance, on the other hand we pay the price of ranking. It is important to realize that setting a cut-off on $P(w_1/x)$ still maximizes the expected number of relevant documents in the retrieved set.

3.4 Measurement of effectiveness

The measurement of retrieval effectiveness within an experimental set-up is beset with many difficulties. These difficulties have been with us for many years, and are likely to remain unresolved for many years yet. A typical retrieval experiment has been described in Chapter 2, so I shall not repeat it here, except to emphasize that its aim is usually to establish the absolute or relative effectiveness of some search strategy, information structure, ordering process, etc., within the context of an overall retrieval system. The output for such an experiment may be a ranking (partial or full) of documents, or simply an unordered set. Each query will have associated with it some output for which retrieval effectiveness measures can be calculated. In comparing the results for different tests with the same queries and document collection, one aims to produce statistical summary data which will enable statements to be made about the comparative merits of differently designed subsystems. In the main, experimentalists have concentrated on two types of statements.

(1) What is the probability that a retrieved document is relevant for the *system* operating at that level of recall?
(2) What is the probability of a retrieved document being relevant for a *query* at a particular recall value?

How well these two questions are answered depends on the method of evaluation adopted. The discussion will concentrate on evaluation of rankings, evaluating unordered sets is a special case of this.

In the main there are two distinct ways of looking at the problem of measuring effectiveness. One way assumes that the effectiveness of a given system for a set of queries is a direct function of the effectiveness for individual queries without reference to how these might have been arrived at; let us call this the **predictive** approach. Another way is to insist that to represent the effectiveness for a set of queries one should set up a correspondence between levels of effectiveness of different queries based on some control variable (e.g. match value) used to generate the ranking of documents in the first place; let us call this the **descriptive** approach. A minor variation of this latter approach is to use the rank order number of the document as the control variable, the actual value is then ignored.

Perhaps if the above is illustrated by describing what happens in terms of precision and recall it will be clearer. I shall limit the example to these two parameters as any other two parameters would be treated analogously. The first approach attempts to summarize, by simple averaging, precision values at given recall values. The second approach averages both precision and recall at a given value of a control parameter, e.g. co-ordination level. Both methods have problems, the first requires the precision value to be defined at recall values not necessarily achieved at any control variable value. The second method requires a decision about which value of the control variable for one query will correspond to what value of the control variable for another query, so that averaging may be done across queries for precision–recall values at corresponding values of the control variable. This still leaves open the question, for either method, how might these averages be computed? The predictive approach requires interpolation and extrapolation of precision values so that averages can be computed at given recall values. On the other hand in the descriptive approach one need not calculate precision–recall values for individual queries at any given value of the control variable; instead one pools the documents and calculates what are known as micro-averages. In other words for all queries one pools the documents retrieved and the relevant documents retrieved and then calculates an average recall and precision. Once the averages have been calculated it would appear that the descriptive approach answers question (1), whereas the predictive approach answers question (2). Of course once the averages have been calculated we still only have a *set* of average precision recall values; a final step is to link these points into a continuous curve.

There are arguments for both approaches. In an earlier publication I have strongly argued in favour of the predictive approach[7]. Sparck Jones[8] has argued in favour of the descriptive approach.

Although the above discussion has assumed that retrieval output is subject to a control variable leading to a sequence of nested sets of retrieved documents, some strategies will only retrieve one set of documents. For example, output from a boolean search, or from a cluster-based retrieval strategy, will be just one unordered set of documents. Problems arise when attempting to compare 'set retrieval' with 'ranked retrieval', which requires some statement about the comparative performance of two retrieval strategies, one for which the effectiveness is represented by a graph, the other by a point. It is for cases like this that a single number effectiveness measure, call it E, can be useful. If one assumes that for every point of the graph a single number measure can be calculated, then one way of comparing

effectiveness is to select a point in terms of E, this might be the best, worst or some other point, and then compare it with E for the set output. Once a single number measure has been adopted statistical summaries for sets of queries become straightforward, and interpolation and extrapolation are not needed. Of course in conflating the precision–recall (P–R) graph to one value there is loss of information, but this is not as severe as it may appear at first sight. There is a certain amount of evidence now that no matter what model is adopted for retrieval, the precision–recall graphs are constrained to some extent. In fact it is not difficult to prove that under the probability ranking principle expected recall and expected precision are inversely related. This means that given any E value for a point on the P–R graph, E values for other points are constrained by this trade-off. This is not to say that there is a functional relationship between P and R but that there is 'almost' one. In this sense the loss of information is not so severe, although it must be admitted that this loss has not been quantified.

Further difficulties arise in evaluation when attempting to measure the comparative effectiveness of relevance feedback strategies. In these strategies certain documents are looked at on a first iteration to establish the parameters for the second iteration. Typically the documents looked at are the top documents in a ranking, the remaining documents are unsighted. To establish the effectiveness of the feedback, we must somehow measure how feedback improves retrieval. The most sensible way of doing this is to generate a *residual ranking* for the second iteration which is a ranking with the n feedback documents removed. This can then be compared with the ranking for the first iteration with the same n documents removed, in this latter case they are of course the top n documents. From these rankings, one for the first and one for the second iteration, precision–recall graphs can be generated. This method has been used extensively by Harper[4] and Ide[9] for evaluating feedback experiments. It neatly measures the effect of feedback on documents the user has not previously seen.

3.5 Limits to retrieval

In evaluating the results of retrieval experiments, it is often important to establish the bounds on retrieval. Trivial bounds obviously exist in that retrieval effectiveness cannot exceed precision and recall jointly being 100 per cent, nor can it fall below both being 0 per cent.

One interesting speculative question to ask is whether in fact we wish to design retrieval systems that achieve 100 per cent precision and recall. It is not too difficult to argue that this could be achieved for some specific query. Achieving 100 per cent precision and recall on the average for some unknown set of queries is a different matter. In designing any retrieval system we use certain models for the structures and processes involved. These models are necessarily an imperfect reflection of the reality they are trying to model. In particular any model for relevance we might invoke will have built in an inherent uncertainty. Therefore one would hypothesize that perfect retrieval is impossible, or to put it differently, that a retrieval system cannot be all things to all men. Let us now look at a possible objection to the above claim of the impossibility of perfect retrieval. One might claim it is the primitive

manner of representation that leads to the limitation on effectiveness. In one sense this is true. However, consider the situation where we could characterize and discriminate our documents perfectly, and let us assume that the system is very large and that queries can be arbitrarily complicated. Of course I do not believe that for an almost infinite collection we can represent documents perfectly with anything less than the complete semantic content (disregarding arguments as to how this might be done). To achieve perfect retrieval in such a case, one would need to know already what one was looking for, and hence the retrieval system would be irrelevant. The point of a retrieval system is to tell you what you do not already know in response to an incomplete statement of your information required.

The computational process of locating likely relevant documents is not unlike the mechanical process of proving theorems within a formal system. If we assume, not unreasonably, that a computational model for retrieval is in power equivalent to arithmetic, then Gödel's theorem by analogy also leads one to suspect that certain statements will be undecidable; or to put it more precisely, there will be documents whose relevance or non-relevance is 'undecidable' within the formal system for retrieval, i.e. documents which the system cannot guarantee to find. What I am arguing is that there may be an essential incompleteness about retrieval systems, which although not apparent in current implementations, may well become apparent with increased sophistication of our systems.

At the other end of the scale we have performance trivially limited by 0 per cent precision and 0 per cent recall. A less superficial bound is the one set by random retrieval, which is represented most naturally by recall = fallout. Any retrieval strategy worth its salt will do better than random.

The limit set by random retrieval can be discussed at two levels: a **global** and a **local** level. At the global level it simply indicates that we can expect to do better overall by some retrieval strategy than retrieving randomly from the collection. At a local level a limit can be used in two ways. First, to set a natural cut-off when ranking with respect to the probability of relevance; at some point down the ranking the estimated $P(\text{relevance}/\boldsymbol{x})$ will equal the prior probability $P(\text{relevance})$ which is the point at which we are retrieving randomly, and beyond which point the strategy becomes useless. Of course this 'random' cut-off point will almost always exceed any cut-off set by the user. But since in most experiments we are dealing with simulated users it is as well to rank to this conservative cut-off point. Secondly, we can use random retrieval at a local level to motivate a particular way of interpolating between precision–recall points. It was pointed out earlier that in the predictive method of evaluation we needed to interpolate. One possible method is by step-function, the jumps occurring at actual changes in recall. One way of motivating, and to some extent justifying, this method is to argue in reverse, and say that given any recall–precision point which has been calculated for the retrieved set then it is always possible to retrieve randomly from this retrieved set. Therefore it is always possible in this way to generate points at the same precision value increasing in recall up to the recall of the given point. Thereby we can interpolate points between any two given points by defining firstly:

$$G = \{(R_\theta, P_\theta)\}$$

as the set of points at which there is a change of recall, and secondly an interpolation function

$$P(R) = \{\sup P : R' \geqslant R \text{ such that } (R', P) \in G\}$$

This is simply an algebraic expression for the interpolating step-function with the jumps occurring at (R_θ, P_θ). One consequence of defining the function in this way is that certain points belonging to G may be 'ignored'. To see this consider two neighbouring points (R_1, P_1) and (R_2, P_2) such that $R_2 > R_1$ but $P_2 < P_1$ (normally one would expect $P_2 < P_1$ because of the trade-off). Then in interpolating for a value R_0 immediately preceding R_1 the P_1 value will not be used since $P(R_0) = P_2$. In an earlier publication I referred to (R_1, P_1) as an **anomaly** emphasizing that it could legitimately be ignored[7].

Other limits to retrieval effectiveness are due to the nature of the mathematical model used to represent any subsystem. Any model is always only an approximation to reality, and will therefore have inherent limitations associated with it. The most important set of models for which limits have been calculated are the probabilistic ones. These models require the estimation of certain probabilities from sample information. If one assumes that these probabilities can be calculated perfectly then in some sense the model cannot be improved. Therefore using perfect or complete information for the probabilities should lead to the best possible retrieval under the model.

One must be very careful in thinking about the limitations imposed by certain models. For probabilistic models there are at least two levels of approximation. At one level the model is approximating the particular process or structure identified as determining retrieval effectiveness, at another level one is estimating (or approximating) the parameters of the particular model. A good example of this can be seen when constructing the function $P(\boldsymbol{x}/C)$ where C might be relevant or non-relevant. As a first step one decides on the structure of $\boldsymbol{x}$ and its associated distributions. This could be to assume that the components of $\boldsymbol{x}$ are independent, or partially or fully dependent. This implies a series of models one of which may be a better approximation than any of the others. Once the particular model has been settled one attempts to estimate its parameters. Now it is impossible to establish which is the *correct* model; one can only demonstrate by experiment that one model will lead to better performance than some other model, and assume that because our experiment is random any future experiment will show the same result.

Although in the above discussion I have blithely talked about perfect or complete information for estimating parameters, in reality we never have this. Our information fails to be perfect in at least two important ways. First, our documents are only a sample of a population of potential documents, and so even though we use all the information in the collection available to calculate our so-called perfect estimates, they are not the population parameters. Secondly, in most collections for which relevance judgements have been made, they are rarely 100 per cent exhaustive although they may be 80 per cent exhaustive, and so our perfect information falls short by 20 per cent.

Upper bounds to retrieval effectiveness can be used in different ways, (1) as a general experimental yardstick for performance, (2) to compare the

potential performance of different models, and (3) to study the effects of small changes to any given model. Of course comparing upper bounds can be misleading for the following reasons. It may well be that upper bound A exceeds upper bound B significantly; however, it may not only be that it is not possible to design a reasonable estimation rule for A: it could be that the model for A is so complicated that we can not get sufficient data to estimate its parameters.

3.6 Conclusions

In this chapter I have tried to show the considerable interplay that exists between probabilistic definitions of effectiveness and certain models for retrieval. In the past researchers, including myself, have tried to argue the relative merits of different measures of effectiveness independent of how these might influence the design of retrieval systems. I now think that the most important consequence of defining a measure of retrieval effectiveness in a particular way is the ability it gives us to make theoretical statements about certain models. These statements can then be tested empirically against stimulated or real users.

References

1. ROBERTSON, S. E. The probability ranking principle in IR, *Journal of Documentation* **33**, 294–304 (1977)
2. SWETS, J. A. Information retrieval systems, *Science, N.Y.* **141**, 245–250 (1963)
3. VAN RIJSBERGEN, C. J. *Information Retrieval*, 2nd edn, Butterworths, London (1979)
4. HARPER, D. J. *Relevance Feedback in Document Retrieval Systems*, Ph.D. Thesis, Computer Laboratory, University of Cambridge (1980)
5. ROBERTSON, S. E. and SPARCK JONES, K. Relevance weighting of search terms, *Journal of the American Society for Information Science* **27**, 129–146 (1976)
6. COOPER, W. S. The suboptimality of retrieval ranking based on the probability of usefulness (private communication to S. E. Robertson) (1977)
7. VAN RIJSBERGEN, C. J. Retrieval Effectiveness. In: *Progress in Communication Sciences*, (Ed. M. J. Voigt and G. J. Hanneman), Vol. 1, pp. 91–118, Ablex Publishing Corporation, New York (1979)
8. SPARCK JONES, K. Performance averaging for recall and precision, *Journal of Informatics* **2**, 95–105 (1978)
9. IDE, E. *Relevance Feedback in an Automatic Document Retrieval System*, Master's Thesis, Report ISR-15, Department of Computer Science, Cornell University (1969)