BCS

INFORMER

CEPIS

British Computer Society
Information Retrieval Specialist Group

Volume 9 | Autumn 1999 | ISSN 0950-4974

Reminder....

Springer-*Informer*

Prize for best student paper in IR '98-'99

The Informer in collaboration with Springer-Verlag are pleased to announce the launch of the third annual competition for the best student paper in Information Retrieval. This is an open competition for any student in a European academic institution who has published a paper in a refereed journal/conference/workshop in the period 1st November 1998 - 1st November 1999. Springer-Verlag have kindly donated a prize of £100 worth of Springer-Verlag books (full catalogue at <http://www.springer.de/product/index.html>).

The winner of the competition will be invited to present their paper at the 22nd Annual BCSIRSG Colloquium to be held in Cambridge, hosted by Microsoft. The IRSG will pay the winner's registration fees (and accommodation) for the Colloquium and travel expenses of up to £200 if the winner is able to present his/her paper.

To enter

1. Send a postscript or word copy of your paper to *informer@dcs.gla.ac.uk*, including publication details (conference name and dates/journal name, volume and number)

Rules

1. The paper must have been **published** or **accepted for publication** in the period 1st Nov 98 - 1st Nov 99 .
2. The paper must have appeared in a *refereed* journal, conference or workshop proceedings and should have a significant information retrieval content.
3. The entrant must have been a student at a European institution (university, college, etc) at the time the paper was *written*.
4. The entrant must be the main or only author of the paper.
5. Each entrant can only submit one paper for consideration.

Closing Date: November 15 1999

Contents

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
University of Glasgow
Email: jane@dcs.gla.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

Possible plans for the Y2K problem?

The Informer team is pleased to announce that it has solved the Y2K problem. The solution is to remove all computers from the desktop by Jan 1999. Instead everyone will be provided with an Etch-A-Sketch. There are many sound reasons for doing this:

1. No Y2K problems

2. No technical glitches, keeping work from being done

3. No more wasted time reading and writing emails

4. Substantial hardware cost savings

In order to ease the transition we have compiled a list of possible Frequently Asked Questions:

Q: My Etch-A-Sketch has all of these funny little lines all over the screen. What do I do?

A: Pick it up and shake it

Q: How do I turn my Etch-A-Sketch off?

A: Pick it up and shake it

Q: What's the short-cut for undo?

A: Pick it up and shake it

Q: How do I create a New Document window?

A: Pick it up and shake it

Q: How do I set the background and foreground to the same colour?

A: Pick it up and shake it

Q: What is the proper procedure for rebooting my Etch-A-Sketch?

A: Pick it up and shake it

Q: How do I delete a document on my Etch-A-Sketch?

A: Pick it up and shake it

Q: How do I save my Etch-A-Sketch document?

A: Don't shake it

Sigir 2000 Cfp

SIGIR 2000: Information Retrieval in Context

Department of Informatics, Athens University of Economics and Business Athens, Greece. July 24-28, 2000

http://sigir2000.aueb.gr

SIGIR is the major international forum for the presentation of new research results and the demonstration of new systems and techniques in the broad field of information retrieval (IR). The Conference and Program Chairs invite all those concerned with issues of IR to submit original research contributions, posters, and proposals for tutorials, workshops, and demonstrations of systems, for presentation at SIGIR 2000. All contributions should be submitted to the appropriate Chair, as indicated below (see the Conference web site for further details: *http://sigir2000.aueb.gr*).

Topics

Information Retrieval is contextual. IR functionalities form part of increasingly complex information systems serving a great variety of information tasks and behaviors. SIGIR 2000 seeks original research contributions in the broad field of information storage and retrieval, covering the handling of all types of information, user behavior in information systems, and theories, models, and implementations of IR systems. Topics relevant to SIGIR include but are not limited to:

IR Theory, including logical, statistical and interactive IR models, data fusion.

Experimentation: test collections, interactive IR experiments, evaluation measures, experimental design, testing methodology, scalability.

Natural Language Processing: word sense disambiguation, discourse analysis, summarization for the purposes of IR, use of linguistic resources for IR.

Contextual IR: multi-media IR, cross-lingual IR systems, speech retrieval, dialogue management, (non) feature-based indexing, information seeking and task embedded IR.

Interface issues: user & use modeling, human - computer interaction, search strategies.

Filtering, Extraction, Routing, and Text Classification.

Systems and Implementation Issues: integration with database systems, networked systems and the internet, compression, efficient query evaluation.

Applications: electronic publishing, digital libraries, text mining, WWW-related issues, semistructured document retrieval.

Important dates

January 14: Original research paper submissions due. 5 hardcopies of full papers (max. 5000 words) to be submitted to the relevant Regional chair. **Electronic submissions will not be accepted**.

February 11: Proposals for tutorials, workshops, posters, panels and demonstrations due. Please send submissions by email only to the relevant Chairs. For requirements for submission please see the Conference Web site: *http://sigir2000.aueb.gr*

April 1: Notification of acceptance of all submissions.

May 1: Final camera-ready copy of all submissions due.

Conference chair

Emmanuel Yannakoudakis, Athens University of Economics and Business, Department of Informatics, 76 Patission Street, Athens 104 34, Greece (eyan@aueb.gr), Phone: +30-1-8214145, Fax. : +30-1-8203356

Programme chairs

The Americas: Nicholas Belkin, Professor and Director of the Ph.D. Program, School of Communication, Information & Library Studies, Rutgers University, 4 Huntington Street, New Brunswick NJ 08901-1071, USA (nick@belkin.rutgers.edu). Phone: +1-732-932-8585, Fax: +1-732-932-6916

Europe and Africa: Peter Ingwersen, Royal School of LIS, Birketinget 6, DK 2300 Copenhagen S, Denmark, (pi@db.dk). Phone: +45-32-58-60-66, Fax: +45-32-84-02-01

Asia, Australia and the Pacific: Mun-Kew Leong, (Attn: SIGIR Submission), Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613 (mkleong@krdl.org.sg). Phone: +65-874-7864, Fax: +65-774-4998

Tutorials chair

Alan Smeaton, School of Computer Applications, Dublin City University Glasnevin, Dublin 9, Ireland (asmeaton@compapp.dcu.ie). Phone: +353-1-7045262, Fax: +353-1-7045442

Panels and demonstrations chair

James Allan, Computer Science Department, University of Massachusetts, Amherst, MA 01003-4610, USA (allan@cs.umass.edu). Phone: +1-413-545-3240, Fax: +1-413-545-1789

Workshops chair

Bob Krovetz, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA (krovetz@research.nj.nec.com). Phone: +1-609-951-2773, Fax: +1-609-951-2483

Posters chair

Amit Singhal, AT&T Labs - Research, Rm A-281, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932, USA (singhal@research.att.com). Phone: +1-973-360-8335, Fax: +1-973-360-8970

ECDL 99 Report

Sean Bechhofer
Information Management Group
University of Manchester
Oxford Road
Manchester M13 9PL
seanb@cs.man.ac.uk
Tel: +44 161 275 6145
October 1999

The Third European Conference on Digital Libraries was held in September in Paris. The objective of the conference is to bring together researchers from multiple disciplines to present their work on enabling technologies for digital libraries. Proceedings are published as vol. 1969 in the Springer-Verlag Lecture Notes in Computer Science Series. There were around 240 attendees and the program committee accepted 26 papers from 124 submissions. Jean-Francois Abramatic and Robert Wilensky gave keynote addresses, and other papers covered areas including information retrieval, semantic metadata (with much mention of RDF) and user interface issues.

Please note that the following discussion is very much a personal impression of the conference and should not be taken as a definitive or complete summary of the events.

General impressions

As a relative newcomer to the field of Digital Libraries, I was keen to attend the conference in order to get a feel for the state of the art and the current issues. I was also seeking an answer to the question "what *is* a digital library?" Although there were several interesting papers, I ultimately felt somewhat unsatisfied with the conference and found the question remained unanswered. This is not to say that the meeting was not useful and educational, but I occasionally found myself wondering "where's the *hard* bit in all this?". This may of course be my own misunderstanding of the fundamental problems of the field, but in a couple of cases, it seemed like existing solutions to problems were being reinvented (as discussed below).

As for the general feel, there was often little discussion in the sessions other than the statutory single question, and the conference as a whole was less animated than others I've attended in the past even though the atmosphere was friendly.

Keynote addresses

Jean-Francois Abramatic of the World Wide Web Consortium (W3C) started things off with a keynote on *Challenges for the Web: Universality and Scalability*. This was an overview of some of the activities of the W3C, whose mission is leading the Web to its full potential. The challenges for the W3C are concerned with architectures, best practices and preventing market fragmentation. It produces notes, submissions, working drafts, recommendations and open source code in the four domains of *Technology and Society; User Interfaces; Architecture* and *Accessibility Initiatives*.

Abramatic also discussed some of the issues surrounding the introduction of XML, in particular the idea that XML brings modularization, allowing browsers to choose which parts of the specification they will understand and display. He provided a demonstration of this, using the example of a mobile phone which has very limited display capabilities. Although the separation of content from presentation in XML is a good thing if it allows browsers to select how they are to render information, it's hard to see where this differs from, say, a traditional MVC approach where the data may be presented in different ways.

The second keynote address was given by **Robert Wilensky** from the University of California, Berkeley who spoke on *Re-Inventing Scholarly Information Dissemination and Use*. There is a need to rethink the information "lifecycle", particularly in the light of the increase of non-textual information. The cost of maintaining collections is growing by approx 15% a year and there are differences in publishing methodologies. In the past, because physical publishing was expensive, the tendency was to filter first and publish later, leading to long delays in availability. However, simply digitizing information tends to increase cost and provides little extra functionality. The process of peer review is vital in dissemination, and this will need to persist in an new architectures.

In the traditional state of affairs, authors submit content for review which is then made available to end users through libraries. Little has currently changed, other than the fact that content may now held in repositories, with users accessing these via indexes.

In Wilensky's proposed model, however, the emphasis is much more on collaboration, with everyone interacting through some central repository or library. In addition to the technology required to support collaboration and access to non-textual documents, *economic analysis* is required to understand how everything fits together. In addition test cases and prototypes are needed.

He also described a particular piece of technology known as *multivalent* documents. This adds a framework which allows the description of behaviours of documents along with layers that can extend their functionality. Such documents are conducive to developing a digital library-centric browser. He demonstrated the idea using a browser which added OCR functionality to a document which was actually stored as a scanned gif. This allowed the browser to search the document (by searching the results of the OCR analysis), giving the impression that the document was simply text. Along with that, semantic lenses were used which could combine that OCR functionality with

magnification etc. Annotation such as editors markup could also be added. The document could then be saved as an XML description which records things like where the document originally came from, along with the layers that sit on top of it. By associating the behaviours with the documents, the browser used a kind of "plug-in" architecture which meant that context dependent menus would appear allowing access to the bits of functionality. The current status of their implementation is that they have behaviours implemented for OCR and HTML. There was also a short demonstration of their blobworld system - a state of the art content based image retrieval engine. This didn't seem to fit so well with the rest of the talk, but was an opportunity to plug some fun research!

The multivalent documents and layering described by Wilensky in the talk seem very close to notions in open hypermedia (cf. Yndigegn's paper as discussed below).

Image access

Carl Sable of Columbia University, presented a paper on *Text-Based Approaches for Categorization of Images*.This described a system which attempts to categorize images based on descriptions, in particular using high-level, non-topic concepts, in this case whether images were of indoor or outdoor scenes. They took 21,000 news postings, from which they extracted 1675 image/caption pairs. 14 volunteers categorized the images and where shown either the images, captions or both and were asked to determine whether they were indoor or outdoor pictures. About 80% of the images were assigned as definite indoor/outdoor, the others had differing opiniions.

Their analysis used various parameters, looking at parts of speech, density estimation and used TF*IDF categorization. They also varied whether the analysis looked at the entire caption, first sentence, only the nouns and used a technique called density estimation. This relies on there being only two categories, and involves computing an indoor/outdoor score, then using the images with similar scores to then determine the indoor/ outdoor status.

Just over half images were used for training, of which 30% where indoor, 70% were outdoor. The system then tried determine the indoor/outdoor status of the remaining images and results were compared with those produced by the human categorizers. They reported various statistics on performance, the final analysis being that the first sentence was generally pretty important, normalization of the TF*IDF vectors was good, and the density estimation worked well. There was also some discussion of using the number of people in the picture as an extra help to determining the indoor/ outdoor status, although this is of course a strategy that's specific to the domain.

In *Metadata for Photographs: from Digital Libraries to Multimedia Applications*, **Anne-Marie Vercoustre** from INRIA discussed some of the issues in moving from a digital library to a CD-ROM about French rural houses. Various layers of content description were required: object/ subject: name, location, description; original photo: title, author, etc; digital photo: title, author, format (e.g. jpg), size, etc.

The production of the CD made use of XML descriptions which were then transformed automatically using templates into HTML. The talk was more about schema issues i.e. what to record rather than *how* one would represent content. For example, the Dublin Core allows the description of Date or Format, but in this particular case, the Date could be either the date of the original photograph or the date that the digitization or scanning took place - it's dependent on context. The DC is good for simple metadata, but has difficulties in differentiation between a photograph and its surrogates.

Howard Besser from UCLA discussed the MESL project in *Image and Metadata Distribution at Seven University Campuses*. The MESL project is a large scale project that distributes images and metadata from 7 museums to a collection of universities. Each university then implements its own system on top of this allowing access and retrieval. A typical implementation has a query answered with a collection of thumbnails and/or textual descriptions. They compared the various UIs (identical data sets but different delivery mechanisms) for things like presentation and layout, search options, image display and so on, and evaluated the costs. They found that digital distribution was good for individual usage, but problematic when you tried to do things like group viewing, partly because of issues concerned with projection and quality.

There has to be sufficient critical mass in terms of users and available material to encourage conversion to the digital media. There must also be transparent integration with local content. The usual problems with heterogeneity meant that the integration of the metadata was more of a problem than anything to do with the images.

Audio and video

In *Audiovisual Cultural Heritage*, **Gwendal Auffret**, described work on video broadcast archives at INA. In France, AV material is now considered an important part of their cultural heritage. There's a need to search and retrieve over broadcast archives for French television. However, the archives store more than just images and sound, as the documents have some intensional structure which may, for example, be recorded in the form of production notes. Their desire is to move from an simple audiovisual repository to a digital library, changing

the job from archivist to multimedia publisher.

There are three new challenges: encoding; delivering and navigating around the library and it's the third of these that they're currently interested in. They intend to use metadata and documentation to aid the navigation. Control of the metadata leads to control of the generation of material. There is a requirement for new formats and tools for metadata for AV resources - standardization must be addressed.

Jane Hunter from CITEC in Australia gave a paper entitled *An Indexing, Browsing, Search and Retrieval System for Audiovisual Libraries*. This described an application which aids in the generation of Dublin Core based metadata for videos. Their objectives are to: use metadata to generate online visual summaries; increase the usage of the library; represent DC and video-specific metadata in RDF; and automate links to related resources.

Video metadata can include non-textual information such as keyframes and there are problems with automatic indexing. For example, scene change detection can work ok, but it tends to pick up the first frame of each scene, which may not be the best in terms of acuracy or aesthetics. Closed caption decoders can be used, but (obviously) require that the material has been captioned. There can also be problems with formats and standardization with such things. Bibliographic data is useful, but requires human input.

They are using RDF scheme to constrain metadata using RDF schemas. The Dublin Core is used for the top level descriptions, and at the lower level, they use MPEG-7 as this is more appropriate. A tool allows scene and keyframe selection and produces RDF which can then build summary pages for the library. They are also interested in the possibility of automating schema translations between DC and MPEG-7.

A musical database was the subject of *Music Structure Analysis and its Application to Theme Phrase Extraction* presented by **Atsuhiro Takasu** from the National Center for Science Information Systems in Japan. The database contained musical phrases in MIDI form. Users can play or sing a phrase, and the system uses feature analysis and matching to find results. Various features like pitch, duration of notes, velocity and so on were used to match phrases. The system seemed like it might be a little biased towards Western-style pop music as it used rests to try and identify where phrases started and finished. In addition it wasn't clear how well the system would perform when faced with music that didn't have a simple phrase structure.

Information retrieval

Claudio Carpineto from Fondazione Ugo Bordoni in Rome presented *Effectiveness of Keyword-Based Display and Selection of Retrieval Results for Interactive Searches*. Current problems with results display are that documents are presented one at a time, and the relationship between the document relevance and the query can be obscured. Their idea is to produce an unsophisticated display, with efficient algorithms which would scale. They use the notion of a view which is a subset of documents containing a certain subset of terms. These views are shown using a bar chart display that indicates the number of documents in each view. A prototype of the system can be seen at *http://www.fub.it/viewer*. They performed a couple of experiments, comparing their system to Alta Vista. The conclusions were that the approach helped the users by allowing easier selection of relevant results.

In *Predicting Indexer Performance in a Distributed Digital Library*, **Naomi Dushay** of Cornell University discussed the problems of resource discovery in a distributed environment, and the use of query mediators. An *Indexer* is a search engine, and their basic architecture involves a query mediator sitting inbetween a User Interface and a collection of search engines. The mediator chooses which indexers should be used and adaptively reacts to operational conditions, e.g. machines and services which may be down. The collection service holds information about the collections which are available. The key question is: *Given a query, where do I go to answer it?*

Selection critieria are based on performance, cost, content, licensing issues etc. Their current research focuses on performance, e.g. selecting indexers that can provide a rapid and reliable response. On average, they found that query managers spent 50% of their time waiting for indexers to respond. They applied Prediction Methods which attempt to determine what the performance might be. These use metrics like: Running Average of all the accesses; Single Last Observation and Low Pass Filters, where the recent behaviour is weighted more heavily.

The experiments where for a single indexer w.r.t. one Query Manager. Two measurements are kept. *Availability* - will the indexer respond within a given time? *Response Time* - if it will respond, how quickly will it answer the question. Results suggested that single last observation worked well for availability, while a running average was good for response time. Overall, prediction didn't seem to improve much with the addition of complicated algorithms (like low pass filters). A remark from the floor suggested pursing "anti-social" strategies, i.e. just asking all sources and then waiting for the first one to come back.

User adaptation

The results of an experiment were

reported by **Yin Leng Theng** of Middlesex University in *Design Guidelines and User-Centred Digital Libraries*. The objectives were to: investigate useful design factors; investigate the effects of the lack of those features and produce some guidelines.

The experiment involved taking 10 CS staff and students and getting them to perform two tasks on three sample libraries. The tasks were browsing, or navigation without a specific goal or purpose and searching for a particular document. The subjects then answered 40 questions. The design features of the libraries were divided into nine categories including things like screen display, navigation and so on. These were rated by the subjects using 7-point scale.

Two areas were identified as being important, navigation (subjects got lost and weren't sure where they were, where they'd been etc) and customization. To counteract this, they suggest the addition of web-like features e.g. document headers, hypertext links etc. There was also mention of the principle of Equal Opportunity - outputs should be available for use as inputs with or without modification.

This was interesting, but the experiment seems far too small to be able to claim any meaningful results. As a question from the floor pointed out, they had two tasks and three libraries and only 10 subjects, which provides little statistical significance. The paper and presentation also gave quantitative analyses about "% satisfaction" (based on the 7-point ratings) which is questionable.

Users have both short-term (e.g. who won the match on Saturday) and long-term (e.g. an interest in Information Retrieval) information needs. In *User Profile Modelling and its Application to Digital Libraries*, **Umberto Straccia** of CNR addressed the problems of meeting long-term needs - most sources and searches deal only with short-term needs, while long-term needs are simulated through repeated queries.

User profiles can help to record what has to be gathered and how it is to be delivered. Their approach uses the P3P (Platform for Privacy Preferences) which is a W3C standard for user profile information. The goal of the system is to automatically alert the user when something new (and appropriate) appears in the sources. They described possible pull and push architectures - in the pull case, the profile is used to generate a query for the retrieval engine, in the push case, the addition of any new documents leads to a matching against existing profiles.

Knowledge sharing

Irini Fundulaki of INRIA presented *Integrating Ontologies and Thesauri to Build RDF Schemas*. *Metadata* schemas allow us to express semantics within a given domain in terms of elements and semantic relationships. However, creation of these schemas is difficult. The solution proposed here is to try and integrate existing components. They draw the following distinctions between an ontology and a thesaurus. Ontology: a structural shareable view of information, for example CDOC/ ICOM for the cultural domain. *Thesaurus*: a deep and detailed classification scheme, e.g. AAT.

Ontologies provide structure but not classification. Thesauri provide structure but no explicit constraints or relationships. So, the solution is to integrate ontologies and thesauri. This follows a three step process.

1. Connect thesaurus terms to ontology concepts. **2.** Extract a so-called concept thesaurus structure for each concept. This is produced by taking all the terms labelled with a concept or any of its children, and then adding generalization relations from the original thesaurus. **3.** Generate an RDF schema. This is done by mapping concepts to RDF classes, roles to RDF properties, then mapping the concept thesaurus terms to classes and adding a subclass relation from the root of the concept thesaurus to the concept class.

Steps 2 and 3 are effectively mechanical translations, while step 1 is the interesting stage which involves thought and effort from human beings. The ideas have yet to be fully implemented although a prototype is underway which loads RDF descriptions into O_2. Future work includes the incorporation of other relationships (synonyms, partonomy etc), the addition of a thesaurus query language. Work is planned into providing tools for specifying connections and creating concept thesauri - vital if the approach is to succeed.

In *Dynamic Use of Digital Library Material - Supporting Users with Typed Links in Open Hypermedia*, Christian Yndigegn from Aarhus provided a description and demonstration of a system that uses an open hypermedia framework to support access to a digital library. They want to support users who are working with diverse information sources, who may be collaborating and using several applications. They intend to add structure on top of existing documents, and the demo used a scenario where two teachers preparing a course added links to things showing that they were required reading etc. It didn't seem however, that there was anything particular to digital libraries here, other than perhaps the choice of link types - it's simply standard open hypermedia where the content happens to sit in a digital library.

Modelling and accessibility

Yannis Velegrakis from the University of Toronto presented *Declarative Specification of Z39.50 Wrappers Using Description Logics*, a paper describing a proposal for dealing with access to distributed and

heterogeneous sources. Z39.50 (an ANSI standard) describes protocols for information exchange. It provides a collection of standard messages and formats, a view of the world as a flat vocabulary of fields (known as Access Points), and a collection of primitives allowing the expression of queries in the form of field-value pairs. In order to deal with the Z39.50, a server needs to map from the Z39.50 query to the underlying representation, i.e. provide wrappers.

Their idea is to use a Description Logic (DL) layer between the Z39.50 query and the data source. The source is modelled using the DL, with the Access Points described using concept descriptions. The wrappers can then be checked for quality through the use of the DL reasoning services, in particular consistency checking can be achieved through satisfiability tests on the logical descriptions. There was some discussion of how to deal with the mismatch between the models for query in DL and Z39.50. The translation has to introduce what they path *expressions* to take care of this and ensure that the DL query corresponds to the original query.

Although this is an interesting idea, the system is yet to be fully implemented. So far, they have simulated the use of the DL in wrapping the SIS System and are currently evaluating which DL system will be used for a full implementation.

Michael L. Nelson from NASA gave a paper on *Soda: Smart Objects, Dumb Archives*, an architecture which suggests moving functionality from the archives to the objects themselves, encapsulated in objects which are known as buckets. It wasn't clear to me how this differed from a standard OO viewpoint, and whether the architecture was providing anything more than what could be achieved through the use of, say, CORBA.

The author is a Researcher in the Information Management Group of the Department of Computer Science at the University of Manchester with interests in Knowledge Representation and its use in Semantic Metadata. He was able to attend the conference through the award of an ERCIM fellowship, for which he is grateful.

Jason Farradane Award

The Jason Farradane Award is presented annually to one or more individuals for an outstanding recent contribution within the information field. The 1999 award was made to Michael Keen, who, until his recent retirement, was Reader in Information Science at the University of Wales, Aberystwyth. Michael is a Fellow of both the Institute of Information Scientists and the Library Association, and a member of the Information Retrieval Specialist Group of the British Computer Society. He has made significant contributions to research and practice in information retrieval over the past thirty six years.

After working in public and special libraries Michael joined the celebrated Cranfield research team in 1963, migrated to Cornell University where he worked with Gerard Salton on ground-breaking techniques in experimental information retrieval, before settling at the then College of Librarianship Wales, Aberystwyth.

His research and teaching have covered most aspects of information retrieval, but with special contributions in indexing, vocabulary control, searching, printed index design, text retrieval software evaluation and laboratory evaluation test methods. His high standing in the information retrieval community, underpinned by a fine publication record, was reflected in the Readership conferred on him by the University of Wales, Aberystwyth.

Michael's latest research embraces interactive ranked retrieval systems.Awarded a grant by the British Library he undertook the design of an interactive system which offers multiple match methods. Typically, he subjected the work to careful evaluation, the published results representing a further valuable contribution to the literature ofinformation retrieval and the discipline of information science.

Book review

Predictive Data Mining: A Practical Guide. Sholom M.Weiss and Nitin Indurkhya.

Morgan Kaufmann. 1998. ISBN 1-55860-403-0. £30.95. 228p. Softbound.

This book should be on the list of recommended reading for anybody considering embarking on a data mining project that involves a predictive task. Within the book, predictive data mining is viewed as learning decision criteria for assigning labels to new unlabelled cases. It focuses on problems such as classification, regression and time series analysis.

The book is clear and concise, and it should be understandable and useful to the more experienced data miner and novice alike. Moreover, the amount of ground covered is considerable for a volume of this size; sufficient breadth of coverage is given for the book to be suitable for reference purposes. A wealth of sound advice is given throughout, with useful and practical information being provided regarding the approaches to follow and pitfalls to avoid.

A good presentation of statistical evaluation techniques is given, including the classical hypothesis

testing model and how it can be used to compare the predictive performance of different prediction methods. Measures of performance for both classification and regression tasks are discussed, as is the random sampling procedure that may be used in conjunction with these measures to produce accurate projections of the future. There is also some discussion of the potential pitfalls of evaluation, such as the introduction of bias through too much testing, or the evaluation of incomplete solutions on test data. The discussion of evaluation is put into the context of large amounts of data, which is what data mining is all about, and guidelines and reassurances are given for the correct induction and evaluation of prediction models.

A simple model for the data is described which allows all prediction methods to operate from the same basis. Extensive advice is given on how to transform data so that it fits into this simple model, for example by normalising and smoothing the data. Smoothing operations are not often described in the literature, but are reported within this book to have a positive impact on the posterior learning exercise, making this is an important contribution. The treatment of missing values is somewhat brief considering that this often a significant problem in data mining. Time series and text data are also studied, and some transformations that can be applied to convert them into the simple model are explained. The task of time series or text data mining can be quite daunting given the complexity of the data format and its divergence from the standard format accepted by most learning algorithms, so the presented advice regarding how such data should be transformed is very helpful.

Data reduction techniques are discussed and classified as those that reduce the number of features, cases or values in a feature. Various feature reduction methods are reviewed, including some that operate by assessing features individually and those that examine subsets of features collectively. The technique of Principal Component Analysis, which has been widely used and proven to be successful for many applications, is also explained. A variety of techniques for reducing and smoothing values are discussed; these range from simple rounding operations to more complex clustering operations, which try to minimise the average distance of values in a bin from its mean, to those which take account of the value of the class, so that values of the same class are clustered together as far as possible. Case reduction is considered in the context of different problems such as multiclass classification, regression and low-prevalence classification. Approaches such as using a single sample, incremental sampling and average samples are described. The future performance of a solution is discussed for these methods, taking into consideration issues such as the complexity of the solution and how that may affect the expected error rate on new cases. Data reduction may be essential in data mining due to the size of data, which may be too large for some prediction problems, and is also important because the time taken to induce solutions may be reduced without a significant loss in predictive performance; this aspect of the book is important and well covered.

The book then presents some of the learning techniques in a clear and concise manner. The chosen presentation is to divide the techniques into maths, distance and logic solutions. Maths solutions include linear models, neural networks and advanced statistical methods for regression. Distance solutions include nearest neighbour algorithms. Logic solutions include decision tree and rule induction techniques. For all these methods, advice regarding the required solution representation, data preparation and dimensionality reduction, together with the complexity of solutions is given; an overall assessment of each method in terms of its explanatory capabilities, performance and suitability for different tasks is also presented. Coverage of solution explanation, editing and combination is also given.

The penultimate part of the book is concerned with putting the described approaches into practice using a range of case studies. The authors assess how well different data reduction techniques work in combination with the different prediction methods. The assessment is based on the evaluation methods covered in the book, including the significance testing on relative errors of various experiments. Based on the results, some advice is given regarding which data reduction techniques work well in combination with particular prediction methods. For example, some interesting conclusions are presented such as the non-essentiality of feature reduction for most logic methods, and its effectiveness for maths and nearest neighbour approaches. The methods themselves are then compared, and the use of multiple solutions through bagging and arcing is investigated.

Finally, different types of common applications such as outcome analysis, transaction processing and text mining are discussed. For each application, all the phases of learning from data explored within the book, including data preparation, reduction, modelling and prediction, are discussed, and practical advice tailored to each application is given

This book certainly offers a high quality presentation of predictive data mining. A minor criticism would be that some of the terms employed have a different meaning to those which are commonly used within the data mining literature; however, this is not surprising, since little standardisation of vocabulary has taken place in the KDD community. A more comprehensive index would make the book more accessible as a reference, although this is not a major criticism, since we would expect the index to be in keeping with the relatively small size of the book.

B.de la Iglesia and J.C.W.Debuse,
University of East Anglia

The Informer gratefully acknowledges *The Computer Journal* for allowing us to reprint this review.

Recent awards

Concept-based Interactive Query Expansion Support Tool (CIQUEST)

Duration of grant: October 1999 to September 2002

Contact: Professor Micheline Beaulieu, The University of Sheffield, Department of Information Studies, Western Bank, Sheffield S10 2TN. m.beaulieu@sheffield.ac.uk. *WWW: http://www.shef.ac.uk/uni/academic/I-M/is/home.html*

The project will investigate a concept-based approach to provide user support for query formulation and reformulation in searching large-scale textual resources such as those of the World Wide Web. Given that users generate broad and brief queries and then encounter difficulties in refining initial queries on the basis of items retrieved, the proposal is to explore methods for the automatic generation and organisation of concept structures derived from retrieved documents.

The work will further develop a novel approach to clustering document sets which produces a visualisation of hierarchical menus which offers a highly compressed and comprehensible overview of the underlying documents. The focus will be on improving the concept identification process, based on co-occurrence information by applying information extraction utilities and will also seek to widen the range of concept relationships through various text analysis techniques. The validation and effective visualisation of the generated concept structures will be a prime concern. A user interface will be built to incorporate the display and navigation of the concept structures to support interactive query expansion (IQE). It is envisaged that the concept tool will assist searchers in selecting relevant documents as well as selecting potential terms for query expansion. It could thus serve as a viable alternative or complementary approach to IQE based on relevance feedback. A major emphasis is to include user participation in all the elements under investigation and to take full account of user searching behaviour as well as retrieval effectiveness in the approach to evaluation. By operating on retrieved documents the concept tool can be regarded as a front end to a search engine and will be tested on documents retrieved by different Web engines as well as ranking systems such as Okapi and Inquery. It is anticipated that the project will also contribute to the evaluation methodology for interactive systems and in particular to the design of interactive experiments.

Effects of spatial-semantic interfaces in visual information retrieval: three experimental studies

Duration of grant: October 1999 to September 2001

Contact: Dr Chaomei Chen, Department of Information Systems and Computing, Brunel University, Uxbridge UB8 3PH. chaomei.chen@brunel.ac.uk.

http://www.brunel.ac.uk/~cssrccc2/

Despite the proliferation of information visualisation techniques in the design of visual information retrieval systems, little is known about the effects of these techniques on users' search strategies and on their performance in information retrieval. Furthermore, it is not known whether there is a significant interaction between a particular visual representation and a range of cognitive abilities of individual users. Answers to these questions will have profound implications for the design and use of information retrieval systems.

The proposed research aims to investigate the effects of some of the most influential spatial-semantic interfaces for information retrieval. Notable examples of such interfaces include cone trees, multidimensional scaling models, associative networks, and self-organised feature maps. The research will analyse both quantitative and qualitative data regarding changes in users' search strategies and their performance across a range of visual information retrieval interfaces. The project will build a testbed for subsequent experimental studies which will include a number of data sets, such as the entire ACM Hypertext conference proceedings and a sub-set of the TREC document collection. The project will encompass three experimental studies with representative spatial-semantic interfaces:

multidimensional scaling models,
associative networks, and
self-organised feature maps.

A multi-disciplinary framework for the evaluation of Internet search engines

Duration of grant: October 1999 - September 2000

Contact: Dr Frances Johnson, Department of Information and Communications, Geoffrey Manton Building, Rosamond Street West, Manchester, M15 6LL. f.johnson@mmu.ac.uk. *WWW: http://www.mmu.ac.uk/h-ss/dic/people/fcjhome.htm*

The aim of this Project is to develop a framework for the evaluation of Internet Search Engines (ISEs) based on a multi-disciplinary approach with a strong emphasis on user-centred perspectives. The intention is to provide a contextual characterisation - incorporating users and their information searching tasks - for the comparative evaluation of ISE features. It is proposed that user-centred evaluative methodologies developed for information retrieval systems and in other contexts will be adapted to give a multi-dimensional matrix to structure the assessment of features and performance, attributes and task.

The framework developed and tested in this project will be presented as a toolkit incorporating:

- a taxonomy of ISE system design;
- lists of individual and task variations;
- a methodology for capturing and analysing user evaluation statements which relate to the identified quality dimensions of the system; and
- both quantitative and qualitative analysis techniques.

It is intended that the resulting user-centred methodology for ISE evaluation should provide system designers with a focus for the development of search features which meet both the expectations and the information search behaviour of their end-users.

Retrieving multimedia objects: an approach through synchronisation

Duration of grant: October 1999 - March 2000

Contact: Professor Peter Brophy, CERLIM, Manchester Metropolitan University, Geoffrey Manton Building, Rosamond Street West, Manchester, M15 6LL.

p.brophy@mmu.ac.uk.*WWW: http://www.mmu.ac.uk/h-ss/dic/people/pbhome.htm*

While text-based retrieval is relatively well advanced, the retrieval of images, audio and video is more problematic. For example, considerable work is in progress into the retrieval of images which do not have text descriptions (or where text descriptions are deemed inadequate) using automatic recognition of shape, texture, contrast, etc.

Multimedia (for example, objects which structure and inter-link video, audio, text, graphics and still images) could be even more problematic. However, if it is possible to interpret the explicit synchronisation of these media within the object, it should be possible to enhance capabilities of a retrieval system.

This feasibility study aims to establish whether synchronisation in general and SMIL (Synchronized Multimedia Integration Language) - developed to achieve synchronisation on the Web - offers a potential contribution to multimedia retrieval.

Investigations will focus on the information content of each contributing medium and whether by taking cross-media, synchronised 'snapshots' a series of retrieval clues can be built up into a powerful retrieval methodology.

VIRAMI: Visual Information Retrieval for Archival Moving Imagery

Duration of grant: October 1999 - September 2001

Contact: Dr Peter Enser, School of Information Management, Watts Building, Moulsecoomb, Brighton, BN2 4GJ. P.G.B.Enser@bton.ac.uk. *WWW:http://www.it.bton.ac.uk/research/im.html*

Producers, curators and consumers of moving-image heritage material are challenged by the costly and problematic compilation of metadata in support of their collections, and by the opportunities to migrate from the human-mediated to the computer-mediated communication environment.

In this project a systematic analysis of client information needs for moving imagery will reveal (a) whether there exists the same emphasis on uniquely defined and named visual features as has been found to be the case for archival still imagery; and (b) whether there is a significant pre-iconographic query content which, in seeking to recover images which depict aspects of cultural heritage (e.g. fashion, domestic appliances, industrial processes), invokes layers of specialised terminology, the specificity of which demands expert knowledge and thesaural support. This analysis will inform a view on the role, if any, which content-based image retrieval techniques might play, for example in a new breed of hybrid retrieval system, in alleviating some of the dependency on metadata content implied by the above areas of enquiry.

The analysis of client information needs will be undertaken in the South East Film & Video Archive (SEFVA), which locates, collects, preserves and promotes films and videotapes made in South East England. This archive will also provide the primary research tool for a complementary, client follow-up evaluation of current indexing and retrieval strategies in a representative moving image archive.

Image indexing and retrieval in the compressed domain

Duration of grant: November 1999 - October 2002

Contact: Professor Jianmin Jiang, School of Computing, University of Glamorgan, Pontypridd, CF37 1DL. jjiang@glam.ac.uk. *WWW: http://www.comp.glam.ac.uk/pages/staff/jjiang/*

Data compression is used to save computer storage space but also, more importantly, can be used to improve the efficiency of information retrieval and processing, since unnecessary redundancy has been removed in the compressed data. However, current image information systems store original images as pixel data. As the amount of storage required increases, data compression becomes essential. In such cases, the only possible way of using data compression to reduce the size of the database is to compress the population of images separately after each original image is analysed and indexed. Decompression is then required before any automatic retrieval can be performed.

This project will investigate a number of original ideas to develop an

IRSG 2000 Cfp

IRSG2000

2nd Annual Colloquium on IR Research.

5-7 April, 2000. Sidney Sussex College, Cambridge

This annual colloquium on information retrieval research provides an opportunity for both new and established researchers to present papers describing work in progress or final results.

Topics of interest include (but are not limited to):

:- Evaluation and testing of information retrieval systems
:- Information retrieval and the Web
:- Hypermedia/Multimedia indexing and retrieval
:- Information retrieval from non-text media (including spoken document, image/video retrieval)
:- Voice processing and retrieval
:- Logic and information retrieval
:- User interfaces for information retrieval
:- User interaction in information retrieval
:- Information retrieval in library systems
:- Networked information retrieval
:- Database and information retrieval integration
:- Data mining and information extraction
:- Natural language processing for information retrieval
:- Knowledge-based information retrieval
:- Intelligent information retrieval
:- Commercial applications of information retrieval systems

Submissions

Authors are required to submit their paper, in English, by 10 December 1999. Papers should contain at most 7500 words and be double-spaced. The abstract should not exceed 100 words.

Final papers will be required in the Electronic Workshops in Computing format (*http://www.ewic.org.uk/ewic/*). Authors are encouraged to submit papers in this format.

The submission should include two PostScript copies of the paper: one full copy copy and one anonymous. Both files should be submitted by anonymous ftp to *ftp.scms.rgu.ac.uk* under the */pub/incoming/irsg2000* directory. The file names should reflect the title of the paper but the anonymous copy should have the prefix "anon". (To protect author privacy it will not be possible to list files in this directory).

For the anonymous copy, the first page must contain the title of the paper and abstract, but no indication about the author(s) and affiliation(s).

In addition, authors must send an email message to irsg2000@scms.rgu.ac.uk containing the title of the paper, the name of the file that has been submitted, the author name(s), and the author affiliation(s), plus complete contact information (mailing address, telephone, fax and e-mail) for the author to whom correspondence should be sent.

Any queries regarding submission should be sent to: *asga@scms.rgu.ac.uk*

Further details

For further details regarding travel, programme of events etc. see *http://www.soi.city.ac.uk/~andym/colloq2000/cfp.html*

Important dates

Paper submission: 10 December 1999.

Notification of acceptance: 10 February 2000.

Final copy due: 10 March 2000.

Publication

All papers will be refereed. Following notification of acceptance, authors of selected papers will have until 10 March 2000 to make revisions in the light of referees' comments. Papers will be published in the draft proceedings which will be circulated to all delegates during the Colloquium.

Final papers will be published in the Electronic Workshops in Computing (Springer-Verlag) *http://www.ewic.org.uk/ewic/.*

Contacts

If you have any queries or problems concerning submitting a paper, please contact: Dr. Ayse Goker, School of Computer and Mathematical Sciences, The Robert Gordon University, St. Andrew Street, Aberdeen AB25 1HG, Scotland, UK. Tel: +44 (0) 1224-262713. Fax: +44 (0) 1224-262727. Email: asga@scms.rgu.ac.uk

---

contd from page 11

image indexing and compression algorithm through which automatic image retrieval can be directly operated in the compressed data domain without any decompression being involved. In other words, data compression is embedded inside the indexing and retrieval algorithm and hence concealed from end users. This approach would not only save the cost of storage space for an automatic image indexing and retrieval system, but would also improve its efficiency in terms of search speed, retrieval accuracy and convenience of use.

An image database will be constructed to provide a research platform and algorithms will be developed for image indexing using multiple keys in the compressed data domain. The effectiveness, efficiency and usability of the system will be evaluated.

These projects have been supported by the Library and Information Commission through its Information Retrieval research programme, *http://www.lic.gov.uk/awards/ir-curpj.html.*

The Usable Past: Historical Perspectives on Digital Culture

The University of Iowa Obermann Center for Advanced Studies announces Obermann Fellowships for the Summer 2000 Research Seminar

This interdisciplinary research seminar will address issues of digital culture by examining histories of the social integration of previous new technologies and linking them to present conditions. Precedents for our own digital concerns might be found in technologies as recent as 30 or 100 years ago or as distant as the Industrial Revolution and the Enlightenment, the invention of movable type and the Renaissance, or the invention of paper and Classical Antiquity.

A distinctive focus on four inter-related fields of knowledge will provide important touchstones: (1) audio-visual cultures' challenges or resistance to print, (2) cultures and politics of new information technologies, (3) perception and human experience, (4) the metaphysics of appearances and artifice. By focusing on historical models, each seminar participant will be able to contribute reflections on technology, ideology, and culture - past and present.

Scholars from all fields - history, English, American studies, communication studies, political science, art and architecture history, cinema and media studies, sociology, philosophy, business, engineering,the sciences - are invited to apply.

Applicants must hold a Ph.D. or comparable professional degree and should be ready to produce original, previously unpublished work for publication in a volume and to participate in lively,fast-paced sessions on readings, individual papers, visitors' lectures, and special events.

Participants will be chosen in part to provide sufficient range for a published collection of essays. Some fellowships are reserved for University of Iowa scholars.

Director: Lauren Rabinovitz, Professor, American Studies and Film Studies.

Time: June 12-29, 2000. **Place**: The University of Iowa Obermann Center for Advanced Studies, Iowa City, Iowa.

Fellowships: Ten fellows to be selected. $2,700 stipends, plus $500 for travel/housing expenses for visitors.

Services: Offices, personal computers, internet access, library service,technical support, copying, meeting rooms.

Objectives: Discussion of readings, submitted papers, and presentations by keynote speakers. Revision of the papers for a published book.

Deadline: January 26, 2000, including CV and a paper or prospectus.

Notification by late February.

Funded by the C. Esco and Avalon L. Obermann Fund and by the Office of the Vice President for Research.

Address inquiries to: Jay Semel, Director (jay-semel@uiowa.edu), Obermann Center for Advanced Studies, N134 Oakdale Hall, The University of Iowa, Iowa City, IA 52242-5000

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