Reflections on Mira: interactive evaluation in information retrieval

Mark Dunlop
Centre for Human Machine Interaction, Risø National Laboratory, PO Box 49, Roskilde 4000, Denmark
phone +45 4677 5166, fax +45 4677 5199,mark.dunlop@risoe.dk, www.chmi.dk/people/mdd/

Evaluation in information retrieval (IR) has focussed largely on non-interactive evaluation of text retrieval systems. This is increasingly at odds with how people use modern IR systems: in highly interactive settings to access linked, multimedia information. Furthermore, this approach ignores potential improvements through better interface design. In 1996 the Commission of the European Union Information Technologies Programme, funded a three year working group, Mira, to discuss and advance research in the area of evaluation frameworks for interactive and multimedia IR applications. Led by Keith van Rijsbergen, Steve Draper and myself from Glasgow University, this working group brought together many of the leading researchers in the evaluation domain from both the IR and human computer interaction (HCI) communities. This paper presents my personal view of the main lines of discussion that took place throughout Mira: importing & adapting evaluation techniques from HCI, evaluating at different levels as appropriate, evaluating against different types of relevance and the new challenges that drive the need for rethinking the old evaluation approaches. The paper concludes that we need to consider more varied forms of evaluation to complement engine evaluation.

Introduction

This paper aims to give an overview of the results and discussions that took place during the Mira working group on evaluation in interactive and multimedia information retrieval (IR). Many of the issues discussed here have a long history in one part or another of the literature. However, Mira brought together different evaluation communities in an attempt to develop techniques that are applicable to modern highly interactive IR systems being used to access multimedia information stored in linked, distributed and dynamic collections. This focus on multimedia and modern interactive systems forced attention away from traditional underlying machine measures and back to issues often identified long ago but not systematically progressed.

Periodically throughout the Mira programme reference was made to “removing the straightjacket” of the traditional evaluation approaches in IR (Dunlop 1996, Harper 1997, Robertson 1999). One of the

main goals of this paper is to encourage IR researchers to consider novel evaluation methods that have been developed and tested in other fields. More importantly, throughout the Mira meetings a concern emerged that the IR community was restricting its research to areas that could be evaluated using the traditional techniques. If so, there is a risk for the field in focussing on improving only the inner engine and not supporting end users by making improvements to the whole IR system. After a brief background to Mira and traditional IR evaluation, this short paper will discuss the motivations behind Mira and the main threads of discussion in Mira

Background to Mira

Mira was funded by the Commission of the European Union Information Technologies Programme (ESPRIT) as a Working Group: bringing together sites that work on related topics but were not directly funded for staff or equipment. Consequently no core research work was carried out directly within the *project*. To support the strong focus on meetings, the six workshops and a final conference were organised to include interactive and practical sessions so that work was conducted during the meetings and not simply reported. For example, workshop attendees designed evaluation schemes for various IR scenarios, carried out live think aloud evaluations on each other and analysed videotaped interactions. In addition to the regular 14 Mira members, the workshops had many external attendees (see acknowledgements for details). This paper will focus on the main work discussed throughout the Mira programme, those interested in more general reviews of evaluation in IR are directed elsewhere (Draper & Dunlop 1997, Harter & Hert 1997, Tague-Sutcliffe 1996, Harman 1992, Voorhees 2000, Voorhees and Harman 2000). Further information on Mira, including conference proceedings and on-line copies of many presentations are available at the Mira Workshops Website[1].

The test collection approach

The traditional test collection evaluation approach (e.g. Van Rijsbergen 1979, Sparck Jones and Willett 1997) simulates a retrieval environment in an artificial computer-based benchmark or *test collections*. These are composed of the following four components:

1. a set of documents;
2. a set of questions/queries on topics covered by document set;
3. a set of judgements listing which documents are relevant to which queries;
4. an IR engine to index the collection and run the queries (plus evaluation software).

Evaluation is based on measuring how effectively a system finds relevant documents. This is commonly measured using recall and precision, defined as follows:

$$\text{recall} = \frac{\text{number relevant documents retrieved so far}}{\text{total number of relevant documents}}$$

$$\text{precision} = \frac{\text{number relevant documents retrieved so far}}{\text{total number of documents retrieved so far}}$$

Standardisation on the first three elements has enabled IR researchers to relatively quickly test new engines, or engine variants, against the *collection* and have results comparable with the work of others. In particular, the TREC conference format has led to a focus on overall engine effectiveness over a large standard collection of text. With many participating research groups, TREC has resulted in noticeable and consistent improvements in engine effectiveness over the first seven years of the TREC initiative (Voorhees and Harman 2000, Buckley et al 1998).

[1] http://www.dcs.gla.ac.uk/mira/workshops/

Motivations

Mira was established to examine evaluation of interactive and multimedia IR systems. Early on it became clear there were two forms of multimedia: non-interactive and interactive. The latter category poses many challenges to the traditional approach and is the basis for this paper and most of the work within Mira. Here it is the interactive element that dominates and, so, interactive multimedia evaluation is discussed in the same context as text-only interactive IR. Non-interactive multimedia retrieval is interesting in that it highlights the narrowness of the common notion of relevance in IR. For example, in an archive of fine art images the possible reasons for a user selecting a picture and rejecting another are many more than the traditional *topical relevance* used in text collections. Different types of relevance are discussed later in the context of user evaluation but it is worth noting that this issue also affects test-collection based non-text evaluation.

Interactive evaluation of IR systems introduces a fundamental problem for the classic test collection approach: interaction implies end users, who bring their own knowledge, experience and searching abilities. The human-computer interaction (HCI) community is firmly grounded on evaluating end user interaction with systems. Unlike IR, HCI does not normally have to deal with the problems of also measuring the effectiveness of an underlying engine. Many times during Mira the analogy with cars was used: motor manufacturers need to improve the performance of engines both in terms of power output and fuel efficiency, hence figures for fuel efficiency, top vehicle speed etc. But they are not the full story: comfort, storage capacity, resilience to wear, depreciation and government transport policies are some of the many aspects we use to decide whether car A is better than car B. The challenge for interactive evaluation in IR is to connect the two types of evaluation: engine performance and suitability for end-users.

The final motivation behind Mira was to develop measures that will support a broader range of systems such as recommender systems (Resnick and Varian 1997) or novel browsing based systems (e.g. Golovchinsky 1997 and Campbell 2000).

As mentioned above, standard test collections have four components: documents, queries, relevance judgements and a system to evaluate. The remainder of this paper will discuss the assumptions behind each of them. Finally, it will attempt to bring together the main models developed and presented throughout Mira and fit some of the main work in interactive evaluation into this framework.

Documents / Information Resources

In traditional test collections there has been a clear understanding of what a document is: a single, isolated and independent piece of text. Success of a system has been measured in terms of how many relevant documents have been returned. This is not, however, how many users think of retrieval: for many tasks we seek information, not documents and not all documents contain the same amount of information. When working with hypertext and multimedia collections this problem is compounded by a lack of clear boundaries defining what a document is. For example, in the MPEG-7 initiative on video indexing and retrieval (Martínez 1999) it is unclear how to count relevant documents. Although, IR techniques such as passage retrieval (e.g. Salton et al. 1993) could be used to retrieve appropriate sections of video it is hard to envisage evaluating the success of these techniques without either including users in the evaluation process or introducing some artificial segmentation of the continuous material.

Mizzaro (Mizzaro 1998) presents a four dimensional model of relevance. His work became one of the main foci for Mira and it summarises many of the issues involved in interactive evaluation. In this paper, I will

cover the dimensions separately and bring them together towards the end of the paper. The first dimension covered here concerns the depth at which information resources are represented, and at which subsequent relevance judgements are made. The three levels identified are:

- *surrogate*: "a representation of a document, consisting of one or more of the following: title, list of keywords, author(s) name(s), bibliographic data, abstract, and so on";
- *document*: "the physical entity that the user of an IR system will obtain after his seeking of information";
- *information*: "the (not physical) entity that the user receives/creates when reading a document".

Queries / Representation of the User's Problem

The second dimension of Mizzaro's framework concerns the representation of the user's problem, or information need, and the various representations it can take. Traditional IR collections include one of two types of queries:

- *query*: "a representation of the information need in a 'system' language, for instance boolean" constructions or sets of keywords.
- *request*: "a representation of the information need of the user in a 'human' language, usually in natural language."

On top of these there is the *perceived information need*: the user's internal model of his/her need at the time (s)he writes the query. Clearly, in a task context, this can never be fully externalised as there will be unstated aspects of the information need that will drive the user's interaction and selection of material. For example, in work with the Boeing company, Fidel and Crandall (1997) highlighted that participants would select/exclude documents on grounds such as:

- "it confirmed or validated what the participant already knew";
- "the participant had no influence over the issues the report raised";
- "it was about something Boeing was already doing".

In a wider sense it is arguable that there also exists a *real information need*. Many times we satisfy our perceived need only to discover that we were looking for the wrong thing, or with hindsight we realise we were searching for non-relevant material. Not only can this real information need not be externalised, it cannot be fully realised by the user at query time.

The use of users with real information needs is, however, problematic for much research: to get at such needs the search and retrieval tasks have to be embedded within a user's working life. This implies long-term, workplace studies with associated costs, analysis and repeatability problems. Borlund (2000) is investigating how well imposed information needs stimulate behaviour equivalent to that of real information needs. Clearly the real information need cannot be transferred from one user to another. However, through simulated work tasks Borlund claims that perceived information needs can act as suitable predictors of performance of systems with real users and their needs. In line with Borlund's approach, Jose, Furner and Harper (1998) carried out usability experiments on an image retrieval system. Their experimental condition compared two versions of an image search system (one using spatial information only and one using text queries only). Users were given the simulated work task of finding images to illustrate a tourist board leaflet, with their conclusions being mostly compared on user satisfaction after performing the tasks rather than on number of relevant images found.

Relevance Judgements / Depth of Context

Based on a very extensive review of what is meant by relevance (Mizzaro 1997), for his third dimension, Mizzaro summarises the main themes of the discussion into three

categories of relevance (or the *components* dimension in his papers):

- **"Topic**: that refers to the subject area to which the user is interested. For example, 'the concept of relevance in information science';
- **Task**: that refers to the activity that the user will execute with the retrieved documents. For example, 'to write a survey paper on ...';
- **Context:** that includes everything not pertaining to topic and task, but however affecting the way the search takes place and the evaluation of results. For example, documents already known by the user (and thus not worth being retrieved), time and/or money available for the search, and so on."

Traditional IR evaluation concerns itself almost exclusively with topical relevance. This has been successful and given fairly consistent results, in terms of ranking of systems (Voorhees 1998). However, as mentioned above, topic is not the only reason for judging relevance and for many non-textual collections is unlikely to be even an important reason.

Reid (2000) has developed a variation of the test collection approach that is based on users selecting relevant documents based on a real world task. Once the judgements have been made and refined in a task-based process, it is possible to compile the relevance judgements into a test collection in much the same way that traditional relevance judgements are used.

The majority of test collections have also, for simplicity, viewed relevance judgements as binary: a document is relevant or not. This simplification is helpful to evaluators: forcing them to make binary judgements can help focus their attention. However, some documents will be more relevant than others: either because they contain more relevant information or because the information they contain is highly relevant. For topical relevance, it is also possible to measure percentage agreement of a community on a document's relevance to get an impression of how many users would consider this document relevant to this query. Both these observations lead to the conclusion that it might be more reflective of real-world IR to model both strength of relevance and universality of agreement. (Amati and Crestani in-press; Spink, A., Greisdorf, H., and Bateman 1998; Denos and Berti 1998)

The System

In Draper and Dunlop (1997), we argued that the notion of "system" used in IR experiments is too restrictive and that, as in HCI work, the user should be considered as part of the system. Another major focus for Mira was on a framework for evaluation introduced by Annelise Mark Pejtersen (Pejtersen 1996). It focuses on the levels at which we can evaluate a "system" and is summed up in figure 1. The layers are as follows:

1. **Underlying engine**: evaluation of the performance of the underlying engine (*à la* Cranfield / TREC).[2]
2. **Surface interaction**: Evaluation of the immediate interaction between a user and a computer: essentially measuring the low level match between users' sensori-motor characteristics and the system design. This form of interaction is a classic low level usability evaluation and will highlight problems such as users not understanding icons and not being able to manipulate interface features correctly.
3. **Strategy support**: Does the system support all retrieval strategies? In terms of IR, search strategies include analytical, similarity, browsing, empirical and bibliographical.

[2] I have taken the liberty here of introducing a new central ring representing the performance of the underlying search engine for the purposes of this paper.

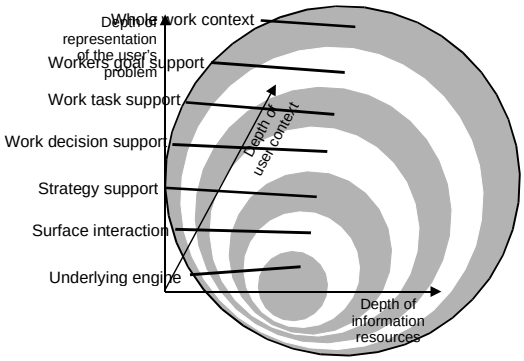

Figure 2: First three of Mizzaro's dimensions

4. **Work decision support**: Does the system support relevant decision tasks? Is the user supported in exploration, situational analysis, goal evaluation and planning for familiar as well as unfamiliar situations?
5. **Work task support**: Does the system support the entire task repertoire? Many IR systems are designed purely to support the task of finding relevant documents, this is rarely the user's end goal and the integration of the IR system with other systems could have a major impact on the success of an IR tool.
6. **Workers' goals**: People have goals within their work life in general (e.g. get *n* publications per year, get promotion etc.). Outside of IR, failure of systems to support these goals, or worse, clashing with these goals, have led to significant failures when new software is introduced. In particular, failure to support personal goals can lead to major problems with collaborative systems.
7. **Whole context**: Company policies and organisational attitude can also have major impact on what software designs work and fail. For example, company policies on sexual harassment can have direct impact on search engine design of pornography filters.

The inner interface layers are typically evaluated in HCI using techniques such as usability experiments, think alouds and cognitive walkthroughs. Usability experiments, much like the TREC interactive experiments, are designed to measure performance of a group of users on system A versus system B. Performance is normally measured in terms of time to complete a set of tasks, number of tasks completed in a set time or number of errors made. Think alouds are typically based around a smaller set of users carrying out tasks while verbalising their thoughts. This approach doesn't often result in numerical data that can be directly compared, but can give great insights into interface design problems (particularly at the inner-most usability layers). Cognitive walkthroughs also attempt to highlight usability problems, rather than collect numbers, and often reveal problems at a higher level than think alouds. The walkthroughs are based around the system designers working through agreed scenarios answering a set of questions on the interface design as they proceed. When evaluating the outer levels of the Pejtersen model, techniques derived from ethnography are typically used to study real people over a long period of time in their actual workplaces. (See most texts on Human Computer Interaction, e.g. Preece et al 1994, for more details.)

Towards a framework

Clearly not all IR work can be evaluated in terms of the full framework described above. However, realising that considerable amount of work in the field of human computer interaction has carried out evaluation in these areas could inspire development and evaluation of IR systems that better support work tasks.

The three dimensions of Mizzaro's model that have been discussed so far can be presented as a 3D space with each dimension representing more in-depth knowledge (figure 2 - although Mizzaro represents the context dimension as a partial order, this simplification is useful here). These dimensions could be mapped onto Pejtersen framework as three of the many dimensions that deepen as more of the environment of evaluation is taken into account (figure 3).

The fourth dimension in Mizzaro's work is time: our view on relevance changes over time and these changes can happen within the space of one interactive session as we learn more about the topic we are researching. As such, we need to consider evaluations that take into account changing views of users (figure 4) and their *information problem shift* (Spink 1999).

The main TREC initiative evaluates only the central circle of figure 3, with no notion of time. Although having done that gives benefits of direct comparison between systems and groups which is missing as soon as outer framework layers are considered.

Some work reported at Mira attempted to bring aspects of the wider layers into the test collection framework: most noticeably the interactive track of TREC (Lagergren and Over 1998). It has focused on assessing the interaction of a user with a system and has tackled many of the problems in designing interactive evaluations that can be compared across sites, a fundamental aspect of traditional IR evaluation, and make use of extensive test collections. Also work by Reid (2000) and Dunlop (1997) took different approaches to widening the scope of test collections. As discussed above, Reid is developing task-based test collections while Dunlop's work introduces aspects of the surface interaction into the test collection methodology. This leads to models of evaluation based on estimating time to complete a task and using test collections as a means for estimating how many documents a user will have to examine.

Mira also showed strong interest in two pieces of work focussing on performing interactive evaluation in a lightweight manner. Harper's *evaluation light* focussed on running small targeted experiments to help in the design of new systems (as opposed to large scale full-understanding studies) (Harper and Hendry 1996). Green (1996) introduced Mira to the notion of *discussion tools*: a set of tools or approaches to challenging design decisions and approaches through discussion among designers. While not giving the depth of understanding of end users, these tools are invaluable for initial hypothesis testing and

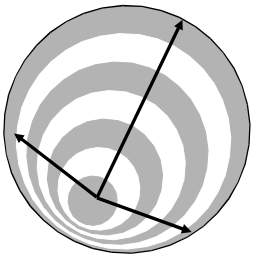

Figure 3: Representing Mizzaro's dimensions on Pejtersen's framework

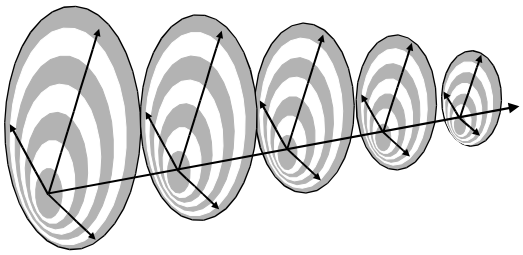

Figure 4: Adding an explicit notion of time to the framework.

design.

Sormunen, Markkula and Järvelin (1999) worked closely with journalists in an investigation into how they judge similarity of photos. Again within the journalist domain, Macaulay (2000) reported her work on a long-term study of journalists in the newsroom. While these investigations are clearly time consuming, they include more of the model in figure 4 and can lead to deep findings about how people actually work. For example, the notion of *trust* is very strong with journalists and they often have problems using general web search engines because they are unsure how much they can trust sources on the web: no matter how "relevant" a story may be to a newspaper's readers - if the source is untrustworthy, it is useless.

Concluding remarks

Usability evaluation has approached the assessment of users and their interaction with computers from many different angles: ranging from laboratory experiments, through co-operative evaluation on simulated tasks to long-term workplace studies of systems in use. Adopting these techniques within IR, in parallel to the traditional test collection approaches, should lead to a wider range and style of evaluation work. This larger palette of techniques should, in turn, lead to different avenues of research and design being followed and to better IR systems as a whole.

Acknowledgements

Thanks are due to Mark Sanderson (Sheffield) and Stephen Draper (Glasgow) for their extensive comments on drafts of this paper.

The Mira working group was funded by the EU information technologies programme ESPRIT (ref: 20039) and led by Keith van Rijsbergen, Steve Draper and myself. The success of the working group was mainly down to the organisers of the meetings: Catherine Berrut, Yves Chiaramella, Marion Crehange, Fabio Crestani, Norbert Fuhr, Silvia Gabrielli, Jane Kernan, Ian Ruthven, Brigitte Simonnot, and Alan Smeaton, plus our many helpers.

The members of Mira were: Glasgow University (prime contractor), City University in London, Dortmund University, Dublin City University, GMD Institute IPSI, IEI-CNR Pisa, Joseph Fourier University in Grenoble, Nancy University, Padova University, Royal School of Librarianship in Copenhagen, Robert Gordon University in Aberdeen, Union Bank of Switzerland, Swiss Federal Institute of Technology Zurich, and Tampere University. In addition there were many participants from outside the working group who attended one or more meetings. In particular thanks are due to James Allan, Bill Arms, Nick Belkin, Giorgio Brajnik, Raya Fidel, Gene Golovchinsky, Thomas Green, Lynda Hardman, Rob Koenen, Catriona Macauley, Wendy Mackay, Stefano Mizzaro, Desai Narimsahalu, Annelise Mark Pejtersen, Amanda Spink and Ross Wilkinson for their contributions to the Mira meetings.

More information, including workshop proceedings, can be found on the Mira website: http://www.dcs.gla.ac.uk/mira/

References to Mira presented material

Amati, G., and Crestani, F., "Probabilistic Learning by Uncertainty Sampling with Non-Binary Relevance". In: F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval: techniques and applications*. Physica Verlag, Heidelberg, Germany, in-press.

Borlund, P., "Experimental components for the evaluation of interactive information retrieval systems", *Journal of Documentation*, 56(1), pp 71-90, January 2000.

Campbell, I., "Interactive Evaluation of the Ostensive Model Using a New Test Collection of Images with Multiple Relevance Assessments", *Information Retrieval*, 2(1) pp. 89-116, February 2000.

Denos, N., and Berti, L., "Document quality criteria for Collaborative IR", slides from presentation at Mira Dublin, www.dcs.gla.ac.uk/mira/workshops/dublin/ October 1998.

Draper, S. W., and Dunlop, M.D., "New IR - New Evaluation: The impact of interactive multimedia on information retrieval and its evaluation", *The New Review of Hypermedia and Multimedia*, 3, pp 107-122, 1997.

Draper, S.W., Dunlop, M.D., Ruthven, I., and Van Rijsbergen, C.J., (Editors), *Proceedings of Mira 99: Evaluating Interactive Information Retrieval (Glasgow April 1999)*, published in *Electronic Workshops in Computing*, April 1999.

Dunlop, M.D., "Time Relevance and Interaction Modelling for Information Retrieval", *Proceedings of SIGIR-97, Philadelphia, USA*, pp 206-213, ACM Press, 1997.

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

Golovchinsky, G., "What the query told the link: The integration of hypertext and information retrieval", *Proceedings of Hypertext '97 (Southampton, UK)*, ACM Press, pp. 67-74, April 1997.

Green, T.R.G. and Petre, M., "Usability analysis of visual programming environments: a 'cognitive dimensions' framework", *Journal of Visual Languages and Computing*, 7, 131-174, 1996.

Harper, D.J., "Information Retrieval Evaluation: Lifebelt, Security Blanket or Straitjacket?", abstract in Smeaton 1997.

Harper, D.J., and Hendry, D., *Evaluation Light*, Section 4.2 of Dunlop 1996.

Jose, J.M., Furner, J., and Harper, D.J., "Spatial querying for image retrieval: a user-oriented evaluation", *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne Australia*, pp 232 - 240, ACM Press, August 1998.

Lagergren, E., and Over, P., "Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment", *Proceedings of SIGIR 98, Melbourne Australia*, pp 164-172, ACM Press, 1998.

Macaulay, C., Benyon, D. and Crerar, C. (in press) Ethnography, Theory and Design: 'Uncovering' New Artefacts. To appear in *International Journal of Human-Computer Studies*, 2000.

Martínez, J.M., *Overview of the MPEG-7 Standard* ISO/IEC JTC1/SC29/WG11 N3158 report, http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm, December 1999.

Mizzaro, S., "Relevance: the whole history", *Journal of the American Society for Information Science* , 48(9), pp.810-832, 1997.

Mizzaro, S., "How many relevances in information retrieval?" *Interacting With Computers* , 10(3), pp.305-322, 1998.

Pejtersen, A.M., "Empirical Work Place Evaluation of Complex Systems", *Advances in Applied Ergonomics. Proceedings of the 1st International Conference on Applied Ergonomics. (ICAE'96), Istanbul, Turkey*, USA Publishing Coorporation, (version in Dunlop 1996), May 1996.

Reid, J., "A Task-Oriented Non-Interactive Evaluation Methodology for Information Retrieval Systems", *Information Retrieval*, 2(1) pp. 115-129, February 2000.

Robertson, S., "Process and outcome: On the evaluation of IR systems in the age of interaction, GUIs and multimedia.", in Draper et al 1999.

Smeaton, A.F., *Report on Schloss-Dagstuhl Meeting on 'Evaluation of Multimedia Information Retrieval'*, www.dcs.gla.ac.uk/mira/workshops/dagstuhl/ , 1997.

Sormunen, E., Markkula, M., and Järvelin, K., "The Perceived Similarity of Photos - A Test-Collection Based Evaluation Framework for the Content-Based Image Retrieval Algorithms", in Draper et al 1999.

Spink, A., “Towards a theoretical framework for information retrieval evaluation in an information seeking context”, in Draper et al, 1999.

References to other cited work

Buckley, C., Mitra, M., Walz, J., & Cardie, C. “Using clustering and superconcepts within SMART: TREC-6. In E. Voorhees, & D. Harman, *Proceedings of the sixth Text REtrieval Conference (TREC-6)* (pp. 107±124), NIST Special Publication 500-240. 1998.

Fidel, R., and Crandall, M., “Users’ perception of the performance of a filtering system”, *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia, USA*, pp 198 - 205, July 1997.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., and Carey T., *Human-Computer Interaction*, Addison-Wesley, 1994.

Resnik, P., and Varian, H.R., “Recommender Systems”, *Communications of the ACM* 40 (3) 1997.

Salton G., Allan J. and Buckley C., “Approaches to Passage Retrieval in Full Text Information Systems”, *Proceedings of ACM SIGIR 1993, Pittsburgh USA*, ACM Press, June 1993.

Sparck Jones, K., and Willett, P., *Readings in Information Retrieval*, Morgan Kaufmann, 1997.

Spink, A., Greisdorf, H., and Bateman, J., “From highly relevant to not relevant: Examining different regions of relevance”, *Information Processing and Management*, 34(5), 599-622, 1998.

Van Rijsbergen, C.J., *Information Retrieval*, Butterworths, 1979.

Voorhees, E.M., “The Sixth Text REtrieval Conference”, introduction to special issue of *Information Processing and Management* on TREC-6, vol 36(1), pp 1-2, 2000.

Voorhees, E.M., “Variations in relevance judgements and measures of retrieval effectiveness”, *Proceedings of SIGIR 98, Melbourne Australia*, pp 315-323, ACM Press, 1998.

Voorhees, E.M., and Harman, D., “Overview of the Sixth Text Retrieval Conference (TREC-6)”, *Information Processing and Management*, vol 36(1), pp 3-35, 2000.