New IR — New Evaluation:
The impact of interaction and multimedia on information retrieval and its evaluation

Stephen W. Draper

Department of Psychology, University of Glasgow, Glasgow G12 8QQ U.K.
email: steve@psy.gla.ac.uk URL: http://www.psy.gla.ac.uk/~steve/

Mark D. Dunlop

Department of Computing Science, University of Glasgow, Glasgow G12 8QQ U.K.
email: mark@dcs.gla.ac.uk URL: http://www.dcs.gla.ac.uk/~mark/

Abstract

The field of information retrieval (IR) traditionally addressed the problem of retrieving text documents from large collections by full-text indexing of words. It has always been characterised by a strong focus on evaluation to compare the performance of alternative designs. The emergence into widespread use both of multimedia and of interactive user interfaces has extensive implications for this field and the evaluation methods on which it depends. This paper discusses what we currently understand about those implications. The "system" being measured must be expanded to include the human users, whose behaviour has a large effect on overall retrieval success, which now depends upon sessions of many retrieval cycles, rather than a single transaction. Multimedia raise issues not only of how users might specify a query in the same medium (e.g. sketch the kind of picture they want), but of cross-medium retrieval. Current explorations in IR evaluation show diversity along at least two dimensions. One is that between comprehensive models that have a place for every possible relevant factor, and lightweight methods. The other is that between highly standardised workbench tests avoiding human users vs. workplace studies.

Introduction

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 focus on evaluation: on methods of measuring retrieval performance, 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. This paper discusses what we currently understand about those implications. Although the views developed here are those of the authors, they have been extensively informed by the activities and participants of the MIRA working group, which was set up by the Commission of the European Union to study these problems (1, 2, 3). These in turn reflect a growing awareness in the IR community as a whole of the challenges implied by these technical developments.

IR typically deals with large datasets ("document collections"). Although initially developed as a method for matching users' information needs with abstracts of academic papers, IR software has expanded to cover massive collections of text (the latest benchmark test environment contains roughly 2 gigabytes of uncompressed text comprising around 1 million documents). Furthermore, the texts are now taken from a wide range of sources, for example full academic papers, full newspaper articles, television news broadcast transcripts, literature and personal papers. Retrieval is based on matching documents to words entered as a query. Internet search engines are probably the first exposure most users currently have to the kind of service (and interface) offered by information retrieval techniques. Although based around standard techniques, these engines perform significantly more poorly than the best laboratory engines at matching relevant documents, but succeed at meeting the unprecedented size and speed requirements demanded of web searching. The next two sections introduce standard techniques from IR before analysing the challenges posed to the field by recent developments. We cannot go into details on how search engines work — the fierce competition in this area has led to some secrecy over the techniques used — and do not have the space to go into standard IR techniques in great detail. For a full overview of many of the techniques in IR see (4), for a more theoretical angle see (5), and for an excellent collection of classic IR papers see (6).

An outline of Information Retrieval

In a typical IR system, the data objects are text documents such as newspaper articles or journal papers. The user enters a query consisting of a number of words in the same language as the documents. The software then returns a long list of documents that match the query to some extent (i.e. all documents that contain at least one of the query words somewhere), ordered by goodness of match.

IR stands in contrast to database retrieval, where the data is fully structured and retrieval is based on queries using that structure. In IR, the data's structure is usually ignored[1]: both the internal structure of the document (e.g. title vs. abstract vs. main text), and the structure of the language (its syntax and semantics, word order, whether two words are in the same

[1] Note however that, as with almost everything we describe, there is research exploring numerous variations to the most common approach.

sentence or paragraph). IR software builds an index as an alternative way of accessing the documents. This increases the total size of the data significantly (normally by at least 30%) but precomputes some of the work involved in retrieval, cf. program compilation. An important feature is the automatic creation of this index (as opposed to assignment of keywords for each document by a human expert), which allows IR software to be automatically applied to data not originally created for searching. A further feature of IR is that the queries are typically made in unstructured natural language, i.e. without artificial constructs such as AND and OR, again in contrast to database retrieval. Furthermore, most work is focused around the idea that users will be looking for documents on a specific topic.

The following is a typical approach to indexing for a basic IR system: extract all individual words from the text, remove stop words (words which have no meaning when taken out of context, e.g. "and", "the": see (5) Ch. 2), conflate the remaining words to their base stem form (e.g. treat "place", "placing" "places" as the same term (7)), and weight words inversely to how often they occur in the collection (8) (a word that occurs in only one document would be given a high weight so that the document that contains it is ranked highly if the query contains that term, while a term that occurs in every document should get zero weight since it cannot discriminate relevant from non-relevant documents). Queries are then matched by processing the query text in a similar way and using a scoring function that takes into account the number and weight of the matches between query terms and their appearance in each document. This typically results in a large list of partial matches, but, again unlike database retrieval, ordered by their score so that the best are presented to the user first.

Much of the work in IR has focused on improving the indexing and matching algorithms described briefly above. Techniques such as different weighting algorithms, matching phrases as well as single words, and thesaurus expansion of query terms have all led to measurable improvements in the ability of search engines to match the topic of the user's query to documents in the collection. One extremely successful technique, which relies on user interaction, is relevance feedback (9).

When a user is presented with a list of suggested documents, relevance feedback systems provide the opportunity for the user to mark which of the documents are actually relevant to their information need. This information is then used to produce a revised list of documents which reflect both the initial query and the marked relevant documents. (The software selects words from the marked documents and composes a new query by adding and reweighting terms.) Although a conceptually simple technique, relevance feedback is very powerful because it is a way for the user to give more information to the system on what kind of documents are likely to satisfy his or her information need simply by pointing to documents — objects that are meaningful to both human and software — without having to think about what terms (words) are both used in the desired documents and are likely to discriminate them from non-relevant documents.

Evaluation in textual IR

Given the above description of information retrieval, how can we evaluate how good an engine is at finding documents which the user considers relevant? The most common 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 (so high precision means few unwanted documents are suggested), and *recall* measures the fraction of those documents the user would consider relevant which were actually retrieved (so low recall means many relevant documents are missed). Obviously, the practical importance 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…").

The overall performance of engines is, in most work, expressed and compared in recall-precision graphs like figure 1. An ideal engine would perform at the top-right of the graph but in practice recall and precision are usually in a tradeoff relationship, as in the typical curves shown in the figure, where as recall rises precision falls. Thus as users go further down the ranked list produced by a retrieval they will find more relevant documents, thus increasing recall, but usually at the cost of an increasing proportion of non-relevant documents, hence reducing precision. Figure 1 shows a typical recall-precision graph comparing two retrieval engines. Engine A (the higher line at the left hand side) performs very well for low recall (when the engine is finding a small percentage of the relevant documents) but performs worse than the other engine when having to find more than 50% of the relevant documents. On the other hand, engine B has lower precision for low recall but its precision drops less for higher levels of recall.

These recall-precision graphs are the standard measure in IR. Note however, that they do not capture some of the information you might want. For instance, you cannot extract from these graphs (nor, therefore, from many published experiments on IR engines) what

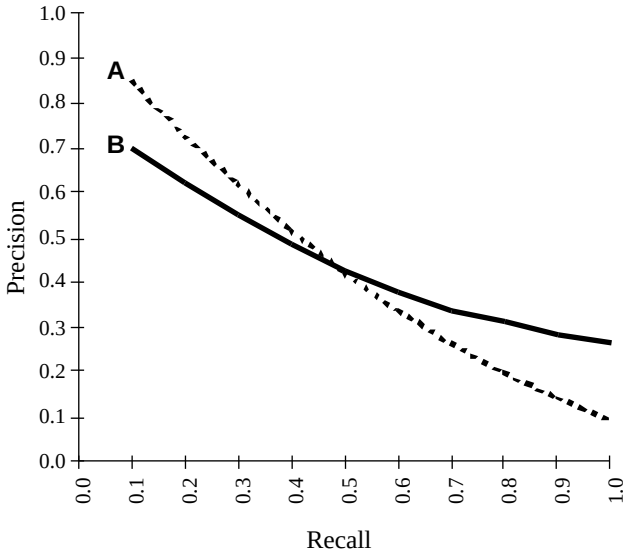

Figure 1: A Recall-Precision graph (fictitious data)

recall and precision would be obtained by using, say, the first 10 documents retrieved: something of central importance to many users and hence to designers.

These graphs are actually produced by calculating a recall-precision value pair for each relevant document on the list generated by each retrieval being considered in the test. Recall and precision pairs are typically averaged at ten standard recall points (0.1, 0.2, ..., 1.0) over all queries in the set. To average the value pairs, all recall values in the range, say, $0.1 \leq r < 0.2$ are considered to contribute to the 0.1 standard recall point. The precision of this point is then calculated as the average of all precision values for recall-precision pairs with a recall value in this range. (For more details see (5), which also explains standard smoothing approaches.)

To plot a recall-precision graph 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 for the large sets of interest to IR researchers, and open to criticisms of bias in the judgements. Considerable efforts have been invested in building large standard test collections: notable test collections include Cranfield (10) and TREC (11).

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, user term selection, and improvements in ranked list descriptions of documents.

Expanding the concept of "system" to include the user

Information retrieval systems were initially designed to be used by intermediates 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 clearly 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 (whereas librarians often have a very good idea of the content of the collections they are searching). 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 (12)).

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 instance, features of the user interface may prompt users to formulate better or poorer queries, to try more or fewer cycles (perhaps stopping before the best retrieval has been achieved), and so on. The quality of a single retrieval cycle may be relatively unimportant in determining whether the session converges on a high quality final outcome. Instead, software features that have been observed to prevent users noticing how a document is relevant (by failing to scroll forward to the parts where the required terms occur), or that it is relevant at all (by using "surrogate" summaries in the ordered list that fail to communicate its relevance to the user and so lead to them not opening the document at all) often have a bigger effect on how well the overall task is performed than the engine's matching algorithm.

New Evaluation models for IR

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 as has been pointed out periodically and with increasing frequency (12, 13, 14, 15, 16, 17) and has begun to be addressed by the interactive track of TREC (18)

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. For instance, when we have observed school children doing a task such as finding the date the Queen was crowned from a newspaper archive, it has had a large effect whether the task was specified as "What date was the Queen's coronation?" as opposed to "When did Elizabeth II become queen?", since the documents are more likely to contain "coronation" than "become queen". The latter form of the task specification would be a harder test of the user plus supporting

software features such as a thesaurus to help users reformulate queries. 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. Pia Borlund (19) is currently researching the use of "simulated work tasks" as a way of addressing this problem.

However this would still only address those tasks where the user begins with a definite and articulated retrieval task ("information need"). 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. Many uses of the WWW (World Wide Web) are examples of this. For instance you might look up someone's web page, not to retrieve a definite fact such as their fax number, but to see if they had anything interesting there, and end up printing out a paper. You might tell an evaluator that this was a highly satisfactory result, but it could not be measured against a panel's agreement about which were all the documents of interest to you personally that you might have found given better software.

A first need in this area, then, is to establish better concepts to use as "discussion tools". How shall we think of user tasks, needs, goals and queries? Thomas Green (20) argues for the importance of discussion tools in related areas such as human computer interaction (HCI) and the design of programming languages. The same doubtless applies in IR.

The next need is to introduce some HCI type evaluations to the IR field. There have already been a number of user studies of information retrieval software (e.g. 21, 22), often published in the HCI 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 (23, 24). 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. Figure 2 shows the model as a set of concentric layers. Many of the layers relate to traditional elements of evaluation in HCI work. However, a difference in the evaluation of IR systems 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 that complex framework, Harper & Hendry presented the notion of Evaluation Light (25): concentrating on using very focused small experiments to answer constrained questions concerning users' interaction with IR systems. This is similar in spirit to Andrew Monk's work on lightweight HCI evaluation techniques (26). Another possible lightweight technique in IR is to use limited user modelling combined with the test collection approach (27). Harper's framework includes a series of rules of thumb, such as:

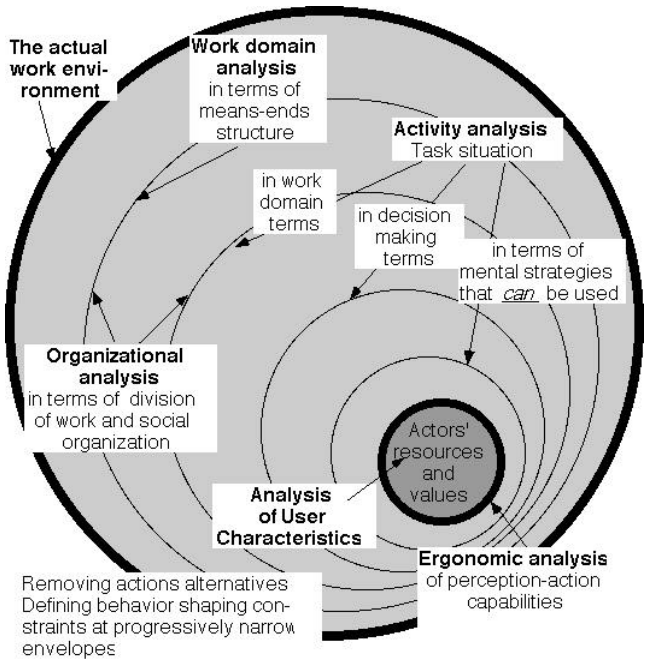

Figure 2: Rasmussen, Pejtersen and Goodstein's framework

- Evaluate using small, controlled, tasks directed at evaluating a given claim or hypothesis;
- Break down a large hypothesis or design conjecture into discrete, smaller hypotheses or conjectures;
- Know your intended users, and brief test users accordingly;
- Use other people's results where possible.

The framework indicated in figure 2 – a "heavy" evaluation approach – claims to be comprehensive and to apply to all human-machine interaction from nuclear power stations to work processors. "Light" approaches concentrate on maximising useful data gathering that guides improvements to software design in a short timeframe, compatible with the small resources that are all that are allocated in many cases. Clearly this is a good starting point for changing IR design practices – to aim first for large benefits for small costs. However equally clearly, such a "light" – i.e. cheap but best-first – approach could not claim to be comprehensive.

Multimedia information retrieval

So far we have only discussed matching textual documents to textual queries. A whole new set of challenges are introduced for the design of suitable interfaces, algorithms and evaluation methods for non-textual collections such as images, video or mixed media.

Over the last few years the spread of digitisation equipment and very large capacity disks have led to many large collections of images being stored on-line. These range from

massive collections of photographs held by photographic supply bureaux through large collections of photographed paintings held in art gallery collections to personal collections of snaps held either on Photo-CD or as the output of digital cameras. How do we access these image collections? There are, broadly, two approaches: querying text associated with the images, or supporting direct querying of intrinsic image properties such as colour.

Thus, 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. While this could be regarded as cheating, it is actually important and in fact essential in some work domains: for instance, in the art world where retrieving by artist and picture title is required all the time. (Having to sketch the artist in order to retrieve paintings by Rembrandt is not what these users need.) On the other hand it is not only slow to create the keyword indexing, but it suffers from the same consistency problems that led to the development of automatic indexing approaches for textual documents.

A second use of associated textual descriptions is to index text which already exists and is, somehow, related to the image (e.g. 28, 29, 30), which avoids the problem of a human having to create text just to make the images indexable. For example in web-based art collections there is usually considerable textual descriptions surrounding images which not only can be used to describe the images (31) but is also likely to be of a nature that would support art historians searching these collections.

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, whether we are indexing high or low level attributes, 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 (32, 33, 34) 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.

For text IR, attention has been focused, mostly subconsciously, on two assumed tasks: academics searching for research papers or journalists searching newspaper archives. These tasks are both purely about topic matching and disregard many attributes of relevance (e.g. Fidel (35) worked with Boeing and discovered relevance criteria such as the user having influence over the issues the document raised). While these assumptions are strong, users can often succeed in using a text-IR system despite their limitations. With images, however, it is even less clear what the real user tasks would be and there may be a

much wider range of tasks to be supported by the same image collection.

An important avenue to investigate to address this problem is the use of relevance feedback versus explicit user queries. In relevance feedback, users inspect items retrieved by an earlier query or even by a random retrieval, and mark them as relevant or not (good or bad with respect to the user's real goal); the software then reformulates the next query tacitly by using whatever hidden attributes those items have, without the user ever having to know or express or recognise those attributes. It may be that IR systems for images can succeed using this technique, exploiting peoples ability to scan images very quickly. Certainly it avoids the need for inventing an image query language that the user has to input, such as sketching the kind of picture wanted.

Some general problems raised by multimedia retrieval may in fact be reflected in the text domain when multi-lingual IR is considered. Firstly, it is not entirely clear how well IR techniques generalise to languages other than English, which has an unusually large vocabulary and a relatively low reliance on surface syntax which suits the standard IR technique of ignoring syntactic structure and distinguishing (only) word-stems. Secondly, cross-language retrieval (e.g. 36) is like cross-medium retrieval (query in one language to retrieve documents in another). This is important in practice for instance in Switzerland, which has four official languages, and although government officials may be able to read all four, they typically will be able to compose queries most fluently in one while needing to retrieve all documents in any language that match the query's meaning.

The above discussion concentrated on image retrieval, but the points made will apply again to each new medium considered (e.g. speech retrieval). True multimedia retrieval will then further raise the importance of cross-medium retrieval.

Workplace studies

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, among other things, that in this business at least, image retrieval is done by text queries not because that is the only thing current technology supports (a "cheat") but because that is how the customer specifies what they want, 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 (information needs) really exist and really matter in practice, workplace studies can collect them. Studying new classes of IR user, for instance WWW users or school children (1), shows how these users do not come with any prior search skills at all (for instance the idea of paraphrasing the task into terms more likely to occur in the target collection): success of IR software here will depend either on having the interface communicate such skills or else by avoiding the need for them altogether (e.g. by relying on relevance feedback as the main user input mechanism).

Workplace studies are expensive to do, as they absorb many hours of investigator time,

although they are invaluable for the above reasons. Their expense however 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 may become important too: for instance, inviting participants into the lab not to use the whole program but to test a small part of it against benchmark measures. For instance, one subtask users of most IR software have to do is to select from the ordered list returned which items they wish to "open" and inspect further: clearly if they never look at an item it will not be one of those ending up as a product of the session. This selection is based on "surrogates": the short (often one line) descriptions of each item shown on the lists. Such lists can be tested alone with users using focused experiments (37).

Future directions

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 affordable in practice. One of the main directions in which we must hope for progress and certainly expect substantial efforts 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 to "browsing" goals of just looking for something interesting, as well as looking for a "nice" or "novel" or "beautiful" picture.

Acknowledgements

The authors are grateful to the Commission of the European Union for funding the Mira working group on evaluation frameworks for interactive multimedia information retrieval applications (Esprit Working Group 20039), to the stimulus of which the views in this paper are largely due. We acknowledge the contribution of all participants at the Mira workshops, but particularly that of Keith van Rijsbergen of Glasgow University for directing it, and Fabio Crestani, Norbert Fuhr, Alan Smeaton, Marion Crehange, Yves Chiaramella, Catherine Berrut and Peter Ingwersen for their work in organising the Mira meetings.

We also thank an anonymous reviewer for moving this paper further beyond being a report on Mira and nearer to being an adequate reflection of the field as a whole.

References

(1) MIRA (1998) *MIRA: Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications*, http://www.dcs.gla.ac.uk/mira

(2) DUNLOP, M.D. (1996) (editor). *Proceedings of the Second Mira Workshop* (Monselice, Italy). University of Glasgow Computing Science Research Report TR-1997-2, and http://www.dcs.gla.ac.uk/mira/workshops/padua_procs

(3) FUHR, N., VAN RIJSBERGEN, C.J., AND SMEATON, A.F., (Editors). *Evaluation of Multimedia Information Retrieval*, Dagstuhl-Seminar-Report-175, Schloss Dagstuhl, April 1997.

(4) FRAKES, W.B., AND BAEZA-YATES, R., (Editors). *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992.

(5) VAN RIJSBERGEN, C.J. *Information Retrieval* (second edition). Butterworths, London. http://www.dcs.gla.ac.uk/Keith/Preface.html 1979.

(6) SPARCK JONES, K., AND WILLETT, P. (Editors). *Readings in Information Retrieval*, Morgan Kaufmann, 1997.

(7) PORTER, M.F. , "An Algorithm for Suffix Stripping", *Program*, vol. 14(3), pp. 130-137. July 1980.

(8) SPARCK JONES, K., 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, 28, 111-21, 1972.

(9) SALTON, G., AND BUCKLEY, C. "Improving retrieval performance by relevance feedback". *Journal of Americal Society of Information Science*, **41**, pp. 288-297, 1990. (reproduced in 6)

(10) CLEVERDON, C.W., MILLS, L., AND KEEN, M., Factors Determining the Performance of Indexing Systems, ASLIB, Cranfield Project, Cranfield. 1966.

(11) HARMAN, D.K. "The TREC Conferences". In R. Kuhlen and M. Rittberger (Eds), *Hypertext - Information Retrieval - Multimedia: Proceedings of HIM 95*, Konstanz, Germany. 1995.

(12) COOPER, W.S. "On selecting a measure of retrieval effectiveness". *Journal of the American Society of Information Science*, **24**, pp. 87-100, (reproduced in (6)) 1973

(13) SARACEVIC, T, "Relevance: A review and a framework for the thinking on the notion in information science". *Journal of the American Society of Information Science*, **26**, pp. 321-343.1975

(14) BELKIN, N.J., AND VICKERY, A. *Interaction in information systems: a review of research from document retrieval to knowledge-based systems*, British Library – Library and information research report 35, 1985.

(15) BORGMAN, C.L. "All users of information retrieval systems are not created

equal: an exploration into individual differences", *Information Processing and Management*, **25**(3), pp. 237-252,1989.

(16) BATES, M.J. "Where should the person stop and the information search interface start?", *Information Processing and Management*, **26**(5), pp. 575-591, 1990.

(17) MIZZARO, S., "How many relevances in information retrieval?", *Interacting with Computers*, to appear 1998.

(18) BEAULIEU, M., ROBERTSON, S. AND RASMUSSEN, E. "Evaluating interactive systems in TREC". *Journal of the American Society of Information Science*, **47**(1), pp. 85-94, January1996.

(19) BORLAND, P., AND INGWERSEN, P., "The development of a method for the evaluation of interactive information retrieval systems", *Journal of Documentation*, **53**(3) pp 225-250, June 1997.

(20) GREEN, T. "An Introduction to the Cognitive Dimensions Framework", *Proceedings of the Second Mira Workshop* (Edited by M.D. Dunlop), http://www.dcs.gla.ac.uk/mira/workshops/padua_procs, Glasgow University Research Report, 1996.

(21) KOENNEMANN, J, AND BELKIN, N.J. "A case for interaction: a study of interactive information retrieval behavior and effectiveness", *CHI96 Conference Proceedings*, (Edited by M.J. Tauber, V. Bellotti, R. Jeffries, J.D. Mackinlay, and J. Nielsen), pp. 205-212, 1996.

(22) PIROLLI, P., SCHANK, P., HEARST, M., AND DIEHL, C. "Scatter/gather browsing communicates the topic structure of a very collection", *CHI96 Conference Proceedings*, (Edited by M.J. Tauber, V. Bellotti, R. Jeffries, J.D. Mackinlay, and J. Nielsen), pp. 213-220, 1996.

(23) RASMUSSEN, J., PEJTERSEN, A.M. AND GOODSTEIN, L.P. *Cognitive systems engineering* (Wiley: New York) 1994.

(24) PEJTERSEN, A.M. "Emperical work place evaluation of complex systems", *Proceedings of the 1st International Conference on Applied Ergonomics. (ICAE'96)*, pp21-24, Istanbul, Turkey, May 1996.

(25) HARPER, D. AND HENDRY, D. "Evaluation light", *Proceedings of the Second Mira Workshop* (Edited by M.D. Dunlop), Glasgow University Research Report, http://www.dcs.gla.ac.uk/mira/workshops/padua_procs,1996.

(26) MONK, A. F. "Lightweight techniques to encourage innovative user interface design". In L. Wood (Ed.), *User interface design: Bridging the gap from user requirements to design* , CRC Press. pp. 109-129. 1998.

(27) DUNLOP, M. D. "Time Relevance and Interaction Modelling for Information Retrieval", *Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR97)* (Edited by N.J. Belkin, A.D.

Narasimhalu and P. Willet), Philadelphia, pp. 206-213, 1997.

(28) DUNLOP, M. D. *Multimedia Information Retrieval*, PhD Thesis, Glasgow University Computing Science Research Report 1991/ R21, October 1991.

(29) FRANKEL C., SWAIN M. J., AND ATHITSOS V. *WebSeer:An Image Search Engine for the World Wide Web*, University of Chicago Technical Report TR-96-14, 1996.

(30) SMEATON, A.F. AND QUIGLEY, I. "Experiments on Using Semantic Distances Between Words in Image Caption retrieval", *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR96)*, Zurich, pp.174-180, August 1996,

(31) HARMANDAS, V., SANDERSON, M., AND DUNLOP, M.D. "Image retrieval by hypertext links", *Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR97)* (Edited by N.J. Belkin, A.D. Narasimhalu and P. Willet), Philadelphia, pp. 296-213, 1997.

(32) FLICKNER, M., SAWNHEY, H., NIBLACK, W., ET AL. "Query by image and video content: the QBIC system", *IEEE Computer*, **28**(9), pp. 23-30, September 1995.

(33) EAKINS, J.P., HARPER, D.J., AND JOSE, J. *Proceedings of The Challange of Image Retrieval* (Edited by J.P. Eakins, D.J. Harper and J. Jose), *Electronic Workshops in Computing*, to appear 1998.

(34) FOUNTAIN, S., AND TAN, T. "Content based annotation and retrieval in RAIDER", *Proceedings of IRSG98* (Edited by M.D. Dunlop), *Electronic Workshops in Computing*, to appear 1998.

(35) FIDEL, R., AND CRANDALL, M. "User's perception of the performance of a filtering system", *Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR97)* (Edited by N.J. Belkin, A.D. Narasimhalu and P. Willet), Philadelphia, pp. 198-205, 1997.

(36) HULL, D.A., AND GREFENSTETTE, G. "Querying across languages: a dictionary-based approach to multilingual information retrieval", *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR96)* (Edited by H.P. Frei, D. Harman, P Schäuble and R. Wilkinson), Zurich, pp. 49-57, August 1996.

(37) TOMBROS, A., SANDERSON, M., AND GRAY, P. "The advantages of query-biased summaries in Information Retrieval", *Proceedings of AAAI'98 Spring Symposium on Intelligent Text Summarization*, to appear 1998.