Performance measurement of information seeking environments

Contact: Steve Draper, Department of Psychology, University of Glasgow, Glasgow G12 8QQ
(More contact details: http://www.psy.gla.ac.uk/~steve/)

Introduction

An IRC (interdisciplinary research collaboration) is proposed for exploring a wide range of evaluation methods for information seeking environments (ISEs), with a view to establishing new standards of evidence for assertions of quality. These methods will range from computer science benchmarking methods requiring no human users, through HCI (human computer interaction) methods such as thinkaloud protocols and controlled experiments with human users, to workplace and domestic ethnographic field studies. These methods draw on a range of disciplines.

The information retrieval (IR) field has for long had a consensus view that evaluation should be based on test collections, and this has long been accepted as the normal standard for comparing IR software and so resolving claims about what design is best. This consensus on what counts as good evidence supported steady progress for several decades. There is now however a growing recognition that the impact of new technology (e.g. multimedia documents, interactive software, the WWW, digital libraries) means that that traditional standard, involving test collections but not users, will no longer suffice. There is however little agreement about what methods should replace it. The proposed IRC will allow multiple evaluation approaches to be developed and compared intensively, working towards an eventual new consensus on evaluation. Without an IRC, work is likely to consist of scattered cases of mixed evaluation methods, each constrained by the particular skills of the researcher and the features of the case being studied, but is unlikely to lead to a systematic exploration, comparison, and understanding of methods in this field. An IRC in contrast could sustain the latter, lead to an expanded delimitation of the field, and re-unify it with a new consensus on what is acceptable as evidence of "good" performance.

Since an essential part of the new focus is the recognition that the contribution of the human user is a key determinant of the success of a retrieval episode, the goal is to evaluate not a retrieval engine (a software function) but the whole environment in which the user operates: hence "ISE" rather than "IR". The evaluation approaches studied will thus apply not only to IR software (using indexing of all words in documents, ignoring any structure) but also to databases and to attempts to combine both these technologies in single facilities. The proposed work is complementary to that on "digital libraries"; the latter term referring mainly to the development of basic storage facilities for multi-media document collections. In the absence of concentrated attention on the effectiveness for users of attempts to retrieve from such collections, digital libraries could develop technically, yet in practice only be usable by librarians and a few other information specialists, and not by the much larger number of the users of the information itself. Appropriate evaluation techniques are important to guiding the design of ISEs that are both effective and usable. New technical developments require new approaches to evaluation, while retrieval systems are becoming ever more important not just in libraries and archives, but in the internet and the intranets now becoming central to the way many companies and organisations operate. Thus at a time when a new attack on evaluation has become necessary for scientific and technical reasons, the field of information retrieval has become much more important to society as a whole, in terms both of the numbers of people using it and of the commercial importance of ISEs. For the same reasons it is also no longer a preserve of specialist users but is impacting a very wide range of people, and so requires much higher standards of usability.

In summary, ISEs are becoming pervasive and so enormously important in many aspects of work and home life: that is the ultimate social relevance of this research. The technical means addressed here for improving their quality are to develop measures of their effectiveness both in relation to their underlying technical functions (as in traditional IR evaluation) and in relation to the role of the human users and furthermore in relation to human-human and social interactions associated with their use. Scientifically this corresponds to the issue of how to update the long-standing orientation of the IR field to evaluation with approaches that can address the many new technical features of digital libraries that have emerged in recent years. This will involve both relatively old (e.g. textual IR and relational databases) and relatively new technologies (e.g. image collections, and emerging digital library developments); and in a variety of task contexts (e.g. lab. tests, workplace and home computing field studies). An IRC is appropriate and probably essential for this because these two ranges — of evaluation methods and of new data technologies — are each more extensive than a typical research group can cover by itself, yet should be considered together in order to work towards a new unified evaluation approach. The proposed IRC will build on a nucleus of existing collaborations among both academic and commercial organisations to explore a full and wide range of techniques adapted from a variety of original disciplines to a single IT area — that of information seeking and retrieval.

The information retrieval perspective

The field of information retrieval (IR) began in the late 1960s, addressing the problem of retrieving text documents from large collections, by computer, based on full-text indexing of words. It has always been characterised by a strong focus on evaluation: on methods of measuring retrieval effectiveness, traditionally just the performance of the software engines. This dominates how most research is now done and reported, but probably stems from the peculiar problem of not being able to judge the quality of any retrieval by simple inspection of the results: you can only judge them if you have extensive knowledge of what might have been retrieved by that query on that collection (using a perfect engine) and that is inherently expensive information to acquire, requiring a much more systematic and formal approach to evaluating test results. The emergence into widespread use both of multimedia rather than only text documents, and of interactive user interfaces has extensive implications for this field and the evaluation methods on which it depends which are far from being worked out.
How can we evaluate how good an engine is at finding documents which the user considers relevant? The traditional measures used are precision and recall. Precision measures the fraction of the retrieved documents which are in fact relevant to the user’s information need, and recall measures the fraction of those documents the user would consider relevant which were actually retrieved. Obviously, the practical measures of each of these measures varies widely with the type of task being done (e.g. contrast “find any two examples of...” with “find all papers on...”). Furthermore, to calculate recall-precision figures an experimenter must have a collection of documents to search, a collection of queries to find documents for, and a set of independently made decisions as to which documents are “really” relevant to which queries. Constructing such a “test collection” of documents, queries, and relevance judgements is very time consuming and open to criticisms of bias in the judgements. Considerable efforts have been invested in building large standard test collections (e.g. Cranfield (1) and TREC (2)). By creating a standard set of queries and relevance judgements, the test collection approach has removed the end users from the evaluation loop, representing them by the queries and judgements stored in the test collections. This may be acceptable when the search techniques are non-interactive and it allows fast experimentation, but it also makes it extremely hard to evaluate the worth of interactive techniques such as user relevance feedback, user query expansion and changes to the interface.

Information retrieval systems were initially designed to be used by intermediaries in a library setting. These trained searchers would interview a user to build up a model of their information needs and then carry out searches at a later date — often specifying very clearly the information (or topic) that the user was looking for. This is not how people search the Internet — users often have only a loosely formed notion of what they are looking for when they start a session and often have very little idea of what the collection will contain on that topic. This stark difference in user population from the traditional models of IR is one of the challenges facing modern IR researchers. As well as Internet searching, the widespread use of encyclopaedias on CD, large volume hard disks and cheap, very fast personal computers has led to many end users with no computing training using search engines on fairly large collections of text. Furthermore, the speed of the engines, the spread of mouse and window user interfaces, together with non-specialist users has made repeated exploratory retrievals the normal procedure, rather than single carefully designed queries. The net effectiveness of a session, or at least a set of retrieval attempts, has become much more relevant than the performance of a single retrieval cycle. Users typically find interesting information at many different steps in a session which not only is used to modify their query formulations but may also modify their goals and relevance judgements (a point made as early as 1973 (3)). This means that to study, measure, and optimise the useful work done with an IR program, we must measure the retrieval done by an interactive user over a set of retrieval cycles. This will depend partly upon the software, but also partly upon what the user does. The “system” being studied is not the function computed by one call on the retrieval engine, but the combined human-computer interaction over as many cycles as the user is observed to initiate in the course of one task. This redefinition of “system” affects how evaluation must be done, what measures can be used to compare designs, and of course the designs themselves — for example, features of the user interface may prompt users to formulate better or poorer queries or to try more/fewer cycles.

At first sight, this redefinition of the system to be designed and measured might not seem to require much change to the evaluation method. Simply set the user the retrieval task, take what they finally select at the end of a session as the result, and again consult the stored “answers” in a test collection in order to measure the combined performance. In addition, direct observation of the users (for instance, by think aloud protocols) would yield useful formative information about how user interface features affect performance and could be improved. However, things are not so simple (3, 4, 5, 6, 7, 8) and have begun to be addressed by the interactive track of TREC (9) and Mira (10). The first problem is that of how to “set the user the retrieval task”. In test collections, these are often specified by the query that would be typed directly into the software. But the formulation of that query, given a goal in the user’s mind, is one of the major steps in the overall task, and it strongly affects the outcome. Consequently in many cases test collections would have to be rebuilt with search tasks specified in more realistic ways; and furthermore should be backed by other studies of what tasks occur in actual work places, and in what forms. Borlund (11) is currently researching the use of “simulated work tasks” as a way of addressing this problem while Reid (12) is investigating including task aspects in test collections. However this would still only address those tasks where the user begins with a definite and articulated retrieval task. But only a little observation of real users shows that a lot of retrieval concerns browsing, not just as a method but as a type of goal where the user just looks for something “interesting”, not something definitely known in advance. There is a growing need to increase the usage of HCl type evaluations to the IR field. There have been a number of user studies of information retrieval software (e.g. 13, 14), often published in the HCl rather than IR literature. These need to become a standard component of IR research. Rasmussen and his collaborators have developed a comprehensive framework of the issues in human-machine-work interactions that can be used to guide evaluation (15, 16, 17). It covers a wide spectrum of evaluation from low level issues such as the user noticing and understanding the output from an interaction through to measuring how the IR system has helped them achieve their work goals. However, a difference in the evaluation of IR systems, which has started to be addressed at Risø, is the linking of interface design evaluation with the performance of the underlying IR engine, whereas the design of the interface of a word processor, say, may require a lot of improvements but there is little need to measure the accuracy of the word storage facilities. In contrast to this complex framework, Harper & Hendry presented the notion of Evaluation Light (18, 19): concentrating on using very focused small experiments to answer constrained questions concerning users’ interaction with IR systems (similar in spirit to Andrew Monk’s work on lightweight HCI evaluation techniques (20)). Another lightweight technique in IR is to use limited user modelling combined with the test collection approach (21).

Up to this point these arguments for major changes in our evaluation approach have been based simply on taking the interactive nature of modern IR software seriously. However the technical basis of IR has changed in many ways in recent years as computing has advanced. A fuller list of the important features of this includes:

- Interactive software.
- The WWW (world wide web): by far the most important collection, but huge, dynamic, and unsearchable by any current engine. (Current engines admit they cannot cover the whole web.)
Each of these developments entails its own evaluation problems — technical developments require matching developments in evaluation methods, which however have not advanced so rapidly. Our first impulse was to develop a multimedia test collection. However it is now apparent that this is much too ambitious, and also in some ways backwards looking. Firstly, the effort just of collecting and organising the documents is a major research direction of its own (now often called "digital libraries"). Secondly, assembling expert judgements of relevance for the whole of such collections seems likely to be beyond reach, given that technology can now support such big collections. Thirdly and most importantly, however, is that many of the objections to test collections that were worrying for text documents now seem overwhelming in the context of new technology: could anyone really identify a “representative task” for all users and all media?

Analysing retrieval from non-textual collections, such as collections of paintings or photographs, provides an insight into the problems of evaluating state-of-the-art IR systems. One approach is to use traditional meta-tag indexing to provide access to the image by attributes such as photographer, date of photograph and similar external information. To access the content of the images, we could add a set of keywords (“meta-tags”) to each image. In this way, textual queries are used to retrieve non-textual documents: cross-medium retrieval. A second use of associated textual descriptions is to index text which already exists and is, somehow, related to the image (e.g. 22, 23, 24) as found in, say, web-based art collections (25). This approach avoids the problem of a human having to create text just to make the images indexable. The alternative to using text in any form is to analyse the content of the image but this leads to a multitude of problems: high level attributes are very hard to extract, low level attributes may bear little resemblance to items users would wish to search for and there is a much wider set of possible relevance relationships for images than for texts. Most image search systems currently use techniques such as colour histogram and texture matching (26, 27, 28) between query and document images, possibly in combination with main object shape detection. While these approaches have shown considerable success in finding images which are visually similar to each other, it is extremely hard to move away from this visual similarity to a more semantic matching: there are only a limited number of tasks in which you are looking for an image and know the texture and colours of the matching set. It is extremely difficult to devise evaluation approaches which categorically answer questions such as which of these techniques is most suitable for which users.

Field studies of how IR is used in real work are particularly important as we face the problems posed by the new horizons in IR. For instance, a study of a commercial image bureau showed that in this business at least, image retrieval is done by text queries not because that is the only thing current technology supports but because that is how the customer specifies and thinks about what they want. Similarly, as mentioned above, they can uncover kinds of relevance that IR engines, so far, have almost no way of representing. As new approaches to IR evaluation worry about what kinds of user task really exist and matter in practice, workplace studies can collect them. Studying new classes of IR user, for instance WWW users or school children (29), shows how these users do not come with any prior search skills at all: the success of IR software here will depend either on having the interface communicate such skills or else by avoiding the need for them altogether. However, workplace studies are expensive to do, as they absorb many hours of investigator time (although they are invaluable for the above reasons); but their expense means that they will not replace other kinds of study. Thus HCI studies in which participants representing users are invited in to use software will retain a place in IR evaluation e.g. for rapid improvement of the user interface. It is likely too that the benchmark style of study using recall and precision will retain some place. Combinations are likely to become more important: for instance, inviting participants into the lab not to use the whole program but to test a small part of it against benchmark measures.

Future work will be characterised by attempts to explore basic tensions in direction. One is the tension between highly standardised workbench tests using precision and recall and no human users (fast and highly comparable with the work of others thus good for competitions, but with doubtful relevance to any real work applications) vs. workplace studies (highly valid, but expensive and of doubtful comparability with each other). Another is the tension between comprehensive evaluation using the Rasmussen et al. framework exhaustively vs. lightweight techniques that are more often affordable in practice. One of the main directions to address is that of characterising, and measuring performance with respect to, other types of user task than those specified by explicit pre-given queries. If evaluation is to correspond to large amounts of current retrieval in practice, it must find a way of measuring how well a retrieval session went with respect, not just to concrete goals like “Rembrandt’s last painting”, but to “browsing” goals of just looking for something interesting, and also to explicit but vague goals such as looking for a “nice” or “novel” or “beautiful” picture.

All these considerations imply a basic shift in evaluation to consider the whole system — not just the software functions, but the active role of the human user in a repeated cycle of query formulation, retrieval by the software of a set of documents, and selection by the user from that list. It follows that the resulting evaluation methods should apply to all information seeking environments: not just to IR but equally to database software. We have not found any published studies that directly compare the two kinds of software for effectiveness, but the evaluation approach proposed here would allow this. As both kinds of software are currently being presented through very similar interfaces on the WWW, and as in our experience many users are unaware of the differences both in function and in the effective user query strategies required by each, such studies will be important and timely. The proposed IRC will therefore include work on (interfaces to) databases.
Our proposal, then, is to adapt and apply a wide range of evaluation methods to ISEs: methods originating from psychology and sociology as well as computer science, and already often applied in areas such as HCI and CSCW as well as in home and workplace studies done for other reasons. It is clear on the one hand that new methods are already proving valuable and need to be widely explored in ISEs, and on the other hand that much exploration is needed to identify what the important issues are in each new technological context. Only then can we begin to formulate the questions that evaluation of ISEs should address, and the methods by which each question might be answered. These are our aims, and the usefulness of the methods in other areas and in some studies in the IR field (e.g. 19) allow us to expect substantial progress. Finally, in addition to developing methods and accumulating cases of their application, we would work towards an evaluation framework capable of specifying criteria for selecting a method, or mix of methods, for any given situation.

The HCI perspective
The main case (above) has been written from an IR perspective: that of designers of ISEs in general and IR software in particular, and the evaluation needed to measure the effectiveness of the resulting software, and so support its improvement. It is worth commenting briefly on the proposal from an HCI research perspective.

To a considerable extent, this topic could be seen as applying HCI to a particular application area. We hope to apply evaluation methods from HCI to ISEs, and also to apply some more theoretical ideas from HCI: in particular Green's Cognitive Dimensions of notations (29, 19). This should not only benefit the area of ISEs but conversely serve as a testbed for developing HCI practice. Certainly the area exhibits a feature that has emerged as crucial in HCI in the last decade: the "turn to the social" i.e. the realisation that effectiveness often depends not just on an individual user's activity, but on the social context of computer use comprised of other users and the human-human interactions that result, and of a workplace and its demands. These aspects have led to the increasing adoption of methods such as ethnography capable of addressing them. In this respect, this application area is at the centre of current HCI.

It has another important feature: the continuing importance of the functional quality of the software. Although some frameworks have always focussed on utility (31), much HCI in the past has tended to focus on usability costs only and neglected utility benefits as an issue, because functionality in, for instance, word processors, is seldom a limiting factor in overall effectiveness. In the IR field, performance remains an important limiting factor: IR engines continue to return a considerable proportion of irrelevant documents while missing some others that would be very useful to the user. (The permanent failure to attain or even approach functional correctness is very unusual in computer science as a whole. It is of course why evaluation has been important from the beginning in the IR field.) This is important for evaluation methods — it is not enough (unlike in many other HCI application areas) to measure only user problems with the interface: machine failures are significant here, and must equally be measured and taken into account when deciding on improvements. This requirement to combine measures of utility and usability also makes ISEs a particularly good topic for concentrated HCI research with a view to generalisability.

Since this topic is defined by ISEs not HCI, we will not limit ourselves to HCI in the methods we draw on: we will also draw upon traditional IR test collection approaches, and on ethnography (which seldom sees itself as part of HCI, even though many in HCI now see it as valuable). Nevertheless ISEs should be an important test application for HCI with some critical features that will explore and test the adequacy of its current techniques, and so can be expected to contribute to it. We anticipate that a considerable proportion of the papers produced by the proposed IRC would be published in the HCI literature, and would serve to advance it.

Training
We anticipate approximately equal numbers of research students, funded research fellows, and supervising academic staff, and plan for full participation by the students in the IRC workshops, and for some other travel by them to other sites, as appropriate. A pending application to the EC for a research training network, led by Glasgow, on a related topic may bring additional young researchers to the Aberdeen, Glasgow and Sheffield sites, the possibility of sending research students to European sites working on ISE evaluation, and the possibility of summer schools to which all could send IRC-linked students.

This collaboration
The Aberdeen, Glasgow and Sheffield groups in this proposed IRC have collaborated in the past, most recently in the Mira working group (32), which culminated in a recent international conference (33), and which also involved Steve Robertson, now at Microsoft Research, Cambridge, and who has agreed to an active participation in the IRC.

The collaborators and their roles are as follows.
- Steve Draper, Dept. of Psychology, University of Glasgow. From the HCI research group "GIST": applying HCI theories and evaluation methods.
- Keith van Rijsbergen, Dept. of Computing Science, University of Glasgow. From the IR research group; formal theories of IR.
- David Harper, Dept. of Computing & Mathematical Sciences, the Robert Gordon University, Aberdeen. IR, databases, and ISE design; "light" evaluation methods.
- Micheline Beaulieu, Dept. of Information Studies, University of Sheffield. Participant in TREC; design, formal evaluation, and user studies of IR systems.
- Richard Harper, Digital World Research Centre, University of Surrey. Ethnography, field studies of domestic technology.
Thomas Green, Computer Based Learning unit, University of Leeds. Theory of Cognitive Dimensions; controlled HCI experiments.

John Eakins, Dept. of Computing and Maths, University of Northumbria. Image retrieval.

Steve Robertson, Microsoft Research Ltd., Cambridge. Has agreed to sit on the advisory committee, and to attend the workshops. Further joint research will depend upon other commitments.

Existing research resources
Each of the academic partners, represented above by a single lead person, in fact consists of an important research group with extensive research experience in this area. In addition, the EPSRC-funded Revelation project at Glasgow (which concerns high capacity multi-media servers) is likely to prove an additional important collaboration, and Northumbria has important image collections and the hardware to hold them.

Involvement of users of the research
Representatives from industry will be invited and encouraged to attend the IRC workshops, which will ensure the delivery of research results directly to some of the most relevant users. Besides informal, but probably influential, feedback from them there, they will be represented on the advisory committee. In addition, we hope to carry out some joint research directly with them, as they will have access to either equipment or ISE users directly relevant and helpful to the proposed research. Organisations with which one or another of the IRC academic partners currently have active collaboration in this area include those listed below. Many of these are expected to participate in the workshops. Some have already been approached and have agreed to attend at their own expense.

Steve Robertson, Microsoft Research ltd., Cambridge
John Snyder, Muscat ltd., Cambridge
BT Cellnet
Vodafone
Orange
Granada

Dr. John Davies, BT Labs, Martlesham
Glaxo Wellcome
ICL
BBC
One-2-one
The Post Office

Organising the multi-site, multidisciplinary collaboration
The IRC will be composed of geographically separated sites, each with different research emphases. The main mechanism for sharing and exchanging results and approaches will be workshops held approximately twice a year.

In our experience, multidisciplinary work requires participating researchers to operate on two fronts: maintaining their involvement in their core discipline by publishing and attending conferences central to that discipline, and by participating in other events across disciplines. Here, each site's budget will allow for substantial conference attendance in each researcher's own main discipline, while the IRC workshops will be the focus for cross-discipline interactions with a further allowance for personal site visits within the IRC as particular joint research activities arise. Organising the workshops in a way that maximises useful exchanges, comparisons, and constraints will be one of the IRC Director's main functions (as opposed to slipping into the academic default of juxtaposing unrelated papers, delivered with little time for questions or discussion). In the Mira group, workshops often involved exercises organised by one or another participant, which not only proved popular but served to introduce and illustrate key issues or methods to other participants, and this may prove useful here too.

Management
Since each site will have a separate contract with EPSRC, only EPSRC will ultimately have the power to change resource allocations. However there will be the need for an active role in promoting the multi-site, multidisciplinary interactions, and in coordinating the IRC as a whole, particularly in disseminating its results outside the IRC. It is proposed to have a "director" to do this, who will devote half his time to this: Steve Draper. His research career has been entirely involved in interdisciplinary activities of one kind or another: a PhD in Artificial Intelligence within a cognitive studies group, a post doctoral post at San Diego in a cognitive science group from which emerged an HCI group, collaborative HCI research at Glasgow between the Computing Science and Psychology departments.

The management of the IRC will be organised, besides the Director's role, as follows:

- EPSRC review about the 3 year mark, which alone could change resource allocation. The IRC will present its recommendations, if any, for change to the review panel for consideration, along with possible minority statements from individual sites.

- An advisory committee including members from industry would meet yearly. It would review progress and, as needed, recommend changes to research topics or priorities. Steve Robertson from Microsoft has already agreed to sit on it; two more members from industrial collaborators would be appointed, and three from academic sites.

- An operational management committee, consisting of the lead investigator from each academic site, would meet every two months, physically or "virtually" as proves effective.
Outline resource budget

For the purposes of this outline proposal, the budget is structured in terms of "units". One unit consists of one research fellow, plus one academic (not requiring salary but requiring travel and equipment) and one postgraduate student (the studentship not included in these costs, but requiring travel and equipment money in order to join in this IRC effectively).

The average research fellow salary is taken at spine point 8 to allow for a) the higher salary usually needed in computer science posts because of competition from industry, b) the higher salary needed for those with ethnographic training which is long and often means hiring relatively senior researchers, c) to allow for annual increments over 6 years: one of the advantages of IRC funding is that there is a real chance to retain at least a few of the researchers for a long period. Spine point 8 means per year, in round figures, £20k net, £24k gross, £34k with 40% overheads.

The IRC would be organised around twice yearly workshops. With sites as far apart as Aberdeen and Guildford, travel might be £200, and £200 for accommodation, making it £400 per person per workshop. Another site visit independently of the workshops would be £400. A USA conference costs about £1,800 and a UK conference about £800. Allowing 2 workshops, one site visit, one USA and one UK conference per person means £3.8k p.p.a. (per person per year). Basic computing support (a desktop machine plus networking costs and consumables) is estimated at £1k p.p.a.

One unit is one salary and three people's travel and computing: £48.4k p.a. Allowing 3 units each for Aberdeen, Sheffield and Surrey, 4 units for Glasgow, and 1 unit each for Leeds and Northumbria makes 15 units or £726k p.a.

The larger sites would require some technical and administrative support: estimated at £20k p.a. each for 3 sites. In addition there will be some special computing hardware e.g. Aberdeen requires, as a one time charge, £50k of hardware for an experimental IR collection server. For the IRC as a whole, £20k p.a. Buyout of 50% of the director's time would be about £18k net, £22.5k gross, £31.5k p.a. with overheads. About £110kp.a. for these headings.

Total about £836k p.a. or £5M over 6 years.

Suggested Referees

Prof. John Tait, School of Computing & Information Systems, University of Sunderland.

Prof. Alan Smeaton, School of Computer Applications, Dublin City University.

Prof. Nick Belkin, School of Communication, Information and Library Studies, Rutgers University, New Jersey.

References


(32) Mira working group URL http://www.dcs.gla.ac.uk/mira/

(33) Mira final conference URL http://www.dcs.gla.ac.uk/mira/workshops/conference/cfp.html