7

Opportunities for testing with online systems

Elizabeth D. Barraclough

7.1 Introduction

Information retrieval testing in the early years was concerned with finding out what was theoretically possible when retrieving bibliographical records from a database. The databases were small collections of records indexed and searched manually. These early experiments produced a methodology of testing, in particular the two performance measures of precision and recall. The early computer based systems were initially concerned with demonstrating feasibility and then with trying to improve the performance of the system measured in the same way as in the manual experiments. The use of the computer allowed experiments with more complex searching techniques to be tried but most of these were done on relatively small static databases which had long since ceased to provide an information service to users.

Unfortunately, much of this work has been ignored by the system providers. Few of the techniques demonstrated to improve precision or recall, or to provide more efficient computer processing, have been incorporated in any of the large, generally available, online systems. Instead the system providers rely on the provision of extensive databases accessible to the user to sell their system. From the commercial point of view they are very successful. Users do tend to opt for the system with the most data available. The performance of such systems, in terms of precision and recall, has largely been ignored. Many of the users are unaware, or unconcerned, that they are not achieving the best that the system can provide. The time is ripe for experiments on current systems in order to demonstrate to the users the type of service they are really getting. Naturally such experiments are more difficult to perform than those in a static environment and, as we shall see, there are many constraints which can bias the results.

Lancaster, in the previous chapter, has amply covered the methods of evaluating systems including those in a real life environment. The function of this chapter is to complement the evaluation techniques and try to show how these can be brought closer to online systems. Most of the evaluation tests that have been done consider the online system as an indivisible entity. If systems are to be improved then tests must be carried out in much more detail; one experiment by Rouse and Lannom[1] goes some way along this route but not yet far enough.

An attempt will be made to define the environment in which one is experimenting; to point out the restrictions that are imposed by such an environment; and to demonstrate the opportunities that undoubtedly exist for evaluation and subsequent improvement of working systems.

7.2 The environment

The systems that are increasingly important, and therefore warrant attention, are the online bibliographical retrieval systems such as DIALOG, ORBIT, INFOLINE, etc. These systems are normally used from a terminal and with interaction from the user or intermediary, but include in their function the facilities previously available in a batch oriented computer system. The online systems, however, offer many more opportunities for analysis and provide the possibility of performing better searches as well as the pitfall of unknowingly doing very poor searches.

The essential function of any information retrieval system is to satisfy a need for bibliographical information as expressed by the user. The advantage of an online system over a batch system accessing the same database is that it allows the user to modify the expression of his requirement in the light of information provided by the system. We ought to expect that in these circumstances users would achieve better results from an online system. Tests in such an environment should aim to prove or disprove this belief. From the results of such investigations one would hope to determine methods of improving the system and gain a greater insight into the real needs of users.

7.3 Constraints on real life testing

The tests that need to be applied to any information retrieval system are concerned with determining whether a client's information need is satisfied by the system. Both the need and the satisfaction are more difficult to define in an online system than in a batch system. The starting point for either system is the client, who may have a written question or merely an idea in his head. This question is translated into a search statement generally with the help of an intermediary. For a batch system this statement is deemed to be the definition of the user's need; with an online system this is not necessarily the case since the user can, and should, change his search statement in the light of information supplied by the system. The difficulty is to decide which, if any, of the search statements used actually defines the user's need, and in particular whether the need as opposed to the definition of it, has changed during the search process.

Satisfaction is equally difficult to define. A user of an online system can be satisfied in a number of ways. The client requiring all references on a topic obviously needs high recall, and is prepared to tolerate some reduction in precision. Such clients are similar to former batch system users, who, because of the intrinsic delay in the early systems, only undertook literature searches when they required a full bibliography.

Online systems with immedate access to the database also provide for the

user who wants to gain an entry into a subject by a 'quick and dirty' search done at low cost. User satisfaction is almost guaranteed for these clients provided one relevant reference is found within a reasonable time. At the other extreme are clients who want to ensure that their chosen area of research is untouched, i.e. they want to find that there are no references in the literature.

The situation facing the experimenter who wishes to carry out tests in an online environment is thus quite complex. The users can come with a variety of needs; the definition of these needs is not precise in every case; measures of satisfaction depend upon the type of need; and performance of the system is also related to the expressed need. An example of the difficulty, or perhaps futility, of attempting to measure satisfaction is shown by Lawton, Auster and To[2].

However, one should not despair: an online system attracts a large number of users, who, by appropriate methods, can be classified. These methods will rely to a large extent on interviews with the user both before and after the search is undertaken but an attempt should also be made to classify users' needs automatically using parameters such as: search complexity, e.g. number of conjunctions in a search statement; number of references printed; and number of references deemed to be relevant. Similarly, the individual user's need would have to be elicited by interview but should be checked against the various search statements attempted. Producing such a classification of types of requirements and methods of assessing the user's needs is a subject worthy of study on its own and needs to be undertaken before extensive tests can be carried out.

7.4 User constraints

Given that parametrization of the users and their needs is possible, the only major constraint on the user sample is the reluctance or inability of some users to take part in an experiment because of the confidential nature of their research. To exclude such users completely could introduce a significant bias into the results. Users from industry could have a rather different requirement from the system which could be neglected if they were unable to take part. Most users are very willing to take part in experiments particularly if the cost of the service is reduced (preferably to zero) for participants.

Despite all these problems it should prove possible to select various groups of users to test the ability of the system to supply a variety of user needs. Both the function of the system considered as a whole and the facilities provided within different parts of the system, e.g. for search formulation, should be investigated.

7.5 System and database constraints

Real life systems impose particular constraints on experimenters which are associated with changes over time. Working systems do not stand still; the data is continuously changing as new information is added to the database. As the files grow the database is split and only the most recent part is kept in

the current file. Thus any recall and precision assessments must be done at the same time to ensure that the same part of the database is used otherwise comparisons are not valid. System providers can also add new databases to their files and hence extend the search possibilities; again, by making comparisons at the same time, this problem can be avoided.

Investigation of the facilities provided by the system can be affected by changes that the system providers may make over time. Thus they could offer more facilities for inspecting the dictionary and any statistics collected over such a change would be invalid.

Perhaps the most serious constraint imposed by the system on the user and hence on the test possibilities, is the time pressure caused by the method of charging. Many users, or their librarian intermediaries, are very conscious of the cost and therefore look for a quick solution rather than taking time to consider how to achieve a better result. The system providers are in some difficulty here; they have to implement a charging mechanism which covers both the cost of running the system and the cost of creating and making available the databases. They are also constrained by the facilities that are provided by the machine manufacturers for implementing a charging scheme. As a result of these constraints most systems base their charges on the time the user is connected to the system and the number of references retrieved and printed. The second parameter is probably quite justifiable. The first parameter also has benefits for the system provider as it encourages users to access the system at slack times as an overloaded system will run more slowly and thus be more expensive. From the user's point of view, this method of charging is most unfortunate. He is charged for thinking time, term selection and searching in the database at precisely the same rate. He ought, at least, to be able to think 'freely'. Also many facilities that only require access to the dictionary are not costly in computer time or storage, but would, if provided cheaply, offer the means of formulating much better searches.

For users who are accessing the systems from remote locations, particularly in the UK, the charging is confounded by the fact that telecommunications costs also have a component which is duration dependent in addition to the charges related to the volume of data.

The effect of this crude method of charging for system use is to force users to take a gross view of the system rather than considering the effectiveness of different parts of the system. This naturally imposes the same limitations on the tests that can be attempted, unless the constraint can be removed by providing free searches or a local charging mechanism.

7.6 The opportunities of real life systems

The advantages of testing in a real environment, for many purposes, outweigh the constraints that they impose. The most immediate advantage is the possibility that tests carried out with real systems can bring direct benefit to the users in a realistic time-scale as opposed to testing on static systems where the application of the results is not immediately obvious.

The feature of online systems, which cannot be overestimated for testing and evaluation, is the ability to observe user behaviour. This can be done in

considerable detail by collecting all the information from a terminal session, both the input from the user and the responses from the system. Agreement must, of course, be obtained from the user to take part in the experiment before observing his behaviour in this way. Suppression of identifying information should also take place. Having once understood that the experiment is going on, the user ceases to be aware that his interaction with the system is being investigated. It is thus possible to get a completely unbiased user view of the system.

Collection of logged data gives a very clear picture of what the user does with the system, in some cases why he does it is also obvious, i.e. correcting mistakes, but the same difficulty that exists in determining the user's information need from the search statements applies to the analysis of the logged information. Some extra information from the user is necessary.

In addition to the collection of complete sessions which can be subsequently analysed, statistics of overall use of the system can be derived and investigations done on the usage of commands, for example, which are misused, which tend to be repeated and which are hardly ever used.

7.7 System aspects that warrant testing

The overall testing of the performance of the system can be done by treating it merely as a 'black box' with input at one end and bibliographical references at the other. However, if methods for improving the system are being sought, then the separate functions provided within the system must be considered. The majority of operational systems can be thought of as having three functional parts.

(1) Search formulation and checking.
(2) Index search.
(3) Database search and print.

The methods used in each section can vary from system to system with perhaps the greatest differences being in the search formulation area. Briefly, the functions are as follows.

Search formulation and checking

The user with his intermediary approaches the system with an information need which can be merely thoughts in the user's head or, more likely, a written statement of the information need, or a detailed search request with appropriate terms already selected. The function of the system obviously varies depending upon the amount of work that has been done beforehand. The minimum work that the system will do is to check that terms exist in the dictionary or index and provide a count of the number of occurrences. At the other extreme the user can try terms to see if they exist and look for related terms where the relation can be alphabetical proximity or a subject relationship; the user can then select terms suggested by the system rather than having to think of them *ab initio*. Thus in this area the user can get considerable assistance if he is prepared to pay for it. Some intelligence is built into most systems in that suffixes and prefixes can be ignored if

requested. Similarly, word adjacency can be utilized as well as other context indicators.

Index search

The purpose of this part of the system is to provide for the user a count of the number of references which satisfy his formulated request. The system also creates a record of the citation numbers for subsequent retrieval from the database. The user has no control over this part of the system, but he may, of course, return to the formulation stage if he is not satisfied with the number of references retrieved.

Database search and print

The function here is merely to extract references from the database and print them in a form requested by the user. Most systems offer a wide range of alternative styles of printing from the brief reference giving only sufficient bibliographic details to identify the citation to a full reference giving all the data, including abstract or index terms, held for the citation. The user may select the style, whether to print locally on his terminal or elsewhere on a fast printer and how many retrieved references to print. He has no control over the selection of the retrieved citations to be printed, generally the systems provide the most recent references first, but nothing is guaranteed.

The user at any stage can return to amend his search in the light of the information supplied by the system. It is this interaction which should make online systems much more powerful than the original batch systems.

Most systems provide access to more than one database, with the system's functions used in the same way on every database available to the system. However, it is not possible to ignore the database being used when assessing the facilities provided by the system functions. For example, the method by which index terms were assigned can make a vast difference to the efficacy of the systems.

The user can thus apply his search to more than one database in the system and may retrieve the same reference from several databases as there is often considerable overlap between databases. What is potentially worrying is the retrieval of a reference from only one database when it is known to be covered by another database, the different indexing or abstracting methods being generally responsible for this.

7.8 Types of tests

Search effectiveness

This type of test, as Lancaster has pointed out in the previous chapter, is a macroevaluation; the system is being treated as a 'black box' and the evaluation is concerned only with the results the system produces not with the methods used to reach them.

The criteria for evaluation of effectiveness depend upon the type of search being attempted. Possible user needs were previously identified, each of which would need a different set of parameters in the evaluation.

The comprehensive search should be assessed by measuring recall and precision. For this type of search in a real life system one would be attempting to assess the actual performance of the system in its normal operating mode against the capabilities of the system as exploited by the experienced experimenter. The user, or his intermediary, would perform the search in the normal way using whatever interactive tools he wished, e.g. iterating through a series of formulations and prints to get a satisfactory search, finally ending up with a formulation which would be deemed to be a correct expression of his needs. The experimenter would provide a broad search on this topic which would include the user's formulation as a subset and the user would be asked to assess the full output. From this the precision of the user's formulation can be established and an estimate of the recall from the overlap between the full set and the user's subset.

The user who wishes to find a few references to provide an entry into a subject has very different criteria for a satisfactory search. Precision is still of interest but recall assumes a minor role, the most important aspect is finding the references quickly and easily. For this type of search all the references would be retrieved online, so a parallel search by the experimenter is not a possibility. In this case analysis of the whole session data would prove very valuable. One might wish to measure the number of references retrieved before the first relevant one, also the number of formulations that were necessary, with the number of relevant citations for each, before a satisfactory formulation is reached.

A good example of the detailed analysis that needs to be done is the study by de Jong-Hoffman[3] of a single search on the INSPEC database. It provides a lot of ideas for points that should be investigated in a wider study.

System effectiveness

The macroevaluation outlined above gives no indication of the reasons for failure or success. In an online system the investigation of the reasons, or microevaluation, can be carried out as an extension of the macro study. In the first case, where a comparison was possible between the user's search and the experimenter's, the differences between the two searches should be investigated, i.e. the relevant references missed and the reasons for this. It is here that the collection of the complete session data is of value. It can then be seen which terms were omitted that in the broader search retrieved relevant references and, in a more detailed investigation, how these terms were omitted from the search and could have been found from the system.

Many other aspects of the system can be investigated by analysis of complete session data collected as part of an evaluation test. For example, the usefulness of commands can be determined by looking at the sequences of commands in a session. One would expect a command showing related terms in a dictionary or thesaurus to be followed by the selection of some, or all, of those terms. The proportion of terms so chosen is a measure of the value of the command. Collection of statistics of both command use and the time taken for the execution of the commands can lead to proposals for the improvement of the system.

An attempt has been made to point out some of the aspects of online information retrieval systems that could lend themselves to testing and

evaluation. A much broader view is taken by Penniman and Dominick[4] ranging from the theoretical basis for investigating interactions with computers to methods of analysis of the data collected, it provides a very good foundation for the practical evaluation of systems.

7.9 Conclusion

Online systems are becoming one of the major sources of bibliographic information and yet few studies have been done to determine their effectiveness or methods of improving their performance. Such systems can yield, at little cost, a great deal of data which could be used to analyse the systems at both the macro and the micro level. The expectation is that these tests would indicate where significant improvements could be made to current systems, the benefit would thus be much more immediate than tests in a static environment. The opportunity is there and should be used.

References

1. ROUSE, S. H. and LANNOM, L. W. Some differences between three on-line systems: impact on search results, *On-line Review* **1**, 117–132 (1977)
2. LAWTON, S. B., AUSTER, E. and TO, D. A system evaluation of the educational information system for Ontario, *Journal of the American Society for Information Science* **30**, 33–40 (1979)
3. DE JONG-HOFFMAN, M. W. Research into the practical results of on-line information retrieval: an extensive analysis, *Aslib Proceedings* **29**, 197–208 (1977)
4. PENNIMAN, D. W. and DOMINICK, D. W. Monitoring & evaluation of on-line information system usage, *Information Processing and Management* **16**, 17–35 (1980)