2

The methodology of information retrieval experiment

Stephen E. Robertson

2.1 Introduction

Information retrieval systems have been the subject of experimental testing for some twenty years now. Like any field in this position, a fair amount of know-how has accumulated about the proper conduct of such investigations. The object of this book is to distil this know-how; the object of this chapter is to set the scene. Thus I will be introducing the basic ideas, sketching in some of the main problem areas, and generally preparing the reader for the more specific or concrete chapters that follow.

Van Rijsbergen takes up the question of the evaluation of retrieval effectiveness in Chapter 3; in Chapter 4, Belkin considers information retrieval in a wider context; and in Chapter 5, Tague gets down to the detail of conducting experiments.

Definitions

What, then, do we intend to convey by this general rubric, the 'experimental testing of information retrieval systems'?

Information retrieval is generally taken to mean the retrieval of references to documents in response to requests for information (more about documents and requests below). An information retrieval **system** is a set of rules and procedures, as operated by humans and/or machines, for doing some or all of the following operations:

Indexing (or constructing representations of documents);
Search formulation (or constructing representations of information needs);
Searching (or matching representations of documents against representations of needs);
Feedback (or repeating any or all of the above processes, with modifications introduced in response to an assessment of the results of some process);
Index language construction (or the generation of rules of representation).

Document is (in theory, at least) taken as more-or-less synonymous with **text** in linguistics—that is, it describes any piece of linguistic (in the widest sense) material that can reasonably be considered as a unit. (In practice, the vast majority of experiments have used the scientific paper as the normal unit.)

Request (or **query**) has usually been taken to mean the statement by the requester describing his/her information need, but recently (particularly with the development of systems such as on-line which allow immediate feedback) has come to mean simply the act of requesting. It is usually assumed that this act is stimulated by an underlying need for information, which in some sense remains invariant, though the requester's perception and/or description of it may change in the course of his/her interaction with the system.

User and **requester** are synonymous. The notion of **testing** has already been discussed, as has the distinction between **experiment** and **investigation**, in the editor's introduction.

A distinction is usually made between systems for current awareness or the selective dissemination of information **(SDI)** and those for **retrospective retrieval**. In terms of the mechanics of the system, in retrospective retrieval a request is made to a system as a one-off occurrence, and searched against the current collection of documents; in SDI, repeated searches are made against successive additions to the document collection, over a period of time.

What is the purpose or function of an information retrieval system: what is it supposed to do? The simple answer to this question is to retrieve documents in response to requests; but this is too simple, any arbitrary gadget could do that. The documents must be of a particular kind: that is, they must serve the user's purpose. Since (we assume) the user's purpose is to satisfy an information need, we might describe the function of an information retrieval system as 'leading the user to those documents that will best enable him/her to satisfy his/her need for information'.

There are many different aspects or properties of a system that one might want to measure or observe, but most of them are concerned with the **effectiveness** of the system, or its **benefits**, or its **efficiency**. Effectiveness is how well the system does what it is supposed to do; its benefits are the gains deriving from what the system does in some wider context; its efficiency is how cheaply it does what it does. In this book, we are mainly, but not exclusively, concerned with the effectiveness or (synonymously) the performance of information retrieval systems.

Why test information retrieval systems?

This book is mainly about the 'how' of testing. But before we launch into the technicalities of how best to conduct a test, we should (at least briefly) consider the prior question of why.

Starting from the simplest situation, suppose that we have a specified clientele and document collection, and two existing information retrieval systems working in parallel, and we wish to decide which of the two to drop. Then we could imagine conducting a formal experiment to help us make this particular decision. In principle, such testing would be relatively straightforward: with a well-defined, specific question to answer, we would have the ideal experimental situation.

The problems become more complex if, instead of two alternative systems, we have one system which we think might be capable of improvement. In this situation, we might for instance want to evaluate how well it performs against some standard, or in terms of some criteria of success or failure, and then look for possible ways of approaching the standard or reducing the failures. Such a test would be more on the lines of an *investigation* as defined in the editor's introduction.

Moving into the sphere of basic research, we may be concerned with the general principles of information retrieval system design. Again we have the distinction between experiments designed to help choose between alternative general principles, and investigations aimed at the discovery of new principles. Another objective for investigation might be to test the feasibility of some particular design principle: that is, to test whether a system can be designed on the basis of such a principle.

All these are good reasons for wanting to test information retrieval systems. They are also all different, and impose different requirements and constraints on the conduct of the test. So even before we get down to the pragmatic level of how best to do things in different circumstances, we find that the question 'How should we test an information retrieval system?' has many answers. Readers are invited to bear this in mind for the rest of the book!

The archetypal retrieval test

At this point, it is worth describing the general form of a retrieval test, as it has evolved over the last 20 years. This is not to say that this form is correct, or that any of the many variants of it are peculiar in any respect; it is merely to establish a reference point on which to base further discussion.

What constitutes a test of a retrieval system or systems? First of all we have to have the system itself: that is, the set of rules and procedures, and human or mechanical operators of these rules and procedures.

Next, we must have the raw material on which the system works: the documents and requests.

Tests in general, and experiments (in the sense defined) in particular, are normally intended to answer specific questions. An important component of any test is the experimental design: that is, the way in which the test is organized in order to answer the appropriate questions.

All tests involve some kind of measurement, in the widest sense of the word. In most information retrieval system tests, this includes (among other things) some form of assessment of the system's response to each query.

Generally, a document collection will contain documents that might have been useful in the context of a particular query, but which the system does not retrieve. Many experiments include attempts to discover some or all of these documents, with a view to assessing the performance of the system against some standard.

Finally, we must have methods of analysing the results, in such a way as to allow us to draw the appropriate conclusions, to answer the questions with which we set out.

All these aspects are discussed further below.

Operational versus laboratory tests

In order to answer a specific question or questions unambiguously, a test must be designed as far as possible to exclude any extraneous variations which may confuse the results—hence the idea of conducting experiments under laboratory conditions, with all variables controlled as far as possible. On the other hand, in order to answer questions which relate directly to real problems in the design of retrieval systems, and to provide answers which will apply in real situations, a test must be conducted in (as nearly as possible) an operational environment.

The conflict between these two aims is a real and continuing one. As a result, a whole spectrum of testing methods has been developed, ranging from pure laboratory experiments to the study of real systems and users in their operating environment. This spread of methods is reflected in the various comments that follow on components of the 'normal' retrieval test.

2.2 Components of the archetype

The system

That we need an information retrieval system in order to do a retrieval test is not quite as obvious as might at first appear. If we go back to the question(s) that gave rise to the test in the first place, they will very often revolve around some particular component of the system: the index language, say, or the indexing process, or the process of search formulation. Suppose, then, that we are concerned with the indexing process. Do we need to even *have* a searching process in order to do an experiment? Is there no way we could test the alternative indexing processes directly, without doing any searching?

The short answer is: no, there is no satisfactory way. By and large it is not possible (at present) to set up criteria for the indexing process which we can be confident will relate to the overall performance of the information retrieval system in the right way. From which it follows that, if we want to decide between alternative indexing strategies for example, we must use these strategies *as part of a complete information retrieval system*, and examine its overall performance (with each of the alternatives) directly.

This is a severe constraint on any test: it is as if, lacking the necessary theories of mechanics and the strength of materials, bridge designers had to build numbers of test bridges to test each material and each structural component which they might use in their designs. (The recent problems with box-girder bridges suggest that the idea is not so far-fetched!)

The second, related problem area under this heading concerns the definition of the boundaries of the system, particularly in connection with the user. The usual view of the information retrieval system manager is that the (narrow) system is that which is under his control; the user normally falls outside this narrow definition. But there are two strong reasons for including at least some of the cognitive processes that the user goes through in the definition of the system for the purpose of the test:

(1) Some of the processes that users go through can be influenced by narrow-system components such as the index language;
(2) In terms of the arguments immediately preceding, we do not know how to relate narrow-system behaviour to wide-system performance.

Both points relate to the idea discussed above, that the function of an information retrieval system is to help the user to satisfy his/her information need, rather than simply to present documents in response to a formal request.

The documents and requests

Almost all information retrieval system tests use genuine documents.

Are there alternatives? It is possible to do some simulation experiments using pseudo-documents which are generated in some fashion (perhaps involving Monte Carlo techniques) so as to imitate real ones in some specific sense (they may, for example, imitate only the index sets associated with documents, rather than the documents themselves). There is certainly a role for such experiments, to answer specific research questions; but for most purposes it is both better and easier to use the real thing. There is certainly no shortage of documents around!

The only remaining question, then, is: Which documents? Here we run into the vexed question of sampling, about which more below.

With regard to requests, the situation is considerably more problematic. To be sure, there is (in principle, at least) no shortage of requests; the problems with obtaining such real requests are several. First, we have to catch them! Requests (in the sense of *acts* of requesting information) exist only for a short space of time, and have to be trapped at that time. Second, the actual locations of these request-acts are usually dispersed; a mechanism for trapping them that is located at one place may take a long time to trap a reasonable number or range of requests. Third, most test designs require, to a greater or lesser extent, the co-operation of the requester. This co-operation may be needed in connection with the operation of the system itself; it may also be needed for the measurement of system output (as discussed below). Fourth, all the three previous difficulties combine to exacerbate the problem of obtaining a sample of requests that is representative of anything.

These problems have prompted many testers to construct artificial queries. Such artificial queries may vary in their degree of realism. Some examples are discussed below.

Experimental design

Broadly, experimental design is concerned with arranging matters so that the experiment does answer the question(s) it is intended to answer. Obviously all the other components discussed above also come under this broad heading. But the phrase is used in a somewhat narrower sense, referring to those aspects of the design that determine whether, from a logical or statistical point of view, appropriate inferences can be drawn. The simplest question one can ask in this context is: How large a sample of documents (say) do I need for this experiment? A more detailed example would be: How should I use the different (human) searchers available, with the different requests and alternative systems, so as to separate the effects I want to measure (the difference between the systems) from those that I don't (any differences between searchers or between requests, or interactions between searchers and/or requests and/or systems)?

Even within this narrower scope of experimental design, it has many aspects, and it is a little difficult to generalize about the design of retrieval tests in this sense. There is, however, one fairly clear-cut example which can be discussed here. Most experiments to date have involved using just one set of requests, and trying each request on both or all the systems to be compared (i.e. 'replicating' the searches). There are clear statistical reasons for doing this, if possible: since requests are difficult to obtain for the reasons discussed above, one is usually working with relatively small numbers of them; and any statistical significance testing to be done on the results can be made much more efficient by a 'matched pairs' procedure, whereby the performance of the two (or more) systems on any one request is compared.

However, there are some circumstances under which this is not possible. If one wishes to compare highly interactive systems, for example, where the user is encouraged by the system to provide additional information about his/her need, then one cannot put the same 'request' (i.e. user need) to two different systems, since the user will have learnt too much from the first system.

Statistical aspects of retrieval testing are discussed further below, and by Tague in Chapter 5.

Measurement: performance

What are the basic measurements with which a retrieval test is likely to be concerned? Most information retrieval tests are ultimately concerned with the effectiveness or performance of each system, or the benefits which derive from its use, or cost-effectiveness or -benefit. Central to all of these questions is the question of how well the system responds to each query presented to it. This 'how well' can be looked at in many different ways: how closely each document output by the system matches the user's need; how useful each document is in satisfying the need; how satisfied the user is with the output as a whole; and so on.

It may seem strange, to anyone more familiar with the harder sciences, that I refer to such an obviously subjective matter under the heading of 'measurement'. However, it is clearly a direct consequence of my definition above of the function of an information retrieval system, that *some* such subjective notion must enter into any assessment of information retrieval system performance.

Most commonly, documents output by the system are individually assessed for **relevance** to the user's need. The word 'relevance' has been used in many different ways, but broadly it corresponds to the first of the three questions above: that is, how well does the document match the user's need. Both the notion itself and its appropriateness to retrieval tests are the subject of much debate and also some experiment. Generally speaking, the assessment of relevance allows of a 'harder' form of analysis than any other assessment in this category of subjective responses to system output, since for example it allows one to ask the question: Why did the system fail on such-and-such a document? On the other hand, utility or user satisfaction may be regarded as being closer to the true objective of an information retrieval system, and therefore better or more valid measurements to make when trying to assess system performance. The debate continues.

Measurement: costs and times

If the object of a test is to determine something about cost-effectiveness or cost-benefit, then clearly we must measure costs or some related factor. Generally speaking, one is not concerned with the overall costs of the entire system, but with the costs of certain specific parts. Thus an operational system manager might want to know what happens (to both costs and performance) if a certain part of the system is changed. I argued above that for effectiveness, one must treat the entire system as a whole. For costs, generally, the opposite is true: that is, since costs are in a strict sense additive, it is easiest and most sensible to cost only those parts of the system that may change.

This is not the place for an extensive discussion of how to go about costing a system or parts of it. It may be helpful, however, to note the almost universal use of the equation cost = time. If the difference in cost between two systems depends only on a difference in the time spent on one particular operation (say human time on indexing or machine time on searching), then one can do the appropriate cost-effectiveness comparison without ever bringing in explicit costs, simply regarding the time spent (by human or machine) as equivalent to cost. This avoids many accounting problems, and is normally the only method of including costs that is open to the laboratory researcher.

Measurement: coverage and currency

One group of variables that may be measured in connection with a particular information service consists of those which relate to the collection of documents, or to the systems for selection and acquisition, rather than to the system which retrieves from the collection. This group includes such variables as coverage and obsolescence. A considerable amount of attention has been devoted to these variables in the information science literature, under the general heading of bibliometrics. This work is, by and large, outside the scope of this book. However, one specific connection should be made.

One of the properties of a retrieval system which one might want to find out from an experiment is recall, or the proportion of the relevant documents in the collection that are retrieved. If coverage (for a particular user) is defined as the proportion of the relevant documents in the universe that are included in the collection, then it is clear that coverage (of the collection) and recall (of the system) together determine how many relevant documents the user sees, given how many there are in the universe. In other words, collection properties and retrieval system properties interact.

A second area of interaction concerns currency. In an SDI service, for example, the delay between a document being published and a user becoming aware of it is determined both by the selection and acquisition system and by the indexing and retrieval system. As these examples indicate, in the final analysis the properties of a retrieval system should not be considered in isolation from other aspects of the information service of which it is part.

Measurement: explanatory variables

One may be concerned, especially in laboratory tests, with variables which might explain or predict the final performance of the system. These variables may be subject to direct experimental control (such as a threshold in a clustering experiment); or they may be intermediate variables which need to be measured (such as the number of terms in an index language, or inter-indexer consistency). Some of the latter (e.g. number of terms again) can easily be measured and do not affect the way the test is conducted; others (e.g. inter-indexer consistency) would impose some requirements on the design or conduct of the test.

Under some circumstances, such intermediate variables might be regarded as alternatives to performance variables. Thus if we *assume* that high inter-indexer consistency goes with good performance, we can test some aspects of the system by measuring the former instead of the latter (in contradiction of my assertion earlier of the necessity of testing the entire system). The important (and in fact unsolved) problem here is of course the validity of the assumption; certainly it would seem dangerous to rely on such an untested hypothesis. However, good use has been made of similar intermediate variables in explanatory experiments or investigations.

Measurement: performance limits and failure analysis

I have said that we would normally be interested in how well the system responds to each query presented to it. But the answer to this question may well beg answers to other questions, such as: What is the best possible response to this query? How well does the system's response measure up against this ideal? What are the reasons for falling short?

If we have measured performance in terms of relevant documents retrieved, this suggests two ways in which the response of the system may have fallen short of ideal: by retrieving non-relevant documents and by failing to retrieve relevant ones. The former kind of failure will be apparent immediately if all the documents retrieved by the system are assessed for relevance. The second is more problematic—indeed, it is one of the major headaches of information retrieval system testing.

How do we find out about those relevant documents which the system fails to retrieve? In a laboratory experiment, with a small collection of documents, it might just be feasible for the requester or a substitute to scan the entire collection. But if there are more than a few hundred documents, this will be out of the question. An obvious alternative would be to sample the collection and scan the sample, but if one takes a typical operational collection and extracts a sample that is small enough for a requester to scan comfortably, it is unlikely to contain any relevant documents at all (since relevant documents are generally very sparse in such a collection).

Most tests rely on methods that are not so satisfactory in a formal sense, but are dictated by pragmatic considerations. In fact, if the object of the test is simply to make a decision between two (or more) existing systems, then there is no need to find these unretrieved relevant documents—one need only compare the relevant documents retrieved by one system with those retrieved by the other. If, on the other hand, we are testing more than one system with a view to analysing failures or assessing absolute performance, we might use the relevant documents retrieved by system B but not system A to investigate the failures of system A, and vice versa. This procedure may suffer from a form of bias: those relevant documents retrieved by B but not A may well not be typical of all the relevant documents missed by A. But we can go some way towards minimizing this bias by making systems A and B as different as possible, and/or by using many different systems (B, C, . . .) to help find other relevant documents missed by A. We can also ask the requester, before putting the request to the system, for any relevant documents that he/she already knows about.

All such methods have limitations, and unfortunately it is not known, in general, how good they are. It seems likely, for instance, that there are some relevant documents that are never retrieved, and presumably have particular characteristics that are not detected by such methods. However, no practical alternatives exist.

Obtaining relevance assessments

The very vague definition of relevance given above (how closely the document matches the user's need) is certainly not sufficient as a basis for an experiment. What aspects of the process of obtaining relevance assessments do we need to consider in more detail?

The first question is: Who is to make the assessments? In the ideal case, where the request is stimulated by a genuine information need, clearly the requester should be the one to decide on relevance. This may cause problems, since the requester may not be prepared to assess as many documents as desired (for good experimental reasons) by the experimenter. In the past, many experiments have relied on third parties, particularly when assessments are required of documents not retrieved. The third party may act as substitute or as pre-selector for the requester him/herself in the matter of relevance. This practice is regarded with increasing distrust, though it is hard in some cases to see any alternative. How reliable it is is not known.

Next, how much of the document should the relevance judge see before making a judgement? Again, the ideal is clearly the entire text of the document; but again, this is usually out of the question: usually titles or abstracts are used. There *has* been some work on the prediction of relevance (of full texts) on the basis of titles or abstracts, and it tends to show that titles alone are very bad indicators, abstracts are better but still leave a lot to be desired. It might be reasonable to postulate that for *some* tests, such discrepancies will not matter too much, as they will affect all the systems being compared equally. But it remains just that—a reasonable postulate.

The question of which documents should be judged has in effect been discussed above. One would often like the whole collection assessed, but this will usually be impossible. More likely, the judged set for each query will consist of the pooled output of various searches on different systems, including perhaps systems other than those under test, or possibly a sample from such a pool.

The order in which the documents are presented to the judge may be important. In some sense it is obviously an over-simplification to regard relevance as something which can be judged for each document independently of the others: one might more reasonably expect the judgement on any one document to be affected by which documents the judge has already seen. Ideally, one would try to devise an evaluation method which took this into account; in practice, no such method has yet been used. In these circumstances, one should try to avoid any bias that may be introduced by the ordering: one should not (if this is compatible with other aspects of the experimental design) present the output of one system and then the output of another; instead, the two should be mixed together.

Many of these recommendations may in fact conflict with other aspects of the experimental design. Thus in testing highly interactive systems, one may need to obtain relevance judgements from the requester in the course of the search. In such circumstances, it may be necessary to introduce new experimental techniques in order to avoid some of the problems mentioned above.

Finally, what instructions should be given to the judges, and in what form should the assessments be obtained? The usual method is to describe a small number of categories, such as 'Answers the question completely', 'Is of major importance in answering the question', 'Is of marginal importance', 'Is of no help at all'. Normally more than two categories are provided, although they are usually conflated into just two (relevant/non-relevant) at the analysis stage. This may seem a strange procedure, but it may be easier for a judge to use more than two categories, even if there is no experimental reason for obtaining the additional information. Also, there remains a feeling that we *should* have methods of analysis that take account of degrees of relevance; but, on the whole, no such methods exist. Some experiments on relevance have included attempts to get the judges to rank the documents rather than assign them to categories, and indeed there is some evidence that different judges are more consistent in their rankings than in assignments to categories. But again, no suitable methods of analysis exist for using ranked relevance judgements in retrieval tests.

The problem of relevance is discussed in a wider context by Belkin in Chapter 4.

Analysis

Having obtained the basic measurements, one then has to analyse the data in such a way as to answer the questions which were the *raison d'ètre* of the project. Such analysis may involve several stages: for example, we might successively:

(1) calculate, for each request and system, an appropriate measure of the effectiveness or efficiency of the system's response to the request;
(2) average this measure over the request set, for each system;
(3) compare the averages for the different systems; and
(4) perform a statistical significance test on the difference.

In fact, the subject of how to analyse retrieval test data has been, with the problem of relevance, one of the two most highly debated topics in the field. The debate was originally simply about the choice of appropriate measure (of effectiveness, cost-effectiveness, benefit or whatever), but lately it has come to include all three other aspects as well. Indeed, it is difficult to separate the four: for example, there is a statistical significance test which has been used in this context which requires that the comparison between different systems be made at the individual request level, rather than after averaging.

That said, the measures of performance or effectiveness used in the majority of retrieval tests are the well-known **recall** and **precision**. Ignoring for the moment the problems concerned with averaging over requests, these measures are usually defined as follows:

Recall = Proportion of the relevant documents that are retrieved;
Precision = Proportion of the retrieved documents that are relevant.

Clearly, these measures relate closely to the ideas discussed above concerning performance limits and failure analysis; a relevant document not retrieved or a non-relevant document retrieved is to be regarded as a failure, and the implicit suggestion is of an ideal performance of 100 per cent recall and 100 per cent precision. It should be noted, however, that there may be other (lower) limits to performance—reasons why some of these 'failures' are inevitable.

These questions are taken up again, in more detail, by van Rijsbergen in the next chapter. The bulk of the rest of this chapter is concerned, in rather general terms, with the problem of making inferences from the results of retrieval tests.

Aside from questions of performance, the main category of measurements discussed above was that of costs. Generally speaking, measurement of costs does not present the same kind of intellectual problems as measurement of performance, in that (for example) the final measure is not in dispute, and the problem of averaging is replaced by the relatively simple procedure of accumulating. (This is not to claim that costing has no problems—on the contrary—but the kinds of problems that arise are more pragmatic than conceptual.) Many of the intermediate variables such as inter-indexer consistency, however, present much the same kinds of problems as retrieval performance—though they have not received the same amount of attention.

2.3 Some examples

Detailed discussions of particular experiments are well represented in the chapters that follow, and I do not wish to pre-empt such analyses. However, it is appropriate at this point to look briefly at some experiments that have taken place, in order to illustrate the above account of the 'normal' or archetypal retrieval test, and some variants on the archetype. References are given in the bibliography at the end of this chapter.

Cranfield 2

The second Cranfield experiment (which is described much more fully by Sparck Jones in Chapter 13) was a laboratory experiment, undertaken with the object of shedding light on the construction of index languages, and the effect of different rules of construction on retrieval performance. Thus almost all of the 'system', with the exception of the translation of raw indexing into a formal language, was chosen to be as simple and unobtrusive as possible. The translation step, on the other hand, was done in a large number of alternative ways, thus generating a large number of alternative systems. The main aim of the project was to decide which of these alternative systems performed well and which badly.

The documents were 1400 real documents on the subject of aeronautics, selected rather than sampled. The 221 requests were obtained by asking the authors of selected published papers ('base documents') to reconstruct the questions which originally gave rise to these papers.

The experimental design was quite simple: each query was searched against every system. Since the searching part of the system was controlled by simple rules, there was no problem in relation to replicating searches or the order in which the systems were tried.

Measurements were made of relevance and of a number of explanatory variables. An attempt was made to obtain complete relevance judgements. The procedure adopted was as follows: students of the subject searched the entire document collection (starting with titles but consulting the full document if there seemed any possibility of relevance) against each of the requests. Documents selected by them as possibly relevant to any request were subject to final judgement by the author/requester, together with (a) the references given in the base documents, and (b) documents retrieved by one very different kind of retrieval technique.

The analysis was chiefly directed at calculating recall and precision averages, and relating these to the variables built into the experiment (concerning the construction of the index language) and to various explanatory variables such as exhaustivity of indexing and specificity of the language.

Medlars

The object of the Medlars test was to evaluate the existing Medlars system and to find out ways in which it could be improved. A few variables were built into the experiment, notably the form of interaction between the user and the system, so that the results obtained with different forms of interaction could be compared; but the main feature of the test was a detailed analysis of the reasons for failure.

The document collection was that currently available on the Medlars service, and consisted of about 700 000 items. 302 genuine queries were obtained by a form of stratified sampling. Because the requests were real ones, it was not possible to replicate searches with different forms of interaction between system and user. Hence the comparisons in relation to this variable had to be based on different request sets.

Relevance judgements were provided by the requesters. Since in such a situation there could be no question of scanning the entire collection, the testers went to considerable effort to discover *some* relevant documents that had not been found by the system (such documents were necessary for the failure analysis). The sources for these documents were (a) those already known to the requester, and (b) documents found by Medlars staff through sources other than Medlars or Index Medicus. Thus each requester was asked to judge a sample of the output from the Medlars search, together with selected documents from other sources.

After the relevance judgements had been obtained, the measurement process continued with an analysis of failures (non-relevant retrieved and relevant not retrieved). A classification of reasons for failure was devised.

Cranfield 2 and Medlars are two of the classic experiments, both playing a major part in creating the archetype as I have described it. Clearly they lie at opposite poles of the operational-laboratory spectrum, Cranfield 2 being a highly controlled and artificial experiment, and Medlars being an investigation (in the sense defined in the editor's introduction) of an operational system, as far as possible under realistic conditions. However, between them they illustrate well the main characteristics of the archetype.

In particular, Cranfield illustrates the necessity for having a complete system, even if only part of it is under test. Both tests used genuine documents; Medlars used genuine queries and Cranfield artificial (or reconstructed) ones. Cranfield used an experimental design involving replicated searches; Medlars could not. Both tests used relevance judgements by the requester; in both cases this precluded exhaustive scanning of the collection, though for Cranfield one might assume that the relevance sets are almost complete; and so on.

The two experiments described below are on a smaller scale, with more limited objectives (each, in fact, forming part of a PhD project).

Oddy: Thomas

R. N. Oddy developed a program for computer searching, called Thomas, with a strong interactive facility. The basic idea was that the system should build up, from its dialogue with the user, an internal image of the user's need. Oddy conducted a test of the program, designed to establish its feasibility and some approximate idea of its quality, rather than to measure in any very refined sense its performance.

For the document collection, a selection of 225 references (complete with indexing) was made from the Medlars data base; for the queries, 32 searches resulting from genuine requests put to the Medusa system were used. Since relevance judgements on the output of Medusa searches were obtained as a matter of course by the system, these were available to Oddy for the test of Thomas.

The test itself involved simulating a user interaction with the system, on the basis of all the information available to Oddy (statement of the request and record of the search process on Medusa, and relevance judgements on the output). Clearly this information is incomplete, in respect of both the search process that the user might have followed with Thomas and the relevance judgements (the relevance judgements affect the search process as well as the evaluation of the results). Also the very limited selection (not sample) of documents makes generalizing from such an experiment even more dangerous than usual. However, in the context of the limited aims of the test, Oddy's methods are appropriate. His insistence on using genuine requests and relevance judgements, while remaining unconcerned with the artificiality of other aspects of the test, is strictly in keeping with the philosophy of Thomas, and seems eminently reasonable in the circumstances.

Harter: probabilistic indexing

S. P. Harter has developed a theory which can be used to derive rules for automatic extraction indexing. In order to subject the theory to test, Harter performed an experiment comparing the indexing derived automatically by means of these rules to human-assigned indexing, and to a simpler automatic method. The comparison was intended to be indicative of the possible quality of the technique, rather than a definitive test of the theory.

Because of the scale of the project, and because of the difficulty of setting up a test of a complete information storage and retrieval system using different kinds of indexing, Harter decided to restrict the experiment to the indexing stage only. He therefore required a collection of documents, indexed by a human indexer, but no actual requests. The document collection also had to be in some sense realistic (as collection, not just as individual documents), and the documents had to be in continuous text form. Harter chose an existing collection of 650 abstracts of the work of Sigmund Freud. The whole collection was used for the statistical analysis of term occurrence on which the automatic indexing rule was based, and the actual comparison of index terms was carried out on a random sample of 38 of these.

For the purposes of the experiment, the human-assigned indexing was regarded as the norm, and the object of the two automatic methods was assumed to be to duplicate as far as possible this norm. Thus the rationale of the experiment depends heavily on the assumption that the human-assigned indexing is 'good'. Indeed, one might regard this procedure, in terms of the archetype, as using the set of human-assigned index terms as artificial single-term queries, and the human assignments as relevance judgements. Harter's use of a genuine collection of documents but highly artificial queries is justified by the aims and circumstances of the test.

So Harter and Oddy each chose to make certain aspects of their respective experiments as realistic as possible, but to allow artificiality in others, in effect selecting from the archetype in a manner appropriate to their objectives and resources.

Portable test collections

It will be clear from all that has gone before that any retrieval test involves a considerable amount of effort, much of which goes into setting up the test collection—that is, the collection of documents, requests and relevance judgements. Even in an operational environment, where the document collection (with indexing) is given, the queries have to be trapped at a suitable point, and the relevance judgements obtained. Many laboratory tests also involve some kind of indexing; and in any case, laboratory testers seldom have easy access to sources of queries and relevance judgements.

For these reasons, it has become common for complete test collections to be passed from researcher to researcher, and re-used many times. The best-known collection to suffer this fate is certainly the collection used in the second Cranfield experiment, described above; indeed, it would be fair to say that this collection has been grossly over-used, in the sense that it has been used for experiments which were far removed from those for which it was designed. On the other hand, given the existence of such a collection, a researcher in a laboratory environment is unlikely to feel justified or motivated to set up a new one.

There are in fact a number of collections which are used in this way: indeed, there are researchers who have become, *de facto*, the brokers for such collections, notably K. Sparck Jones in the UK and G. Salton in the USA.

Collections are normally communicated in machine-readable form (on tape); documents are usually available as the texts of abstracts, and/or some form of index representation.

The existence of these collections has had a considerable influence on the direction of research in the field, for the simple reason that some processes (such as automatic indexing from full text) are not possible on these collections as they currently exist. In these circumstances, it is at least arguable that the research community should set up one or more genuinely portable test collections: collections that are *designed* as general-purpose research tools, rather than taking on that role by accident. Although some work has been done in the last few years on the desirable characteristics of a portable test collection, no such collection has been built. But this is clearly a direction in which future laboratory work in document retrieval might move.

2.4 Statistical ideas and questions

Why statistics?

A test of a retrieval system necessarily involves, as we have seen, some kind of measurement (in a general sense of the word) of certain aspects of the way the system works. But this information about the system is of necessity historical—it concerns acts of retrieval which have already happened. The only ultimate reason for testing a retrieval system must be to discover or infer something about future acts of retrieval, either in the sense of future requests put to the same system, or in the sense of general principles (from which particular deductions about the future might be made). Such inferences are the subject-matter of statistics.

More particularly, having performed a comparison of two systems on specific samples of documents and requests, we may be interested in the statistical significance of the difference, that is in whether the difference we observe could be simply an accidental property of the sample or can be assumed to represent a genuine characteristic of the populations. Further, we may want to enlist the aid of statistical methods in discovering the underlying reasons for what we observe.

We can illustrate the peculiar difficulty of applying statistical methods to information retrieval test data by first describing an unrealistically simple situation. The rest of this chapter is devoted to an examination of the underlying problems that emerge as we try to deal with reality. More concrete recommendations and suggestions are provided by Tague in Chapter 5.

A simple case

Consider the case of an operational test which is designed to decide between two existing alternative systems, for a particular collection of documents and a particular clientele. Assume further that (a) the collection of documents is complete, and will not be added to or changed in the future, and (b) the characteristics of the clientele, and of the kinds of requests that they make, will not change in the future. Then we have a reasonably good case from the point of view of statistics: if we use a random sample of the incoming requests, and the entire existing document collection, then inferences can be made by standard statistical techniques (such as significance tests). Indeed, we can to some extent reverse this procedure, and calculate what sample size is required in order to establish a certain difference between the two systems at a given level of confidence.

Unfortunately, the situation is rarely so simple. The complications, as can be guessed from the specification of the simple case, are many and various. To a large extent, the problems are as yet unsolved; some of them admit (in principle, at least) of a statistical solution; some of them would certainly require other ideas to be combined with the statistical ones, ideas which might for example be described as linguistic, psychological, epistemological or even simply retrieval-theoretic.

Two populations

I assumed in the simple case that, in moving from the situation we are measuring to the situation about which we wish to make inferences, the set of queries changes but the set of documents remains the same. It is possible to imagine an experiment in which the two roles are reversed: an experiment concerned with certain specified SDI queries, with the document collection being completely new each month. In such a case, we would regard the document collection as a sample and make statistical inferences accordingly.

But far more commonly, we have the situation in which neither the query set nor the document collection remains the same. Even in most straightforward tests on operational systems, the document collection changes more or less gradually with time; and one is seldom in a position where one wants to know only about existing queries. So the normal case is one in which we have to consider both the test set of queries and the test collection of documents as (in some sense) samples from a population.

Suppose, then, that we can regard both samples as random: that is, in both cases, the sample is representative of the population, with no systematic differences or biases. In these circumstances, can we call in standard statistical techniques in order to make inferences about the two populations and their interactions from the measurements that we make on the samples?

Even for this (still comparatively simple) case, the answer is no: although in principle the problem remains a purely statistical one, very little exists in the way of standard methods which are formally valid under such conditions. As a result, many testers have tried to apply statistical methods which assume only one sampling process, and have simply ignored the second. Early work on these lines tended to use the document as the critical unit: that is, to regard the test collection of documents as a random sample from a population, and to ignore the problem in connection with requests. However, more recent work has tended to follow the reverse view. There are two reasons for this change. The first is that some of the measurements that have been used are query-oriented, and in order to make any inferences at all with such measures one must consider the queries as a sample (whatever one does about the documents). The second is that in general, the number of queries tends to be a much more critical quantity than the number of documents: for reasons which will be clear from earlier discussions, the tester usually has access to many more documents than requests.

One might say, then, that the state of the art consists of a number of more-or-less standard statistical techniques applied to the query set. It should be remembered, however, that such an approach deals with only part of the statistical problem.

Parametric and non-parametric statistics

Most methods of statistical inference (such as significance tests) in common use are based on assumptions about the population from which the sample is drawn, and in particular on assumptions about the distribution (in the population) of the particular variable being measured. Thus for example many significance tests assume an underlying normal (gaussian) distribution. Any such statistical method is described as 'parametric'.

Unfortunately, many of the variables that one commonly wishes to measure in retrieval tests do not satisfy these criteria. A good example is *recall*: that is, the proportion of the relevant documents that are retrieved. Because there may well be few relevant documents for any given query, and because the values of recall for individual queries may be very widely spread, the distribution of recall values over queries tends to look very strange indeed. In particular, one tends to find many occurrences of the extreme values (0 or 100 per cent), and many occurrences of those values that happen to be low-denominator fractions (e.g. 75, 33, 60 per cent).

Under such circumstances it is often difficult to find suitable parametric assumptions, and one has to have recourse to non-parametric methods. This is a fairly severe limitation: the range of non-parametric methods is somewhat restricted.

Sample size

Even supposing that the variable we are measuring would allow us, in principle, to apply some particular statistical test, are we likely to be able to obtain adequate samples of documents and queries for the test? This question has several aspects; I will consider first the purely statistical aspect of sample size.

As implied above, the documents seldom represent a problem in this context: it is normally easy enough to get hold of, and to input into the system(s), quite sufficient numbers of documents. (This is easiest if the documents are available in a suitable form; most difficult if some fundamentally new form of indexing has to be applied to them; but either way can be done given only sufficient resources.)

The real problem arises with the queries. I have suggested that 'trapping' the queries at an appropriate moment of their existence, and obtaining the necessary co-operation of the requesters, is by no means a trivial task. There is some evidence to suggest that the results of many past tests, relying on tens rather than hundreds of queries, are of doubtful validity for that reason if for no other. The problem is compounded by the large range of variation between queries of almost any variable of interest, and the comparatively small differences between systems that seem to be common.

But the question of sample adequacy is very much wider than that of numbers. We have to consider whether we can take a genuinely random sample, and if that is not feasible (or not desirable for other reasons) whether any particular method that we might use to gather the test sets is likely to introduce biases of any sort. Consider first the case of a test on an operational service which is designed to answer questions about the service itself, not generalizations. How might we take samples, and what biases might be present in them?

Time and related variables

Perhaps the most obvious problem relates to time. The tester must necessarily use documents that already exist, and queries that either occur during the course of the test, or exist in some archive at the start (or perhaps are manufactured in some way for the purpose of the test). But he/she will be concerned with the future—probably with new documents which will enter the system at a later date, almost certainly with queries that are put to the system in the future.

Thus in one strict sense, the samples cannot be representative of the situation about which the inferences are to be made. How much of this is likely to matter is an open question (and is certainly outside the realm of formal statistical inference). It is a question which has scarcely been investigated in the past. One could, however, think of ways to investigate it: for example, one could study the absolute and relative performance of different systems over a period of time. Such tests would help later researchers to assess the dangers of predicting from the past to the future, but would provide only indirect evidence on this score.

It seems likely that many possible biases introduced by time will be not so much direct consequences as indirect effects relating to other variables which are themselves time-dependent. Two examples follow.

The samples that are used for a test may not be representative of a future situation because the type of subject covered by the service may change with time. To some extent this may be a matter of deliberate planning, but it might also be because the nature of some subject that is already covered, or of the queries concerning it, change as the subject develops. Such changes may be reflected in the language of the subject, or in the internal organization of the documents about it, in a way which may have a direct bearing on retrieval.

Another change which may happen to a document collection over a period of time is that the proportions of different types of documents (books, journal articles, research reports, conference proceedings, etc.) may vary. It seems likely (although this has never been tested) that different types of documents have different retrieval characteristics: so again such a change could affect retrieval performance.

Effects of biases

It is worth looking at the last example in a little more detail, so as to see why such a bias might be important and what we might do about it.

Suppose that our document collection consists entirely of journal articles and research reports, and suppose that we are testing alternative systems A and B. We will take the existing collection (which is 90 per cent journal articles) as our test collection of documents, but suppose that five years hence the proportion of journal articles will be more like 50 per cent.

System A, as it happens, is based on the title of documents, whereas system B involves some intellectual indexing. Because research reports are on the whole longer and more substantial documents than journal articles, they are represented (on the whole) by more index terms in system B; but their titles are of very similar length, so in A the two types of documents tend to have similar size representations.

Under these conditions, we might surmise, system B is a good deal more expensive than system A, and works considerably better on reports but roughly the same on articles. Thus our test will show a marginal performance advantage to B, but at greatly increased cost; on a cost-effectiveness basis, we might well feel justified in choosing A.

But as the proportion of research reports rises in the future, the average performance difference between the systems will increase. So we may have made a mistake, as far as the situation in five years' time is concerned.

The questions that arise from this example are: how could we detect this change in the makeup of the collection; how could we assess its importance; and how could we make appropriate adjustments to our results. These questions are closely connected because we are only interested in looking for changes that may be important. The problem is, we have little idea of which variables may have major effects. Below, I discuss the paucity of results from laboratory tests that might help in this situation.

So, for the tester of operational systems, the only way ahead is to make a guess at any variables that may be important. The question of how to detect changes in these variables is clearly one of observation and further guesswork. In the example discussed above, suppose that we guess, at the time of the test, that the type of document (or the proportion of different types) might be a source of problems. Then we could examine current input to the system (as against the existing cumulated collection) to see whether such a change might already be happening. We could also look at the sources of documents and any changes that may be happening in the publication process.

Having detected a change in some variable, we want to find out whether it may have important effects. We could, in principle, include this question in our experimental design: in the example, we may have to divide the collection into journal articles and research reports, and make separate measurements on the two collections. Finally, we want to make appropriate predictions. This would involve guesstimating the possible proportion of journal articles in five years' time (or at different times over the expected lifetime of the system), and weighting the results of our test appropriately.

Artificial queries

The foregoing discussion of sample adequacy assumes that the samples are taken from a situation X, we wish to make inferences about a situation Y, and we can make some reasonable guesses about the relation between X and Y. Earlier, I suggested that there are sometimes strong reasons for constructing artificial queries rather than acquiring real ones. Obviously, a set of artificial queries is in no sense a sample of any real population, either now or in the future. Is there any way of ensuring that this set of artificial queries is representative of a real situation?

Some ways of obtaining artificial queries are: to ask some actual or potential users for examples of queries they have put to any system in the recent past; to ask intermediaries (librarians or information officers) for examples of the sorts of queries put to them; to construct queries by some random choice of index terms, in such a way as to duplicate known statistical characteristics of real queries. Any such method obviously has at its heart an intention to produce artificial queries which in some sense 'look like' real ones. The problem is, we do not really know what are the important characteristics of real ones that we should be trying to reproduce.

To a limited extent, one may test for the representativeness of artificial queries provided there are some real queries available in some form, by looking at various measurable characteristics of the real and artificial sets (any characteristics that one can think of). If a bias is detected in this way, it may be possible to allow for it in the analysis of results. But such procedures are of only limited value.

One possible statistical justification for using artificial queries is that we could, in principle, generate to order queries of a range of different types (that is, we could directly control some of the variables associated with queries). This would be a stronger justification if we possessed a reasonable typology of queries; at the moment, no such typology exists.

Laboratory tests

Again, in the foregoing discussion I have assumed that there is a specific situation about which we wish to make inferences (even if it is a postulated future situation). In laboratory tests, where we wish to make generalizations about system design, this is not the case. How then can we begin to make inferences?

If we have two alternative general hypotheses, then we can test them against each other by the usual scientific methods. That is, we have to devise an experiment from which the two hypotheses would predict different results. Because of the vagaries of individual documents and queries, almost any general hypothesis is bound to include some (explicit or implicit) statistical element in its specific predictions: that is, no general hypothesis in information retrieval (of any importance, at least) can be expected to make deterministic predictions. (To take an extreme example, we would not expect to be able to support or disprove a general hypothesis on the basis of a test on two documents and one query.) Thus we can expect statistical considerations to play a part in such hypothesis testing.

The problem is, how should we think about the experimental set-up in statistical terms? We have to regard the documents and queries as a sample from something (whether or not they actually are). The only way we can do so in general is to consider the criteria by which they were selected (or constructed), and define notional populations of all the (actual or potential) documents or queries satisfying these criteria. Then we can hope to make inferences about whether or not any general hypothesis holds for these notional populations.

As in any scientific field, although we might possibly reject a general hypothesis on the basis of a single test, one test is certainly an insufficient basis for acceptance: one must look for a number of different ways to test a hypothesis before accepting it, even if only provisionally. In information retrieval, this has generally meant testing on several different test collections (of documents, queries and relevance judgements). The reason for this form of multiple testing is that the most obvious variable (which could cause a hypothesis which works under some conditions to fail under others) is subject: the different test collections are usually in different subject areas. But little attention has as yet been paid to other variables which might cause problems, such as document or query type, or heterogeneity of the document collection in terms of subject matter or date. This lack is partly a function of availability of resources: as discussed above, test facilities which would allow such tests to be made do not exist at present and would be expensive to set up.

As I have indicated, this scarcity of results from laboratory tests on the various variables associated with document and query collections which might influence the results of retrieval tests is also unfortunate from the point of view of operational system testers. It is to be hoped that more work will be done on these problems.

Experimental design

So far, I have assumed the problem to be: 'Given the results of this test, what can we infer?'. But one can also approach the statistical aspects from the opposite direction: 'Given the sort of inferences I am looking for, how should I design my test to ensure that I get suitable results?'.

The obvious and commonest application of this idea is to sample size. Suppose that we want to ensure (at least to a certain level of confidence) that, if system A really performs so much better than system B, then the test results will lead to the correct inference. Assuming we know in advance which significance test we are going to use, and something about the distributions of the variables we are measuring, then it is possible to specify a minimum sample size to achieve this aim.

Because of the difficulties of finding suitable methods, few testers actually do statistical significance tests, let alone define the minimum sample size in advance. So this kind of procedure is not yet common in retrieval tests, though it should become more so.

A second procedure common in experimental design generally is concerned with the control of variables. Suppose that we are to do a test involving a small number of searchers (intermediaries) on a number of different systems. The object of the exercise is to compare the systems, but it may be that the choice of searcher will have a strong influence on the results for an individual query. Further, this influence may depend on the combination of searcher and system, rather than just the searcher. So we must devise a method for ensuring that the variations between searchers do not in any way distort the comparison between the systems. There are well established methods, such as Latin square designs, for coping with this kind of problem; some such methods have been used to good effect in retrieval tests.

Again, suppose we are testing alternative relevance feedback procedures. The problem is to isolate, in some way, the effect of the relevance feedback from the performance of the system without feedback. This is not an entirely trivial problem, since one must use the results of an initial search without feedback before trying the feedback procedure. Further, in this case (unlike the last) there are no obvious solutions to be brought in from outside the field. There are in fact two methods in use at present: 'residual ranking', which involves removing the documents obtained by the initial search from the collection (a different set for each query); and 'half collection' experiments, where the initial search is done on one half of the collection and the feedback is applied to the other half.

But in general, there has not been as much application of experimental design ideas in retrieval experiments as perhaps there should. This may be in part to do with the fact that so many of the variables of interest are difficult to control directly; but we might reasonably expect more such application in the future.

The limitations of statistics

Following this discussion of statistical ideas, two general points may be made. First, statistical problems are pervasive in retrieval tests; second, statistical and other considerations are closely intertwined. The process of drawing conclusions, of any sort, from the results of a test involves calling on various ideas, some of a statistical nature and some not; both sets of ideas are necessary, and they are not easily separable.

Unfortunately, many of the basic statistical problems are difficult ones, not necessarily solvable in terms of textbook methods; indeed many of them have not yet been solved. So the extent to which any experimenter can use formal statistical methods when the situation demands is severely limited. Experimenters have been in the past, and will continue to be, forced to rely on *ad hoc* methods and statistical intuition. I hope, of course, that the necessary basic work will be done for new methods to be developed; but in the meantime, I hope that the above discussion will encourage an awareness of the nature of the problems, as an aid to intuition.

2.5 Conclusions

There is no such thing as a watertight method for evaluating an information retrieval system.

There is, on the other hand, a considerable battery of methods and techniques for dealing with the various problems that arise in this endeavour. Furthermore, each new test throws up new problems, or brings out inadequacies in traditional solutions. So the archetype I have described is a fluid concept, which will no doubt change as much in the next twenty years as it did in the last. If, in 2001, this entire chapter is obsolete, so much the better!

Bibliographic notes

Barring cross-references to other chapters, the text of this chapter has deliberately been left without references, in the interests of readability.

Indeed, the best 'further reading' list must surely be the rest of this book. However, some specific matters discussed in the text must be given sources.

The reports of the Cranfield 2 and Medlars experiments are:

CLEVERDON, C. W., MILLS, J. and KEEN, E. M. *Factors Determining the Performance of Indexing Systems* (2 Vols), Aslib Cranfield Research Project, College of Aeronautics, Cranfield (1966)

LANCASTER, F. W. *Evaluation of the MEDLARS Demand Search Service*, National Library of Medicine, Bethesda, Md (1968)

Another research report of about the same vintage which discusses the methodological problems in some detail is:

CASE WESTERN RESERVE UNIVERSITY. *An Inquiry into the Testing of Information Retrieval Systems* (3 Vols), Comparative Systems Laboratory, Centre for Documentation and Communication Research, Case Western Reserve University (1968)

Some of the general issues are discussed in textbooks, e.g.:

LANCASTER, F. W. *Information Retrieval Systems: Characteristics, Testing, and Evaluation*, Wiley, New York (1968)

SALTON, G. (Ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, N. J. (1971)

The various experimental and theoretical investigations of relevance are surveyed in:

SARACEVIC, T. Relevance: a review of and a framework for the thinking on the notion in information science, *Journal of the American Society for Information Science* **26**, 321–343 (1975)

The other experiments are reported as follows: Oddy's experiment:

ODDY, R. N. *Reference Retrieval Based on User Induced Dynamic Clustering*, Ph.D. Thesis, Computing Laboratory, University of Newcastle upon Tyne (1975)

ODDY, R. N. Information retrieval through man–machine dialogue, *Journal of Documentation* **33**, 1–14 (1977)

Harter's experiment:

HARTER, S. P. *A Probabilistic Approach to Automatic Keyword Indexing*, Ph.D. Thesis, University of Chicago (1974)

HARTER, S. P. A probabilistic approach to automatic keyword indexing, *Journal of the American Society for Information Science* **26**, 197–206 and 280–289 (1975)

Portable test collections:

SPARCK JONES, K. and VAN RIJSBERGEN, C. J. Information retrieval test collections, *Journal of Documentation* **32**, 59–75 (1976)

There is no useful reference on the general nature of the statistical problems—they tend to emerge as asides or specific issues in the various research reports such as those noted above. Another such report which considers some of these problems is:

KEEN, E. M. and DIGGER, J. A. *Report of an Information Science Index Languages Test* (2 Vols), College of Librarianship, Wales, Aberystwyth (1972)

The problems deriving from the fact that there are two populations are discussed in somewhat theoretical terms in:

ROBERTSON, S. E. *A Theoretical Model of the Retrieval Characteristics of Information Retrieval Systems*, Ph.D. Thesis, University of London (1976)