BCS

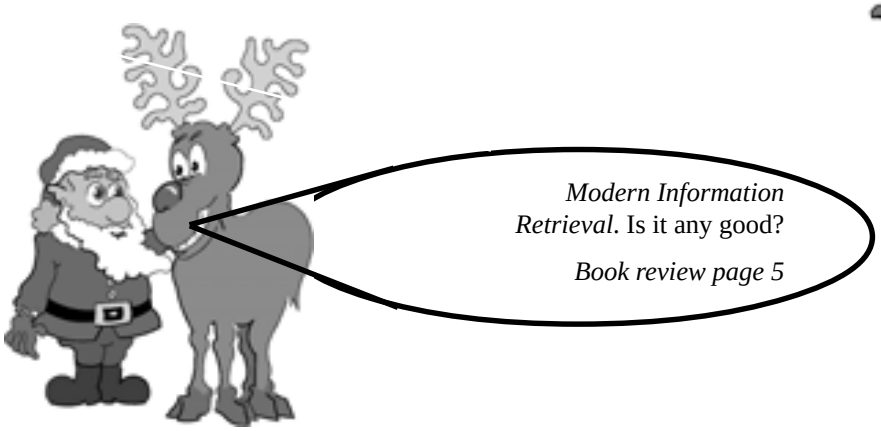
INFORMER

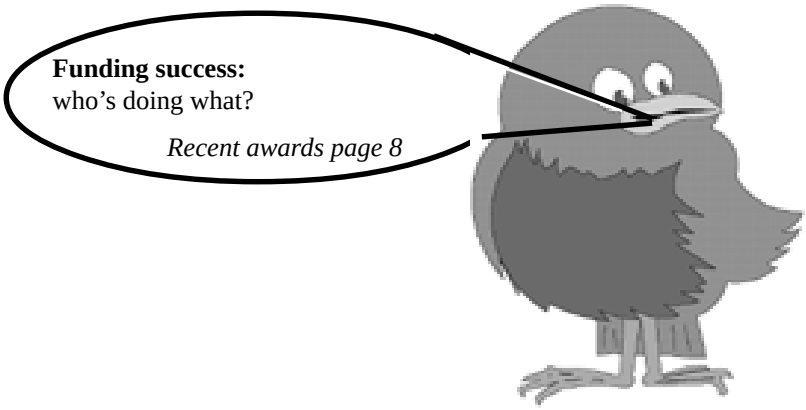
CEPIS

British Computer Society
Information Retrieval Specialist Group

Volume 9 | Winter 1999 | ISSN 0950-4974

Merry Christmas Everybody!





Who's who

IRSG Committee Contact List 1999 - 2000

Wondering who you should contact about what? Well, here's the current list of contacts.

Chair
Jan J IJdens
Sharp Laboratories of Europe
Oxford Science Park
Oxford OX4 4GA
Email: jan@sharp.co.uk (or chair.irsg @bcs.org.uk)

Vice-Chair
Mounia Lalmas
Department of Computing Science
Queen Mary and Westfield College
University of London
Email: mounia@dcs.qmw.ac.uk

Secretary
Jane Reid
Department of Computing Science
Queen Mary and Westfield College
University of London
Email: jane@dcs.qmw.ac.uk

Treasurer
Margaret Graham
Institute for Image Data Research
University of Northumbria
at Newcastle
Email: margaret.graham@unn.ac.uk

Colloquium 2000
Stephen Robertson
Microsoft Research Ltd
St George House, 1 Guildhall Street,
Cambridge CB2 3NH, U.K.
Email: ser@microsoft.com

One-day events
Mounia Lalmas
as above

Ordinary members
John Davies
Information Access Research.
BT Laboratories
Email: john.davies@bt-sys.bt.co.uk

Ayse S Goker-Arslan
School of Computer and Mathematical Sciences
The Robert Gordon University
Email: asga@scms.rgu.ac.uk

David Harper
School of Computer and Mathematical Sciences
The Robert Gordon University
Email: d.harper@scms.rgu.ac.uk

Monica Landoni
Department of Information Science
University of Strathclyde
Email: monica@dis.strath.ac.uk

John Lindsay
School of Information Systems
Kingston University
Email: lindsay@kingston.ac.uk

Andrew MacFarlane
Department of Information Science
City University
Email: andym@soi.city.ac.uk

Kerry Rodden
Computer Laboratory
University of Cambridge
Email: Kerry.Rodden@cl.cam.ac.uk

Tony Rose
Canon Research Centre Europe Ltd
Guildford
Email: tgr@cre.canon.co.uk

Steve Wade
School of Computing and Mathematics
University of Huddersfield
Email: s.j.wade@hud.ac.uk

Informer Team
Jon Ritchie
Department of Computing Science
University of Glasgow
Email: jon@dcs.gla.ac.uk

Ian Ruthven
Department of Computing Science
University of Glasgow
Email: igr@dcs.gla.ac.uk

CIA-2000

Submission: January 24, 2000
Notification: March 10, 2000

Information agent technology is one of the major key technologies for the Internet and worldwide Web. An information agent is a computational software entity that has access to one or multiple, heterogeneous and distributed information sources, pro-actively searches for and maintains relevant information on behalf of users or other agents preferably just-in-time. In other words, it is managing and overcoming the difficulties associated with information overload in the open and exponentially growing Internet and Web.

Although low-level infrastructure has been developed to support interoperability between heterogeneous databases and application programs, this is not sufficient when dealing with higher-level object organizations such as vertical business object frameworks and workflows. Existing multi-database or federated database systems do not support any kind of pro-active information discovery. One key challenge of advanced information systems is to balance the autonomy of databases and legacy systems with the potential payoff of leveraging them by the use of information agents to perform collaborative work.

Development of information agents requires expertise from different research disciplines such as Artificial Intelligence (AI), advanced databases and knowledge base systems, distributed information systems, adaptive information retrieval, and Human Computer Interaction (HCI).

Submission details at the conference homepage : *http:// www.dfki.de/~klusch/cia2000.html*

Workshop report

Searching for information workshop

Glasgow 11-12th November 1999

This event, organised by Mounia Lalmas (Queen Mary and Westfield), Alison Cawsey (Heriot-Watt), and Keith van Rijsbergen (Glasgow), was run by the IEE Informatics Professional Group A4 (Artificial intelligence) in association with the IRSG. The main aim of the workshop was to examine, common ground and complementary goals and techniques in Artificial intelligence and Information retrieval for retrieving relevant information.

This is a personal view of the two day-workshop, an official summary of the event written by the workshop chairs can be found at *http://www.iee.org.uk/Flashpoint/*. The titles given here refer to the titles of the talks, not to the titles of the abstracts and papers published in digest form by the IEE. Any other discrepancies, errors or misrepresentations are, of course, mine.

Yorick Wilkes, University of Sheffield, started the event with a look at *'IR and AI traditions of representation and anti-representation'*, differentiating Information Extraction (IE) from traditional IR and AI approaches, although highlighting the connection to both. The particular feature that characterises IE is one of linguistic rules and templates to index or categorise text. Machine learning is present both to assign documents to templates but also to generate templates. Wilkes also discussed the nature of representation in IR and IE.

Dieter Fensel, of the Vrije Universiteit, Amsterdam, called for standards for representing complex information in *'Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce'* (with Frank van Harmelen). Fensel's survey concentrated on representation of web objects with particular interest on management and e-commerce. He then proposed ontologies - a shared, formal, common representation of a domain - as a solution to the variety of possible representation techniques. Unlike, say, XML which only standardises a syntax, this approach, it is claimed can standardise vocabulary and structure.

John Eakins, of the Institute of Image Data Research, University of Northumbria at Newcastle, gave an excellent analysis of the use of AI techniques for Content-based Image Retrieval (CBIR) in *'How smart are current image retrieval techniques?'*. He set the scene by separating approaches to CBIR into three levels: retrieval by primitive features (colour, texture, spatial location), retrieval by logical or derived feature (incorporating aspects such as scene analysis or object recognition) and retrieval by abstract attributes, which involves more complex reasoning about the semantic content of images.

Eakins then compared each of these three layers against a set of criteria for intelligence; use of reasoning ability, heuristics, learning, higher order knowledge, and ability to match output of human experts. The bad news is that at his first level, retrieval by primitive features where currently most CBIR techniques are operating, few approaches stand up to any of the criteria. Little or no CBIR research was discussed for the third - conceptual/ abstract - level, the issues involved being 'dauntingly complex'. We concur.

The better news is that in the second level - retrieval by derived features - there is evidence of more exploitation of AI techniques. He differentiated between two types of approach - wholly automatic scene or object recognition and a semi-automatic feedback-driven approach. The automatic approaches show significant use of high-level knowledge, heuristics and some ability to reason but with little evidence of learning and a domain-dependent ability to match human judgements. In the semi-automatic approaches there is strong use of learning and reasoning approaches but little use of high-level knowledge and generally poor match to human experts.

The general conclusion was that significant advances to CBIR research can be made using AI techniques but that semantics is an AI problem not an image processing problem *per se*. Hard problems will require sophisticated solutions but these have the potential to generate a lot of useful collaboration.

In *'Fetch me a picture representing triumph or similar: classification based navigation and retrieval for picture archives'*, **Carol Goble** (with Sean Bechhofer) of the University of Manchester, came from a stronger AI direction than some of the other speakers. This talk also presented research based on ontological representations of knowledge, in this case for image retrieval. The system presented by Goble is powered by a description logic which, in conjunction with a controlled vocabulary, and an ontology describing a domain, can create and automatically classify semantic descriptions of images. Queries are also represented in this logic (the translation to DL representation being mediated at the interface). This talk showed how domain knowledge can creatively combine with information representation but also hinted at the difficulty of producing these representations.

Theo Huibers, KPMG, The Netherlands (with Bernd van Linder, Utrecht University), presented a formal approach to a more recent problem in IR, distributed information retrieval. In *'Intelligent information retrieval*

agents', he outlined an approach (based on the logical model of IR) for axiomatising IR systems - representing the characteristics of a particular retrieval model as a series of rules together with a derivation system (to infer whether a document is 'about' a query). Different rules and different derivation systems deliver models of different IR systems. The axiomisation of IR systems in this fashion, it is claimed, allows the evaluation of effectiveness of systems without the need for experimentation (or at least in addition to experimentation).

Instead of running only one IR system, Huibers proposed running a series of agents, each instantiating a different IR model. These *retriever* agents are supplemented by *user* agents, who make presentation decisions based on the retrieval results returned by the retrieval agents. Modal logic was described as a suitable framework to theoretically unify the different types of agents.

The role of personal preferences and individualised searching was introduced by **Marc Moens** of the University of Edinburgh, in *'Personalised information objects'*. This falls into two areas. Firstly being able to create personal collections of resources. Secondly, and more interestingly, was the concept of personalised delivery of information; the content of web pages is not regarded as fixed entity but as something to be manipulated for presentation. This manipulation would be as the result of e.g. existing information about the user or previous interactions with a set of pages. Information can then, for example, become relative to what has already been seen.

Karen Sparck Jones, Computer Laboratory, University of Cambridge, threw down at least one gauntlet in *'IR Lessons for AI'*. The first is that AI has misunderstood the task of IR! IR and AI certainly share a general common goal, that of finding information; conventional AI approaches (tends) towards more explicit, formal, representations of knowledge and knowledge manipulation, IR (tends) towards weaker, approximate statistical approaches. This latter approach can be represented as an anti-classical AI approach with the emphasis on implicit meaning. The reason standard IR techniques, such as term weighting and relevance feedback, are important is not because they are supplementing a rigorous analysis of meaning to be found in AI techniques but because what IR systems are supposed to be doing is underspecified.

Sparck Jones's assertion, in my view, is that most IR situations (or information seeking situations) are poorly defined in terms of expected outcome; the task is not that of finding specific answers to questions but often that of finding general information that may be of interest. What IR is doing correctly is implicitly exploiting aspects of natural language use such as redundancy, coherence, and associative relations to provide 'sufficient-to-the-day' solutions. Detailed classical AI approaches may actually be inferior to IR models in that they are not generalisable to a wide enough range of information seeking situations.

The second challenge set down to the AI community at large was to do better than IR retrieval baselines. She compared sets of retrieval results from the same query set and collection, generated using, successively, no term weight, term weighting, blind feedback and relevance weights to show how the (by now) standard IR statistical approaches can massively increase retrieval effectiveness. These baseline figures indicated what level of results IR techniques can achieve in the current environment of large, relatively diverse collections, and with no associated knowledge modelling. The challenge to critics in the AI community, of course, was - can you do better?

'*Reinforcement learning for information seeking'* by **Susan Craw**, The Robert Gordon University, presents a machine learning approach to predicting good paths through a web site using information from previous paths. This may be seen as a form of collaborative filtering without the need for user feedback; overlapping paths and repeated accesses to the same pages indicating shared interests. Craw uses machine learning techniques to maximise an overall reward in following a particular path. This need not only benefit individual users but may be used in the design process of web sites; using routes through a site to direct design of useful sites. Different algorithms for learning good paths were outlined and discussed, as well as the particular nature of the learning problem

Gianni Amati, of the Fondazione Ugo Bordoni, Rome, and currently at the University of Glasgow, presented a machine learning interpretation of relevance feedback in *'Learning from examples as relevance feedback and relevance feedback as learning by examples'*. He uses this analysis to show how machine learning techniques can explain why sometimes the standard probabilistic relevance feedback approach can make poor choices. This results in particular from the fact that relevance feedback must operate with very few positive cases (relevant documents) and a large number of negative cases (non-relevant document). An alternative is suggested based on the informative power of terms.

Keith van Rijsbergen, University of Glasgow, introduced a new paradigm for IR based on '*Quantum logics and IR*'. The big idea is to develop a 'logic of interaction', one that is query-driven and usage- and context-sensitive but has less emphasis on representation. One promising feature of this model, even though it is very new work, is the potential to unify the main features of the three majors models - vector space, probabilistic and

logical - into one algebraic model.

The final paper was by **Stephen Robertson** of Microsoft and City University. *'Probabilistic retrieval: thresholding for automatic filtering'*, differentiated between two tasks in filtering: scoring (determining the likelihood of a documents relevance) and, the main content of the talk, deciding which documents to retrieve (thresholding). Set a threshold too high and the system will retrieve too few documents, set the threshold too low and it will retrieve too many documents. In the later stages of a filtering profile (after weeks, months, or years) there will be an accumulation of relevance information which can be treated by techniques such as iterative optimisation. However, in the initial stages of a profile there will be generally very little relevance information. Consequently, Robertson argues, we need a stronger model to make good thresholding decisions. The proposed solution is to turn the document score into an absolute value of the probability of relevance (as opposed to the variations in scoring techniques contained within the major probabilistic implementations such as Okapi) by a procedure such as logistic regression.

This event was relatively rich in content, ranging from very specific to very general discussions and the organisers are to be congratulated on producing such a varied programme of speakers. The workshop itself could have been better attended, perhaps the poor turn-out being a result of the IEE's pricing policy of which there was much discussion and criticism.

The overlap between the two areas has the potential to generate rich collaboration between the two groups. Hopefully, future events to this workshop will be forthcoming.

Ian Ruthven

University of Glasgow

Book review

Modern Information Retrieval. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Addison Wesley. 1999. ISBN 0-201-39829-X. £29.95. 513pp. Softbound.

'What's a good Information Retrieval book?' is a question often asked. The answer has never been simple: for many years, the choice was limited to Salton and McGill's great book from 1983 or Van Rijsbergen's classic from 1979. Both had their strengths and weaknesses and the recommendation of one or the other would depend on what the questioner wanted to know about. Given the age of these two books, there has been an increasing need for a replacement and recently a great many books have come along trying to take the place of these two volumes. *Modern Information Retrieval* is one such. Ricardo Baeza-Yates and Berthier Ribeiro-Neto with the help of Gonzalo Navarro and Nivio Ziviani wrote nine of its core chapters. The other six were contributed by Elisa Bertino, Eric Brown, Barbara Catania, Christos Faloutsos, Elena Ferrari, Ed Fox, Marti Hearst, Edie Rasmussen, and Ohm Sornil. Together, the fifteen chapters cover the more technical and mechanical sides of Information Retrieval. Weighting schemes, automatic query expansion, index file construction and compression, matching of multimedia data are some of the 'hard' IR topics covered here. Those looking for a presentation of the 'softer' side of IR, such as user models of information seeking, will need to look elsewhere.

Overall, I feel that the book covers information retrieval well. But I have reservations, which I will come to as I discuss the first three chapters.

Chapter 1 provides both an introduction to the book and to the retrieval process as well. The IR introduction is, however, very basic. Rather than write an 'Introduction to IR' chapter---quickly covering such things as stop word removal, stemming, simple ranking, index files, etc.---the authors choose to cover these topics within more detailed chapters later in the book. For example, stemming and stop word removal is first covered in the 'Text Operations' chapter (**Chapter 7**) and indexing is described in **Chapter 8**, which covers a wide range of text searching and matching techniques. The consequence of choosing to structure the book in this manner is the presence of a great many forward references to later chapters in the earlier sections of the book. There is therefore a danger that absolute beginners to IR may find it hard to get going with this book if they read it from cover to cover. With some directed reading, however, a basic introduction can probably be gleaned from bits of the chapters. Those teaching a course based on this book would have to provide the direction themselves.

Throwing the novice into the deep end, **Chapter 2** dives straight into the topic of models of IR. The classic models of retrieval are there: Boolean, vector space, and probabilistic. This is followed by a description of more sophisticated versions of these three, and the chapter then moves onto latent semantic indexing, neural networks, and an interesting discussion of retrieval from structured documents. One of the problems with this and many of the other chapters is that statements are often not obviously backed up with references. The reason is that the bulk of references often appear in a bibliographic discussion section at the end of each chapter. This format is common throughout the book, which has scant use of references in the body of a chapter (or in the case

of Chapter 4, no use at all), followed by a bibliographic discussion section, where they appear in great numbers. There are 852 references in this book, which makes it a great resource; but detaching the references from the descriptions of work detracts from the readability and usability of the book.

Continuing with the unusual ordering of subjects, **Chapter 3** covers the evaluation of retrieval systems. For me, this was the weakest part of the book. About half of it discusses retrieval effectiveness measures such as recall and precision, but these are briskly covered without going into much detail. To make matters worse, there is a mistake in the interpolation formula on page 77: 'j+1' should read 'max' (I'm told this will be fixed in later print runs). The rest of the chapter is much less interesting, discussing trivial aspects of retrieval test collections and the annual IR evaluation conference, TREC. This part reads like the authors were trying to 'pad out' the chapter to make it look longer. To illustrate, on page 85 the authors write

'The first TREC conference was held at NIST in November 1992, while the second TREC conference occurred in August 1993. In November 1997, the sixth TREC conference was held (also at NIST) and counted the following participating organizations [sic]...'

The book then goes onto list the names of the 47 organisations that were at TREC that year. Not even hard-core TREC groupies are going to find this interesting. When one considers the wide range of evaluation issues that could have been discussed in a chapter like this (where was interactive user evaluation, or consistency of relevance judgements for example?), this second half feels a real waste of time and is the reason why this chapter is the low point of the book. But things start to get better.

Chapter 4 (with Navarro) discusses query languages, such as Boolean query formats, regular expression pattern matching, queries on structured documents, and certain client-server protocols such as Z39.50. As already mentioned, the only thing that lets this chapter down is that all references are shunted to the end of the chapter into the bibliographic section. **Chapter 5** discusses means of query expansion, through such techniques as relevance feedback, Xu and Croft's local context analysis, or Qiu and Frei's similarity thesaurus. **Chapter 6** contains an odd mix of text, image, sound, 3D, and video format descriptions, followed by a discussion of Shannon's information theory (which, bizarrely, is not referenced), and some text modelling issues. Although the authors make it clear that knowledge of these issues is important, I found the composition of this chapter strange.

Chapter 7 (with Ziviani) on text operations describes various rudimentary IR text-based techniques, such as stemming and stop-word removal, followed by a brief mention of document clustering and then text compression. The section on compression is well explained, covering both the processing of collections and of postings files. This last topic illustrates again the unusual ordering of the subjects in the book; it is not until the next chapter that the structure of index files is explained. One significant omission in the chapter is the work on 'lossy compression' of index files, typified by Michael Persin's ACM SIGIR 1994 paper *Document filtering for fast ranking* or Smeaton, Kelledy, and O'Donnell's paper in TREC-4 held at NIST in November 1995.

Finally, in **Chapter 8** (with Navarro) we get an explanation of the structure of index files along with a comprehensive description of other searching methods, e.g.~suffix arrays, signature files, and sequential searching. This section requires a certain level of background knowledge in algorithms and data structures. In places, I found the descriptions of the algorithms heavy going and I would have preferred more examples. Nevertheless, this is a good chapter.

Chapter 9 (by Brown) is a description of the issues of using parallel computing architectures to speed up retrieval. This covers classic parallel schemes such as SIMD and MIMD, but it also covers the more prevalent distributed computing systems. The chapter is comprehensive and well explained.

This can also be said of the following **Chapter** (**10**, by Hearst) on user interfaces and visualisation. I thought this was the best chapter of the book. It covers a wide range of issues relevant to interface design in IR, from support for Boolean query construction to visualisation of document clusters, through to issues of window management and layout of information. Presumably, as a form of promotion for the book, Baeza-Yates and Ribeiro-Neto have decided to put a copy of this chapter on a web site for anyone to download (*sunsite.dcc.uchile.cl/irbook/chapters/chap10.html*). I recommend anyone interested in this area take a look at this. I've already used parts of it in an IR course I teach.

Chapters 11 (by Bertino, Catania and Ferrari) and **12** (by Faloutsos) discuss the storage of multimedia through use of metadata and conceptual structures and the searching of said data using generic indexing approaches. Given the size of the problem of storing and retrieving multimedia data, both these chapters sensibly stick to a very general approach, abstracting away from the specifics of any particular media form. To my mind, this approach works well and the chapters are both readable and understandable. I know little about this subject and so cannot really comment on the comprehensiveness of the chapters---except to say that in both chapters the authors seem to stick a little too much to systems they have produced. **Chapter 12** in particular

reads a too much like a research paper and not enough like a review.

Chapter 13 is the obligatory web searching chapter, which must be tough to write given that much retrieval on the web is being conducted by companies using proprietary systems. Quite how document ranking, porn site filtering, or 'spammed page' detection works will probably never be published by the big web search engine companies. Indeed, in the section describing search engines, it feels as if there is some guessing being done by the authors. As the web is constantly changing, there is also a danger of research results going stale very quickly or whole areas being missed out. To tackle this, Baeza-Yates and Ribeiro-Neto take the novel approach of apologising in advance for any research they may have missed (page 368). Given these problems, however, the chapter reads well, is understandable, and covers most topics in web-based IR with a good number of references.

Chapters 14 (by Rasmussen) and **15** (by Fox and Sornil) discuss physical and digital libraries respectively. Again, these are areas I'm not very knowledgeable in, but both chapters are well written and I found that I learnt a lot from them.

The book ends with a procedural description of Porter's stemming algorithm, a comprehensive glossary, and an index. The inclusion of the stemmer seems a little strange, especially as it seems little different from the description in the original published version. Also, if a stemmer was included, why not a stop word list as well? The authors have also set up a web site (*sunsite.dcc.uchile.cl/irbook/*), which contains online versions of Chapters 1 and 10, bibliographies of the authors, any errata, and a list of online IR resources. Nothing stunning, but it's nice to have.

With this many contributors there may have been a danger of topics being covered twice or a feeling that the chapters are not well integrated. Neither problem arises in this book. It is clear that the main authors went through the contributed chapters carefully and placed copious cross-references to other chapters. I could only find one example of topics not being appropriately linked up.

As for subjects that were missed, text categorisation, filtering, a history of NLP techniques in IR, and the logical models of retrieval all should have been covered to some extent. Apart from the user-related issues mentioned above, the biggest technical subject missing from the book is Cross Language Retrieval. Although there are other publications that cover this topic well, the absence of CLIR is a gap in this book.

Is this the perfect IR book everybody has been waiting for? No, but it's pretty good. Despite the reservations I've expressed, my feeling about this book is that it covers the subject of IR well. Its price (£29.95 in the UK, $40.45 in the US) suggests it is intended for IR courses. For a course composed of students with a computing background it would make a good source text, although some guidance may be required to help students find the introductory material within. I teach a master's course in IR. Some of my students are aiming for a Masters in Arts. This is not the book for them, as it assumes too much technical knowledge. As for me, with its large number of references I suspect that *Modern Information Retrieval* will become my first port of call when seeking further information in IR. I recommend it.

Mark Sanderson
University of Sheffield

The Informer gratefully acknowledges *The Computer Journal* in reproducing this book review.

Recent awards

Evaluation of Content-Based Image Retrieval in an operational setting

Duration: January 2000 - December 2001

Contact: Mrs Margaret Graham. Institute for Image Data Research, University of Northumbria at Newcastle, Newcastle upon Tyne NE1 8ST. margaret.graham@unn.ac.uk. *http://www.unn.ac.uk/iidr/staff/margaret.html*

In recent years there has been enormous growth in interest in the potential of digital images, especially as technological advances make it now possible to store and access large quantities of data relatively cheaply. Coupled to this has been the rapid growth if imaging on the World Wide Web (WWW). Many organisations are taking advantage of various funding opportunities to digitise parts or all of their collections. But the process of digitisation does not in itself make image collections easier to manage or to use. There are several computerised image data management systems on the market which help to organise and view the digital images. Some forms of cataloguing and indexing are still necessary, since browsing is not an option except with small collections. These problems have stimulated research into content-based image retrieval (CBIR), the selection of images from a collection via features automatically extracted from the images themselves. Current CBIR systems

Recent awards contd

typically provide image retrieval by low-level attributes such as colour, texture or shape, and few attempt higher levels of retrieval, such as by semantic content (e.g. the presence in an image of specific types of object, or the depiction of a particular type of event). There has been little systematic evaluation of CBIR system effectiveness on a large scale. Key questions, such as whether CBIR techniques can bring about worthwhile improvements in performance with real-life image retrieval systems, or where and how such techniques can most profitably be used, thus remain unanswered.

The aim of this research is to evaluate CBIR systems in an operational setting. The project will install appropriate commercially-available CBIR software as additional functionality to the image data management systems currently in use in three pictorial libraries in the public and private sectors. An initial user study will be conducted to obtain the first impressions of the CBIR functionality by both staff and other end-users. Following a six months "gestation" period, detailed user evaluations will then be conducted. The outcomes will be three case studies, demonstrating CBIR in practice, and a body of evidence regarding the usefulness and effectiveness of CBIR as a searching tool in the context of the individual organisations. Specific research questions which will be investigated include:

* How successful are CBIR systems in meeting user needs?

* What are the effects of CBIR provision on user search behaviour?

* To what extent can the use of CBIR systems be justified in different contexts?

A system for music information retrieval of encoded music (MuTaTeD 2)

Duration: November 1999 to January 2001

Contact: Ms Carola Boehm. Department of Music, University of Glasgow, 14 University Gardens, Glasgow, G12 8QH. carola@music.gla.ac.uk. *http://www.music.gla.ac.uk/*

The aim of the project is to design and implement a music information retrieval system with delivery/access services for encoded music. The project will build on the experiences and user needs studies of an existing digital library service within the performing arts (PADS, the Performing Arts Data Service) and on the results of MuTaTeD! - a project funded by the Joint Information Systems Committee (JISC) of the UK higher education funding councils - which deals with encoded music information for the web. The prototype system will provide a user-friendly web-based search/browse/query interface to access musical content. It will include:

* provision of an architecture for flexible and controlled access to music content;

* provision of an initial set of powerful music content searching tools;

* provision of a set of user friendly, web-based search/browse/query interfaces to access musical content.

Retrieval through explanation: an abductive inference approach to relevance feedback

Duration: November 1999 - October 2001

Contact: Professor C.J. van Rijsbergen. University of Glasgow, Department of Computing Science, Glasgow, G12 8RZ. keith@dcs.gla.ac.uk. *http://www.dcs.gla.ac.uk/ir/*

Selecting good query terms to represent an information need is difficult. The complexity of verbalising an information need can increase when the need is vague, when the document collection is unfamiliar or when the searcher is inexperienced with information retrieval (IR) systems. It is much easier, however, for a user to assess which documents contain relevant information.

Relevance feedback (RF) techniques make use of this fact to automatically modify a query representation based on the documents a user considers to be relevant. RF has proved to be relatively successful at increasing the effectiveness of retrieval systems in certain types of search, and RF techniques have gradually appeared in operational IR systems and even some Web engines. However, the traditional approaches to RF do not consider the behavioural aspects of information seeking. The standard RF algorithms consider only what documents the user has marked as relevant; they do not consider why the user has assessed relevance. For RF to become an effective support to information seeking it is imperative to develop new models of RF that are capable of incorporating why users make relevance assessments.

The underlying assumption of the vast majority of RF theories is that terms occurring more frequently in relevant documents that non-relevant documents tend to be good for retrieving more relevant documents. However, it has been demonstrated in a number of studies that why users mark documents as relevant is as important as which documents they mark relevant, in deciding what further documents to retrieve. This means, in deciding whether a document is likely to be relevant, we not only have to consider which terms are used in documents: we also have to consider how the terms are used in documents.

In this project we view RF as a process of explanation. A RF theory should provide an explanation of why a document is relevant to an

information need. Such an explanation can be based on how information is used within documents. We will use abductive logic to provide a framework for an explanation-based account of RF. Abductive logic is specifically designed as a technique for generating explanations of complex events, and has been widely used in a range of diagnostic systems. Such a framework will produce a set of possible explanations for why a user marked a number of documents relevant at the current iteration. These explanations will be based on how information is used within relevant documents. From the set of possible explanations, on explanation, known as the best possible explanation, will be selected to reformulate the query. The choice of the best possible explanation is guided by a number of factors, the main factor being the previous search history.

Improved online information access
Duration: November 1999 - October 2001
Contact: Professor Mark Girolami. University of Paisley, High Street, Paisley, PA1 2BE. giro0ci@paisley.ac.uk. *http://cis.paisley.ac.uk/staff/giro-ci0/index.htm*

Since the emergence and explosive growth of the Word Wide Web (WWW) there has been a commensurate growth in the availability of online information. Efficient searching and retrieval of relevant information from the WWW has lagged behind this growth and intelligent information retrieval methods are now required as a matter of urgency. This particular challenge is attracting great interest from the machine learning research community. There are two main approaches to online information retrieval: query based and taxonomic. The query-based approach relies on methods such as search engines which take a user's query and compare it with an existing document collection; the system then returns a list of possible matches ranked in order of relevance. The taxonomic approach relies on manual organisation of the information (online documents) into hierarchical categories. This project postulates that the design of information retrieval systems utilising a number of novel intelligent computing paradigms may provide the key to improved access. It proposes the fusion of recently-developed supervised and unsupervised computational models to adaptively build and maintain suitable document hierarchies and then rank and classify existing as well as incoming new documents based on user queries.

VISOR II - a user-oriented evaluation framework for the development of electronic image retrieval systems in the workplace

Duration of grant: June 2000 - May 2001

Contact: Mrs Margaret Graham. Institute for Image Data Research.

The chief aim of the project is to develop a user-oriented evaluation framework for electronic image retrieval in the workplace. This framework would span the overall process of image retrieval system design, development and implementation - something which includes an ongoing process of evaluation throughout all these stages. It is primarily intended as a practical tool to assist and guide those responsible for conducting evaluations of image retrieval systems from the user's perspective, though there are other potential benefits for researchers, professional groups and organisations as well as users themselves.

This endeavour will utilise and build upon the results of and groundwork laid by the first phase of the VISOR research programme (Information seeking behaviour in image retrieval, LIC project LIC/RE/031). Thus the models of user search behaviour focus on users of images in context and this will lie at the heart of the development of an evaluation tool.

A central focus of the evaluation framework will be the consideration of how well the system supports the user in their work, particularly in terms of functionality, interface/access and decision support. It must be emphasised that a search system and the search process together form the entire information searching and retrieval process and it is not acceptable to evaluate the system alone. The image retrieval systems will therefore be evaluated in context and from a user-oriented perspective.

The evaluation tool will be developed by performing user-centred system evaluations with existing image retrieval systems currently in use in various organisations. The initial evaluation procedure will be informed by the models of image seeking behaviour developed during VISOR I, by a comprehensive review of the relevant literature and by input from experts in the field of evaluation. The results of the practical work conducted during a pilot phase (i.e. initial user-centred system evaluations) will be used to reformulate the evaluation procedure accordingly. The resulting procedure will then be used in a second organisation to verify the approach and make further modifications as necessary. Finally, an evaluation framework to guide the development of electronic image retrieval systems will be formulated, incorporating the experience gained throughout VISOR II. Throughout the project, feedback sessions will be held with both participant organisations and experts in the field.

These grants were funded under the Library and Information Commission's Information Retrieval Programme, *http://www.lic.gov.uk/research/information_retrieval/index.html*

DIAC 00

Shaping the Network Society: The Future of the Public Sphere in Cyberspace

DIAC-00. A Directions and Implications of Advanced Computing (DIAC) Symposium Sponsored by Computer Professionals for Social Responsibility
May 20 - May 23, 2000
Seattle, Washington, USA

Cyberspace may become the dominant medium through which people create and share information and ideas. How their conversations about the environment, culture, leisure, and political decisions, are conducted and how they are resolved are likely to have major social implications in the future. What directions and implications does cyberspace foretell for community, democracy, education and culture? Addressing those questions may be among the most urgent tasks facing humankind today.

The objective of DIAC-00 is to integrate many perspectives, conversations, and people from around the world on the topic of public space in cyberspace: What is it? What should it be? What would we do with it? What can we do about it?

While DIAC-00 will present "best practices" and other lessons learned "from the field" there is an urgent need for theoretical work (or "condensed practice") as well. For that reason, DIAC-00 is strongly encouraging reflective work on strategic and policy levels. There is enormous energy found at the grassroots level and it is growing. The big problem today is framing the idea of public space in cyberspace in a way that engages intellectuals, decision-makers, artists, and citizens. This can only be done by combining "best practice" stories with strong provocative conceptualizations of what is happening in our world and how public cyberspace can play a role. We need theories, concepts that can help us discuss, reflect, and take action on these critical matters. As an integral part of the DIAC-00 conference social scientists, engineers, computer scientists, artists, journalists, and other members of the research community will contribute their thinking on these pressing issues:

Community Informatics ; Civic Knowledge, Civic Infrastructure ; New Tools, Applications, Services, and Institutions ; Theoretical Frameworks ; Methodological Frameworks ; Critical Theory ; Social Economy of the Internet ; Computers, Work, and Cyberspace ; New -- and Retooled -- Media ; Participatory and Community-Centered Design ; Community Initiatives ; Public Access and Community Networks ; Practitioner and Researcher Co-Learning ; Bridging the Digital Divide ; Cyberspace Policy -- Social Policy -- Cultural Policy ; Computer-Supported Community Work ; Localism and Globalism ; International Perspectives and Partnerships ; Social Movements and Collaborations

DIAC-00 will be a multifaceted event. This call for abstracts /papers addresses the research or academic component of the symposium. There are other opportunities for participation within this framework. The guidelines for workshop proposals will be released soon.

DIAC-00 will be the seventh symposium sponsored by Computer Professionals for Social Responsibility in the "Directions and Implications of Advanced Computing" series. DIAC-00 is intended to broaden the discussion and awareness about the future of cyberspace both in terms of topics and in terms of participation. It is also our intent to provide visibility to topics and perspectives that are often neglected by the media.

Each extended abstract should contain a description and outline ofthe work, supporting evidence and data, and references. Abstracts and papers should be written in English. All extended abstracts should be submitted (in plain text only!) electronically to Peter Day (p.day@btinternet.com). Abstracts should be fewer than 2,000 words. Authors should remember that they will be addressing non-academics as well as academics at this conference and avoid jargon whenever possible. Citations should follow the Harvard Citation guidelines.

Important Dates:

February 15, 2000 extended abstracts due;

March 15, 2000 feedback given to authors;

May 1, 2000 revised abstracts due.

May 20 - May 23, 2000 DIAC-00.

The final papers, ready for book / journal, will be due sometime in summer 2000. We are planning to publish all submitted abstracts on our web site. We are planning to publish accepted papers in a book or journal. The academic program will be thoroughly integrated with the rest of DIAC-00.

DIAC-00 is sponsored by Computer Professionals for Social Responsibility and co-sponsored by Friends and Partners. Please contact us if your organization would like to become a co-sponsor or endorser. We'd like to thank the Morino Foundation for their support.

For more information about the symposium, please see the web site (http://www.scn.org/cpsr/diac-00) or contact conference organizer Doug Schuler, douglas@cpsr.org, 206.634.0752.

One-day workshop on Evaluation of Information Systems

Queen Mary and Westfield College, University of London

15 September 2000

http://www.dcs.qmw.ac.uk/~mounia/EIS.html

With a growing amount of electronic, multi-media data being accessed by an increasing number and variety of end-users, it is becoming ever more important to design and build effective information systems which meet users' needs. An essential part of this process is the identification of suitable techniques and systems for particular users, or groups of users, in particular information-seeking situations.

Considerable research has already been carried out into methods of evaluating the effectiveness, efficiency and usability of information systems. However, there are still many theoretical and practical issues that remain unsolved. Much more work is required in order to move towards the development of a comprehensive framework for evaluation of information systems.

This workshop is open to anyone with an interest in information system evaluation, including academic and industrial researchers and practitioners working in the areas of information retrieval, library and information science, databases, artificial intelligence, digital libraries, the Web, and other related areas.

Content of papers

Papers discussing work in progress or completed work on evaluation of information systems are invited.

Topics include, but are not limited to:

* Problem issues in evaluation * Application of existing or traditional evaluation techniques * Novel evaluation techniques and methodologies * System-centred or user-centred evaluation, or integration of these two approaches * Theoretical or empirical evaluation, or integration of these two approaches * Evaluation of quality of results or quality of interaction, or integration of these two approaches * Evaluation of multi-media information systems * Application of HCI principles and techniques to evaluation

Authors are invited to submit three copies of their paper, in English, to be received by Friday 31 March 2000. Papers should be no more than 10 pages (of A4) in length, and should be formatted according to Springer Verlag's formatting guidelines for workshops in the electronic Workshops in Computing (eWiC) series. The guidelines can be found at *http://www.ewic.org.uk/ewic/editors/submitting.cfm.*

Papers will be refereed and, if accepted, will be published in the proceedings of the workshop. There is also the possibility (currently under discussion) of publishing the workshop proceedings as part of the eWiC series.

Papers should be sent to:

Jane Reid, Department of Computer Science, Queen Mary and Westfield College, University of London, London E1 4NS

Important dates

Deadline for submission: Friday 31 March 2000

Authors notification: Friday 9 June 2000

Final submission of camera-ready copy: Friday 21 July 2000

Workshop organisers: Mounia Lalmas, Jane Reid (QMW)

Program committee: Pia Borlund (Royal School of Library and Information Science, Denmark), Nathalie Denos (CLIPS IMAG, France), Mark Dunlop (Risoe National Laboratory, Denmark), Theo Huibers (KPMG Consulting, The Netherlands), Frances Johnson (Manchester Metropolitan University, England), Tony Rose (Canon Research Centre Europe, England)

Local organiser: Sue White (QMW)

Enquiries

Informal enquiries regarding the workshop can be directed to the workshop organisers:

Jane Reid, Mounia Lalmas. Department of Computer Science, Queen Mary and Westfield College, University of London, London. E1 4NS. Fax: +44 (0)20 8980 6533. Tel: +44 (0)20 7882 5236 (Jane). Tel: +44 (0)20 7882 5200 (Mounia). e-mail: {jane, mounia}@dcs.qmw.ac.uk

This event in sponsored by the British Computer Society Information Retrieval Specialist Group, in association with the IEE Informatics Professional Group A4 (Artificial intelligence).

The *INFORMER*

The *Informer* is published quarterly by the British Computer Society Information Retrieval Specialist Group.

Editorial Team: Jon Ritchie/Ian Ruthven, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ. Tel: 0141 330 6292 Fax: 0141 330 4913. Email: informer@dcs.gla.ac.uk

Advertising: Jan J IJdens, BCS IRSG Chair,Sharp Laboratories of Europe, Oxford Science Park, Oxford. OX4 4GB. **Email:** jan@sharp.co.uk (or chair.irsg @bcs.org.uk)

Change of address/removal from mailing list: Specialist groups liason at BCS HQ **Email:** sg@bcs.org.uk