# Informer

**BCS**
INFORMATION
RETRIEVAL

## About Informer

Informer is the quarterly newsletter of the BCS Information Retrieval Specialist Group (IRSG). It is distributed free to all members. The IRSG is free to join via the BCS website (http://irsg.bcs.org/), which provides access to further IR articles, events and resources.

The British Computer Society (BCS) is the industry body for IT professionals. With members in over 100 countries around the world, the BCS is the leading professional and learned society in the field of computers and information systems.

Informer is best read in printed form. Please feel free to circulate this newsletter among your colleagues.

Things never stand still at IRSG HQ. In the last issue we talked about the changes at BCS and how they were going to affect our membership structure going forwards (i.e. SGs for BCS members only, all that stuff).

Well, there's more. We're now about to overhaul Informer itself. Not that there's anything particularly wrong with the current format – it serves a purpose, after all. And, based on survey results, it does appear to have worked reasonably well for the past three years or so.

But, as with any publication, we need periodically to review our design and editorial approach to ensure we fully reflect the evolving interests and needs of our members. In particular, I think we've reached a point in the growth of the group where the production of Informer should be a more inclusive activity (which is a polite way of saying it should involve more people than just Andy MacFarlane and I!).

We've already reviewed some fantastic layout designs for the next issue – great credit goes to Murat Yakici for those. So the next challenge is organising the editorial side. I can see at least four opportunities here:

- Features Editor
- News Editor
- Reviews Editor
- Events Editor

I'd like to think that with more people on board we could grow Informer into a really great publication, with a unique offering both to our own members and to the wider search industry.

But we can't do it alone. We need your help! So, if you're interested in being part of the New Informer, and can spare a few hours a month, get in touch via informer@bcs.org.

# Informer

Meanwhile, a date for your diary: September 23, BCS HQ, London. As you probably know, Search Solutions is our main practitioner event, at which the major search vendors give us an insight into the future in presenting their latest innovations and ideas.

What you perhaps didn't know is that Search Solutions grew out of the Industry Day at ECIR 2006, a tradition which has continued with ECIR 2008 and is reviewed on page 2 of this issue. Also reviewed on page 8 is the workshop on Novel Methodologies for Evaluation in Information Retrieval, which was also held as part of ECIR 2008.

And to complete the picture for this issue, we have two book reviews: "Aspects of Automatic Text Analysis" and "Multimedia Retrieval". Our thanks go to Eric Jukes and João Magalhães for those.

We hope to have a new set of books available for review in time for October's Informer. So if you'd like to be the Reviews Editor next time around (or take on any of the roles above), just drop us a line at informer@bcs.org.

Best regards,
Tony

Tony Rose, PhD MBCS CEng
Editor, Informer
Vice chair, IRSG
Email: irsg@bcs.org.uk

## Conference Review: ECIR & Industry Day 2008

*By Udo Kruschwitz and Dawei Song*

The European Conference on Information Retrieval (ECIR'08), annually organised by BCS-IRSG, has celebrated its 30th birthday this year! It took place at the University of Glasgow from 28th March to 3rd April 2008 and was followed by an Industry Day.

While maintaining ECIR's student-friendly tradition, evidenced by student papers (26 out of 56 accepted papers) and student attendance, ECIR'08 also featured strong academia-industry interaction with authors coming from universities, research organizations and industry, representing 25 countries. The main conference also included pre-conference workshops and tutorials for the first time in ECIR history, and a "greener" conference.

The keynote speeches were given by Nick Belkin of Rutgers University, Bettina Berendt of K.U. Leuven, Belgium, and Amit Singhal of Google.

Nick Belkin addressed a number of "grand challenges for information retrieval", mainly some issues associated to users, such as the methods for inferring user intentions from implicit resources, the IR techniques responding to users' goals/tasks/intensions, user interactions, integration of IR and information seeking, characterizing context, personalised search, formal models for Interactive IR, and new evaluation paradigm for interactive IR.

Amit Singhal presented Google's view on the challenges and directions of web search. The major goal of web search can be summarised as: *"This is what I said, give me what I want".* The major challenges include the scalability, interface, adversarial information retrieval, new ranking algorithms fighting against the ever evolving spamming methods, and evaluation with respect to the dynamic collection and time sensitive queries. Some future directions were also talked about, for example, assistance to users to formulate queries, tailored search results for different

locations and language tools for cross-language information retrieval.

Bettina Berendt explored today's move from simple data retrieval to a more complex world where each information-related activity involves disclosing and withholding information, and where the dichotomy of information-seeking users and information-containing documents has vanished. This all calls for a new understanding of information literacy required to participate intelligently in this new world. The talk included numerous illustrative examples to highlight some of the (sometimes worrying) privacy issues in everyday search tasks that many of us are not even aware of.

## The major goal of web search can be summarised as: *"This is what I said, give me what I want".*

The main conference programme was, as usual, divided into parallel sessions, covering the core topics such as evaluation, IR models, user aspects, cross-lingual IR, web IR, as well as emerging topics such as social media, expert search and Wikipedia (*Wikipedia* being one of the main buzzwords of ECIR'08). The emerging topics (perhaps) also represent the pressing need for dealing with user/community-generated content and searching finer-grained information such as named entities rather than whole documents. The quality of the papers, presentations and posters was impressive.

Consistent with two hot topics (new evaluation methodologies and expert search), this year's best paper and best student paper awards respectively went to "Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions" by Kalervo Järvelin (University of Tampere), Susan L. Price, Lois M. L. Delcambre (Portland State University), and Marianne Lykke Nielsen (Royal School of Library and Information Science*), and* "Modeling Documents as Mixtures of Persons for Expert Finding" by Pavel Serdyukov and Djoerd Hiemstra (University of Twente).

Following the format of ECIR'07 in Rome, this year's ECIR had a short paper session. Each

short paper was given a five minutes introduction slot, followed by poster exhibition enabling Q/A and discussion.

The conference also accommodated the BCS-IRSG annual meeting. Leif Azzopardi, the IRSG chairperson, summarised the IRSG activities in the past year and reported the budget/plan for the next year. One of the important issues was to refine the rules for ECIR organization, in particular the clear relationship between BCS-IRSG and the conference organizers and a clarification of the responsibilities of the conference organizers. In addition, a new IRSG committee were elected during the annual meeting.

What would a conference be without social events? ECIR'08 had quite a number of them, including a banquet at the historical Kelvingrove Museum and Art Gallery, a civic reception at Glasgow City Chambers, and a Google-sponsored reception at Glasgow Science Centre. There was also time to sample the excellent Scottish cuisine in local restaurants such as the "Ubiquitous Chip". On top of that, you should have seen Leif, our local guide, demonstrating his enthusiasm and organisational skills by shepherding a group of 30 conference attendees through Glasgow to end up in a most amazing whisky bar.

## What has language technology done to improve IR? ... Not much

The main conference was followed by an Industry Day, the third one following successful BCS IRSG events in 2006 and 2007. The Industry Day provides a forum for academics to get an insight into IR challenges in the real world and offers commercial organisations to share their work and experience with a wider audience. Attendance had to be limited to about 50 participants. The talks were organised in sessions. The first one (*Language in IR*) was concerned with the reoccurring question of what language technology has done to improve IR. The short answer has so far been *"Not much."* However, all three speakers in this session, Hugo Zaragoza (Yahoo!), Daniel Tunkelang (Endeca) and Antonio Valderrabanos (Bitext), demonstrated that this is slowly changing and

that search engines are going to move closer to (domain-specific) conversation systems.

The second morning session (*Search*) started with a very different focus on analysing query logs. Nick Craswell (Microsoft) pointed out that what can be done with usage logs is underrepresented at academic conferences. However, industry has large amounts of such data and makes heavy use of it. The logs can be used to model relationships between queries (to suggest query modifications for example). Nick stressed that he is very keen on getting Microsoft clickthrough data released to the research community fairly soon (a tricky task after the AOL fiasco). This year's Google talk by Michai Stroe focussed on Google's 20% projects. He also brought up an interesting point about what sort of people Google is typically looking for: engineers who can do research. In the last talk of this session Jeffery Dalton (GlobalSpec) addressed challenges of vertical search concluding that this is a hard problem. Jeffery (as well as Daniel Tunkelang earlier on) saw huge potential in helping users to navigate the search space by offering query modification suggestions derived from domain-specific knowledge. Interestingly from a technical perspective was his discussion as to why Lucene is one of the most widely used search engines in industry summarising that despite its weaknesses it works in an industrial setting without being too academic.

## Lucene... works in an industrial setting without being too academic

The first of two sessions after the lunch break (*Enterprise and Business*) started with a talk by Theo Huibers (Thaesis & University of Twente) who combines both an academic and an industry perspective. He summarized his view of how search and document collections have changed and what challenges businesses face by suggesting *"Forget about the structure. Deal with the chaos!"* Richard Boulton (Lemur Consulting) proved that even as a very small company you can compete in the enterprise search engine market by using open source tools. He presented an example of a sophisticated search engine developed for an industrial customer that combines open source tools ranging from IR and simple text

processing to full-blown NLP and knowledge representation. Francisco Webber (Matrixware) and John Tait (IRF) concluded the session in a joint talk hat nicely brought academia and industry together: The Information Retrieval Facility (IRF), initiated by Matrixware in Vienna, provides large-scale patent data to interested research groups. What large-scale means in this context is that we are talking about Petabytes here!

The day closed with a session about how IR can support real-world problems in modern-day society. Friso Visser (Netherlands Public Library Association) started off by outlining what challenges traditional libraries face nowadays. He sees the future for public libraries in adding value and finding a niche such as the Dutch "Young People Library". The last two talks, by Iain Drummond (Memex) and Antonio Linari (Expert System), were both examples of the value that knowledge accumulated over decades can contribute to addressing real-world problems and competing in the market.

## Faceted search and language processing were major themes of this year's Industry Day

What did we learn? Faceted search and language processing were major themes of this year's Industry Day. Naturally, these topics sparked questions in the audience as to whether this is what people who are used to one-shot, two-word queries actually want. We will have to see! Open source tools featured prominently too. Most importantly, what makes the Industry Day such a great event is that the talks are not given by sales people but by technical experts who are also excellent speakers. Industry Day 2008 was a great success. We are looking forward to next year's event. And don't forget BCS IRSG Search Solutions 2008 in September in London!

In summary, this year's conference once again demonstrated the growth of ECIR in terms of both quality and diversity. There is no doubt that it has become a major information retrieval conference. Thanks to Iadh Ounis, Ian Ruthven, Craig Macdonald and all others for the excellent organisation.

# Informer

ECIR'09 will take place in Toulouse France, 6th-9th April 2009.

Udo Kruschwitz is a Lecturer in the Department of Computing and Electronic Systems at the University of Essex. He received a Diplom in Computer Science from Humboldt University Berlin and a PhD in Computer Science from University of Essex. His main research is in natural language processing, information retrieval and the implementation of such techniques in real applications. He is the author of the monograph "Intelligent Document Retrieval: Exploiting Markup Structure", published in Springer's Information Retrieval series. He is a member of the BCS-IRSG committee.
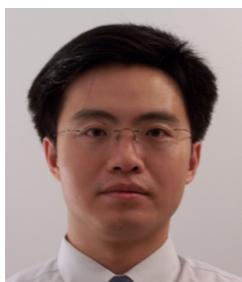
Dawei Song is a Senior Lecturer at the Knowledge Media Institute (KMi) of the Open University. He obtained his PhD in Information Systems from the Chinese University of Hong Kong in 2000. Before he joined KMi in September 2005, he worked as a Research Scientist (2000-2002) and Senior Research Scientist (2002-2005) at the Distributed Systems Technology Centre, Australia. His research interests include information retrieval models, web and enterprise search systems, biomedical information retrieval, text-based knowledge discovery, and management and search of complex data such as biological structures. He is also a member of the BCS-IRSG committee.

## Book Review

## "Aspects of Automatic Text Analysis", edited by Alexander Mehler and Reinhard Köhler

*Reviewed by Eric Jukes*

The importance of natural language texts as the prime information structure for the dissemination and management of knowledge is undisputed. It is of primary importance that relevant texts can be made available in different contexts for efficient task completion in academic and industrial settings. In order to meet this demand there is a need for automatic form and content based processing of texts, which enables the reconstruction or exploration of the dynamic relationship of language system, text event and context type. The new application areas, disciplines and methods such as text and web mining are proof of the importance of this task. The growing area of new media is another motivator on the further development of methods of text analysis with respect to their computational linguistic, information theoretical, and mathematical underpinning.

"Aspects of Automatic Text Analysis" is another book in the series "Studies in Fuzziness and Soft Computing". The series contains publications on significant recent developments in the areas of soft computing including fuzzy sets, rough sets, neural networks, evolutionary computations, probabilistic and evidential reasoning, multivalued logic, and related fields.

The book considers recent developments in automatic text analysis, and provides an overview of linguistic modelling through its collection of contributions of authors who focus on the topic of automatic text analysis from different perspectives (linguistic, mathematical, and information theoretical).

A dedication at the front of the book is "in honour of Professor Burghard B. Rieger on the

occasion of his 65th anniversary", although the dedication does not make it clear what the 65th anniversary is - I assume it is his 65th birthday. Professor Rieger is Professor emeritus of Computational Linguistics and former Head of Department of Linguistic Computing at the University of Trier. He has been active in research for more than three decades and has worked in language studies on topics ranging from history of German literature to linguistics of Germanic languages, with an early interest in computational approaches to linguistics starting in the 1960s. Most of his work is on fuzzy modelling in computational semantics and knowledge representation, focussing increasingly on cognitive models in computational semiotics, and dynamic systems of meaning acquisition and language understanding by machine. Thus, many of the contributions tend to laud Professor Rieger, and make reference to various aspects of his work.

## "Aspects of Automatic Text Analysis" cannot in any way be regarded as a general reader on Automatic Text Analysis

The book is divided into seven parts, and part one - Information Modelling - begins by looking at informational uncertainty and its alternative formalizations. Lofti A Zadeh, well-known for his fuzzy sets, introduces a constraint-satisfaction language for the formally precise and computationally tractable representation of natural language sentences, based on the concept of generalized constraints, which are also considered in a chapter by George J Klir and Kari Sentz, who consider linguistic approximation and linguistic variables. Both chapters look at informational uncertainty from the point of view of fuzzy set theory. However, Harald Atmanspacher, in "A Semiotic Approach to Complex Systems" takes a philosophical position in order to distinguish syntactic, semantic, and pragmatic information.

The second part of the book deals with models of semantic spaces, beginning with a chapter by Peter Gritzmann who introduces Rieger's basic model of semantic spaces, followed by a generalised model leading to a limit process in Euclidean space which can be used for identifying semantic strength and relevance of word tokens in text corpora. This part includes surveys by Edda Leopold of different space models starting from Rieger's Fuzzy Linguistics, through to building a semantic space by means of a family of support vector machines where each of the machines represents a single, pre-established topic category. The chapter by Alexander Mehler develops a formal apparatus for making semantic spaces sensitive to sign structure using the hierarchical structure of texts and their graph-inducing coherence relations.

Quantitative Linguistic Modelling is dealt with in the third part of the book, focuses on a wide area of linguistic structure formation. Stefan Bordag and Gerhard Heyer look at a structuralist framework combining the notion of syntagmatic and paradigmatic learning with its reconstruction in a model of nested levels of linguistic resolution. Reinhard Köhler develops a system theoretical model of order parameters which control quantitative characteristics of syntax. Latent connotative text structure. are considered by Arne Ziegler and Gabriel Altmann who explore structure formation within natural language texts based on reference chains, which are the chains of interlinked lexical items which refer to the same entity. Their methods yield one way of capturing denotative and connotative structures in text. The authors use the "hreb" for partitioning the text in denotative units. The hreb was named in honour of its discoverer, L. Hřebíček, and is the set of all entities of a text referring to the same object in reality or the same object in the text. I found this chapter difficult to follow as the examples of word hrebs were all in German (from Goethe's poem "Erlkönig", although I suppose this is really unavoidable.

Part four covers Corpus Linguistic and Text Technological Modelling. Michael Stubbs considers induction as the underlying method of reasoning in corpus-based lexicography and proposes the possibility of an entirely automatic semantic analysis, whilst pleading for the development of computer-based means supporting corpus linguistic work with respect to selecting and pre-processing large amounts of data. After my struggle with Goethe in the previous part, it came as a relief to find the

example used by Stubbs to be definitions of the word "crony" - although one can appreciate Goethe readers struggling here! Dieter Metzing and Jens Pönninghaus note the introduction of Extensible Mark-up Language (XML) as a standard for the representation and/or interchange of electronic documents, and explore the role of RelaxNG, an advanced document grammar, after first considering aspects of Kilivila verb morphology from the point of view of linguistic analyses.

Instead of grammar-based modelling, Alfonso Medina-Urrea explores morphological units, especially affixes, through Spanish, Czech, Ralámuli and Chuj. (Chuj, in case it is not immediately familiar, is is a language belonging to the Q'anjobalan-Chujean family of Mayan languages spoken by around 40,000 people in Guatemala and 10,000 in Mexico). Finally, in this part, Jürgen Rolshoven looks at Licensing Strategies in Natural Language Processing - licensing being a generate and test process in which each configuration hypothesis is evaluated.

## It is a heavy German work, destined to stand, unopened, on many a library shelf

Text categorisation and classification is covered in part five, and this part begins by considering the surface of argumentation and the role of subordinating conjunctions in a chapter by Winfried Lenders. Unfortunately this is another of those chapters which proves difficult for non-German readers, as the examples are in German.  Janusz Kacprzyk, Slawomir Zadrożny return to Zadeh's computing with words and perceptions paradigms in their chapter, Computing with Words and Perceptions for Text Categorisation bringing fuzzy and linguistic approaches to the problem. The chapter by Leonid Perlovsky on Neural Networks, Fuzzy Models and Dynamic Logic discusses possible relationships between computational intelligence, known mechanisms of he mind, semiotics, and computational linguistics. A particularly interesting chapter which briefly discusses relationships between the mind and brain and applications to understanding-based search engines.

Part six - Cognitive Modelling - starts off with Gert Rickheit and Hans Strohner who look at a cognitive systems approach to automatic text analysis and look at ways in which some mentalistic shortcomings of traditional approaches may be avoided by analysing language processing in the framework of cognitive systems. Hans-Jürgen Eikmeyer, Walther Kindt and Hans Strohner's Extended Experimental-Simulative Method is, according to the authors, no only able to contribute to the methodological integration of cognitive sciences but can also form a basis for theoretical progress. They suggest that if all authors in cognitive sciences could agree that good theories should be transformed into formal models, to be confirmed in human and computer experiments, then efficient criteria for theory testing would be available and that these criteria would be better met by system theoretically derived models than by other types of theory.

Part seven, the final part of the book, deals with visual systems modelling, and has only one contributor in this part, namely Wolfgang Wildgen, who explores the dimensionality of text and picture and he cross-cultural organisation of semiotic complexes. Wildgen analyses a number of works of art, including the force fields in Leonardo da Vinci's "Last Supper" (as, I recall, did Dan Brown in the "The Da Vinci Code"!) and concludes, among other things, that it is theoretically and empirically rewarding to analyse texts (sentences) and pictures as two different ways of solving the same semiotic problem: How can world-information be compressed into a basically low-dimensional representation, and what is the subsidiary system if coding levels which allows for the reconstruction of an imaginistic 3-space (+ time) and thus for the understanding of the picture or the text (sentence)?

The work includes an index of names as well as a comprehensive index. As a scientific book from a respected publisher, it naturally is fully referenced at the end of each chapter.

The book is of a decent font-size, clear black (not grey!) print, and is easy on the eyes - rather essential for a book which requires great concentration.

In terms of content "Aspects of Automatic Text Analysis" cannot in any way be regarded as a general reader on Automatic Text Analysis. It is a heavy German work (with a heavy price tag of £100) from the Springer stable, destined to stand, unopened, on many a library shelf, or, at least, only opened by those who are already deeply steeped in the field, and have the depth of knowledge, to be able to readily grasp the concepts which it covers, and to whom it will, no doubt, prove of enormous value. Its intended audience is students and researchers in soft computing and text analysis, but I hope that my detailed coverage of the contents of this work will enable the reader to form their own opinion as to its suitability for their work or for their current level, or depth, of knowledge.

*Eric Jukes MSc MBCS is Asset & Information Systems Manager in the Centenary Learning Centre at the College of North East London. His MSc in Knowledge Engineering is from Middlesex University, and was concerned with legal text analysis using case-based reasoning. He has an interest in Artificial Intelligence going back to 1984, and his current interests lie in the low cost digitisation and analysis of paper records and documents as well as geodemographics and Geographical Information Systems (GIS). He can be contacted via:* [ejukes@staff.conel.ac.uk](mailto:ejukes@staff.conel.ac.uk)

## Workshop Review
## "Novel Methodologies for Evaluation in Information Retrieval"

*By Michael Oakes*



A workshop on novel methodologies for evaluation in information retrieval was held at the Department of Computing Science, University of Glasgow, on 30th March, 2008, as part of the main ECIR 2008 Conference. The chairman was Mark Sanderson. The invited speaker, Tetsuya Sakai, felt that more could be learned about the success of an interaction with a search engine from click-through data than from a small user study. The searcher's prior knowledge is an important factor, and the list of Barry and Schamber was recommended for non-standard relevance criteria.

Ian Ruthven talked about literature-based discovery systems, where the most important criterion for relevance is the ability of a retrieved paper to provide an entry point into an unfamiliar, but related research area. This idea was tested out using a simulated work task situation, where for example Ph.D. students were asked to find "techniques employed in another area that might share commonalities with your work". Jussi Karlgren introduced the idea of measuring trust in information, as opposed to topical relevance. Thus two papers might both be on the theme of "Is aspartame safe to ingest?", yet score differently in a post-search questionnaire on how interesting they made the topic seem, whether they caused the users to change their opinions on the topic, and whether the users trusted the information in them.

Alessio Orlandi and Sebastiano Vigna led a short session on experimental collections for Web evaluation. One such collection was a snapshot of about 100 million pages harvested from the .uk domain using UbiCrawler, which can be used as a simulated crawling data set. Toshiyuki Shimuzu and Masatoshi Yoshikawa, in the context of XML retrieval, introduced the

concept of benefit versus effort. A user will specify the amount of effort they are willing to make, and by analogy with the knapsack problem, the system retrieves the non-overlapping elements of documents which have greatest benefit to the user within the acceptable effort threshold. The two types of effort considered were the effort of reading retrieved items, and the effort of browsing among retrieved items.

The invited speaker for the afternoon was Martin Braschler, who spoke about evaluating the search functionality of enterprise web portals. In particular, he emphasised the quality of the search index: a query term can only retrieve if it's in the index. Criteria for index quality are coverage / completeness, and up-to-dateness /freshness (where for example no dead links are indexed).

## The approaches described were radical departures from conventional, system-oriented measures

The following session was entitled "Evaluation Challenges for New Tasks", and began with news of the planned iCLEF Flickr Challenge, which will enable large-scale interactive evaluation of multilingual information access systems, based on interactions with content from the online photo-sharing service, Flickr. The other talk in this session was given by Gianluca Demartini, entitled "How many experts?", where a new evaluation task was suggested for expert search. The task was defined as finding the number of experts on a given topic in an enterprise, which can be evaluated by the simple metric of the absolute difference between the system estimation of the number of experts and the actual number.

Kalervo Järvelin talked about the importance of the "overall effectiveness of a session", rather than the system response to a single query. Since the initial query may not be optimised, and may need be modified, we should use measures such as the session "Cumulated Gain" which take into account the whole searching session, including such factors as individual scanning times, clicks made, the cost of query reformulation and issues of handling duplicates.

Leif Azzopardi introduced measures from transportation which could be used to measure the success of a search session. For example, the accessibility of a document far down a ranked list is like a long bus ride. A popular search topic is like a key location in a town. Another aspect of the accessibility of a document is that it depends on the likelihood of submitting queries which access that document. He also introduced a gravity-based measure.

Emma Barker also spoke of the need for measures which do not focus on single documents, but consider task-based scenarios involving multiple documents such as "write an essay" or "generate a hypothesis" over multiple documents. Multidocument relevance considers that documents may only be relevant in the context of other documents. One measure for such extrinsic evaluations is Thomas' (2005) completeness measure. Problems with task-based evaluations are that they are costly, labour intensive, and prone to noise from other tasks. She suggested "angle seeking" as a possible scenario for task-based evaluations of information access technologies, where for example two different angles on the politician Claire Short might be "Claire short sparked a series of resignations" and "Claire Short as a serial resigner who has led a charmed life".

Recall and Precision were hardly mentioned at this workshop, showing that the approaches described there were radical departures from conventional single query, binary relevance, system-oriented measures. It was an extremely interesting day throughout, and concluded with a cheese and wine reception where delegates were able to meet with attendees at other workshops. The workshop web site is at:
www.trebleclef.eu/call_for_papers_ecir08.php

*Michael Oakes is a Senior Lecturer in the School of Computing and Technology at the University of Sunderland. He teaches modules in Information Retrieval to second and third year undergraduates, and a Master's module in Natural Language Processing. He can be contacted via:*
Michael.Oakes@sunderland.ac.uk

## Book Review
## "Multimedia Retrieval", edited by H.M. Blanken et al

*Reviewed by João Magalhães*

Multimedia information retrieval has lacked a textbook as this one for many years. The editors realised this need when reorganizing a Multimedia Advanced Course for MSc and PhD students at the University of Twente. Faced with this need, they took over the task of compiling contributions from several researchers to create this textbook. "Multimedia Retrieval" is an inspiring collection of expertise that presents some of the state-of-the-art work done by researchers in The Netherlands. As such, the presented techniques reflect the contributors work and other existing techniques are presented in the (extremely useful) "Further Reading" section of each Chapter.

After a first look at the table of contents one easily realizes that the focus of the book is on high-level features and their use to support the retrieval process. Probably, the reason for this emphasis lies on the strong background of the contributors. After a second look I noticed the only missing topic to a comprehensive coverage of multimedia retrieval: high-dimensional indexing. Nevertheless, my enthusiasm was too high to let this minor issue change my view about this book. It became evident that the focus is on multimedia high-level features and for that the book is just fantastic.

### Overview

Most of the techniques presented in this textbook are based on pattern recognition algorithms which are properly described in Chapter 3. Information retrieval specific algorithms and techniques start on Chapter 4 with an overview of classic text IR techniques. The book then progresses into image indexing (Chapters 5 and 6), audio and speech indexing (Chapter 7), video indexing (Chapters 8, 9 and 10), and interaction (Chapter 11). The final Chapter discusses evaluation metrics and problems faced by multimedia retrieval.

### Image Analysis

Image content is the topic of Chapters 5 and 6. The first one addresses visual features extraction from images for indexing such as colour histograms, textures, and interest points. The second presents a generative approach to model keywords with visual features. More specifically, each keyword is represented by a Gaussian mixture model of DCT low-level features. Audio features and modelling, addressed in Chapter 7, are equally well covered.

### Semantic Video Analysis

Video analysis is covered by three excellent Chapters. As in the image and audio Chapters the goal of these Chapters are also to detect semantic concepts in video data. News video, tennis videos, and Formula 1 videos are the video data used in each Chapter. It is interesting to read on the different methods designed to tackle the specific concept domains of each type of video. For example, in Chapter 9 tennis motion detectors are used to classify a player movement. This contrasts with the more generic approach of Chapter 8 where concepts are visually very different. To address this variety, the framework is more flexible by allowing the combination of different algorithms for each particular concept.

### Interaction

Interaction is positioned both at the start and at the end of a multimedia retrieval system: it must process user requests and present the information extracted by the multimedia analysis algorithms in a meaningful way. A comprehensive view of interaction algorithms and visualization in multimedia retrieval systems is presented in Chapter 11. The Chapter discusses the different types of user input, relevance feedback based interaction, personalization based on user preferences, and presentation of information.

## Evaluation

The book finishes with a comprehensive presentation of evaluation metrics, methodologies, and evaluation efforts such as TRECVID, ImageCLEF, MIREX and INEX. Editors were "cruel" enough to tempt us with XML retrieval evaluation campaigns but did not dedicate any Chapter to this challenging topic. It is actually a pity given the broad Dutch expertise demonstrated in the INEX campaign. Thus, I leave here my request to the editors of this textbook to include a Chapter on XML retrieval on the next edition.

## Metadata and DRM

A very practical problem faced by production IR system is interoperability. Both metadata languages (Chapter 2) and digital rights managements (Chapter 12) play a central role in interoperability scenarios. MPEG-7, Dublin Core, and RDF are metadata languages that provide the vehicle to distribute collection indexes across distributed IR systems. The Chapter on Digital Rights Management discuss the legal framework to exchange content among users, also present a specific architecture together with an illustrative example.

## Summary

There are few books that fill such a large gap as this one. This is a unique textbook that gathers expertise valuable for any MSc student or PhD researching in the field of Multimedia Information Retrieval. Thus, this book serves superbly its objective as support material for any MIR researcher.

*João Magalhães is currently a PhD candidate at the Department of Computing, Imperial College London, working in the area of semantic-multimedia retrieval. His main research interests are on statistical modelling, information extraction and related applications. He is also on study-leave from his lecturing position at the Instituto Superior de Engenharia de Lisboa (ISEL).*

## Forthcoming Events

*Edited By Andy MacFarlane*

**Tenth International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008)**
Of interest to members who work in the area of disability and search. Halifax, Nova Scotia, Canada, 13th-15th October 2008.
http://www.sigaccess.org/assets08/

**Second Information Interaction in Context Symposium (IIiX 2008)**
Of interest to members who work in the area of context and IR BCS Covent Garden, London, 14th-17th October 2008.
http://irsg.bcs.org/iiix2008/index.php

**Audio Mostly - Third International Conference on Interaction with Sound.**
A conference on Audio, of interest to members working in the area of sound or music retrieval. Piteå, Sweden, 22nd-23rd October 2008.
http://www.audiomostly.com

**ACM Seventeenth Conference on Information and Knowledge Management (CIKM 2008)**
A conference with a number of major themes of interest to members including IR and information management. Napa Valley Marriott Hotel & Spa, California, October 26th-30th 2008.
http://www.cikm2008.org/

**7th International Semantic Web Conference.**
A conference of interest to members working in the area of semantic search. Karlsruhe, Germany, 26th-30th October 2008.
http://iswc2008.semanticweb.org/?page_id=21

**XIV Brazilian Symposium on Multimedia and the Web (jointly with the V Brazilian Symposium on Collaborative Systems) - WebMedia 2008.**
Of interest to members working on Multimedia and the Web. Vila Velha - ES, Brazil, 26th -29th October 2008. http://www.inf.ufes.br/ws2008/

**ACM International Conference on Multimedia Information Retrieval**
A conference on multimedia information retrieval, including interfaces, browsing, machine learning etc. Vancouver, Canada, 30th to 31st October 2008.
http://press.liacs.nl/mir2008/

**16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008)**
A conference on geographical information systems, of interest to members working in the area of

geographical IR or local search. Irvine, CA, USA, 5th to 7th November 2008. http://acmgis08.cs.umn.edu/

**4th International Conference on Digital Arts (ARTECH 08)**
A Conference focused on digital arts, which will be of interest to members working in the creative industries with a focus on IR. Porto, Portugal, 7th-8th November 2008. http://artes.ucp.pt/artech2008/

**1st International Workshop on Quantitative Semantic methods for the Internet (QSI 08)**
A Conference focused semantic web including search. Monterrey, Mexico, 9th to 14th November 2008.
http://www.cs.rmit.edu.au/fedconf/index.html?page=qsi2008cfp

**The 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2008)**
Of interest to members working in the area of semantic search. Monterrey, Mexico, 11th-13th November, 2008.
http://www.cs.rmit.edu.au/fedconf/index.html?page=odbase2008cfp

**Sofia 2008: Globalization and the Management of Information Resources.**
An information management conference with a theme on IR. Sofia, Bulgaria, 12th-14th November 2008.
http://slim.emporia.edu/globenet/Sofia2008/index.htm

**International Conference on COOPERATIVE INFORMATION SYSTEMS (CoopIS 2008)**
Of interest to members working in the area of social search. Monterrey, Mexico, 12th – 14th November, 2008. http://www.cs.rmit.edu.au/fedconf

**Third IEEE International Conference on Digital Information Management (ICDIM 2008).**
A wide ranging conference on issues of interest to members including information management, digital libraries and of course information retrieval.

London, U.K., 13th – 16th November 2008. http://www.icdim.org/

**7th International Conference on Practical Aspects of Knowledge Management (PAKM 2008).**
An including information management conference with a theme on information retrieval. Hiyoshi Campus, Keio University, Japan., 21st – 23rd November 2008.
http://pakm2008.comp.ae.keio.ac.jp/

**1st Joint International Conference on Interactive Digital Storytelling (ICIDS 2008).**
A conference on technology for storytelling which can include IR. Erfurt, Germany., 26th – 29th November 2008. http://www.ai.fh-erfurt.de/icids08/

**3rd International Conference on Semantics And digital Media Technologies (SAMT 2008)**
Conference on multimedia IR of interest to members working in the area of search. Koblenz, Germany, 3rd-5th December 2008.
http://samt2008.uni-koblenz.de/

**IEEE/WIC/ACM International Conference on Web Intelligence (WI-08)**
Conference on web mining of interest to members working in the area of web search. Sydney, Australia, 9th-12th December 2008.
http://datamining.it.uts.edu.au/wi08/html/wi/

**Forum for Information Retrieval Evaluation (FIRE)**
TREC for Indian Languages. Kolkata, India, 12th-14th December 2008. http://www.isical.ac.in/~fire

**The Second International Workshop on Evaluating Information Access (EVIA 2008), A Satellite Workshop of NTCIR-7**
A workshop on evaluation issues in IR. NII, Tokyo, Japan, 16th December 2008
http://ntcir.nii.ac.jp/index.php/EVIA-2008/

## Contacts

Web:            http://irsg.bcs.org/
Email:          irsg@bcs.org.uk
Subscriptions:  http://irsg.bcs.org/membership.php
ISSN:           0950-4974

To subscribe, unsubscribe, change email address or contact details please visit http://irsg.bcs.org/ or email irsgmembership@bcs.org.uk.
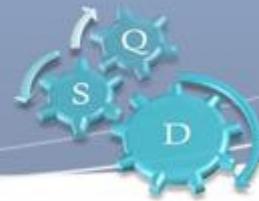
The IRSG is a specialist group of the British Computer Society.
To automatically receive your own copy of Informer, simply join the IRSG via the IRSG website.

# Search Solutions 2008

In conjunction with BCS-IRSG FDIA 2008

23 September 2008
Covent Garden, London

## Innovations in Search and Information Retrieval

Search Solutions 2008 showcases the latest innovations from the major search solution vendors and presents a unique opportunity for information professionals and researchers to view the next generation of information retrieval products and services.

Search Solutions is based on the 2006 BCS IRSG Industry Day (ID2006), an event which was collocated with the European Conference on Information Retrieval. The latest BCS-IRSG event follows a similar format, with a number of invited speakers presenting their latest innovations and sharing opinions and expertise. Last year's Search Solutions (SSE2007) is online and includes pictures and audio from the day.

The full programme is currently being finalized, and should include representatives from Yahoo, Google, FAST, Autonomy and many more.

### Venue

**BCS Central London Office**

BCS, 1st Floor
The Davidson Building
5 Southampton Street
London, WC2E 7HA

### Contact
Please direct all enquiries to Tony Rose email: tgr@tonyrose.net

http://irsg.bcs.org/SearchSolutions/2008/sse2008.php