Informer

Number 13, Winter 2004
ISSN 0950-4974

INFORMATION RETRIEVAL SPECIALIST GROUP
BCS
THE BRITISH COMPUTER SOCIETY

Contents

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.

Feel free to circulate this newsletter among your colleagues.

Welcome to the new edition of Informer. As many of you know, the last edition of Informer came out in 2002, and since then we've waited a long time for a new editor to come forward. Now that's resolved, let me outline a few of the opportunities I see for us going forwards.

Firstly, I'd like us to engage more fully with the practitioner community. There are thousands of information professionals in the UK (and further afield) who regularly use information retrieval (IR) concepts & technologies to get their jobs done, and I'd like to see us do more to address their needs and interests. In this regard I'll be looking to balance our traditional research-oriented content with more practice-oriented case studies, reviews and opinion pieces.

Secondly, as Editor I'll be adopting a more inclusive definition of IR, and looking to explore its relationship with other topical areas, notably:

- Search engines (web & enterprise, text & multimedia)
- Information architecture & information management
- Knowledge & content management
- Data mining & visualisation

This is, of course, in addition to maintaining coverage of our traditional 'core areas' such as text retrieval & classification.

Thirdly, I'll also be introducing a number of new columns, notably "*A Bluffers Guide to...*" (a brief introduction to a topical area, written in an informal, humorous style) and "*My PhD*" (yes, you've guessed it: a concise description of an ongoing PhD project). There are other new columns in the pipeline too, along with regular features such as jobs pages, events listings, and so on. (If you'd like to suggest others, by all means let me know).

One thing we won't be doing is competing head-on with mainstream IR journals: they do a perfectly adequate job already, and I see no particular virtue in us trying to compete with them. Moreover, I see great opportunities for us to cover the spaces left by the traditional journals, in providing more informal, topical coverage of interesting developments in IR theory and practice, and to do so in a style that encourages newcomers to the field to get involved. Incidentally, as part of this approach I'd like to adopt a 'rapid turnaround' policy wherever possible – if you have something to say that is interesting, topical and relevant, I'll do my best to see that it gets published in the next edition. Further details for prospective authors can be found on page 9.

Finally, in amongst all this I'd like to think there would be one common denominator: quality. I'd like to see us publish quality, original content, written in an accessible, inclusive style. Ideally, that content would be described as thought-provoking, compelling, radical even – but I appreciate we're not going to get there from here overnight. It'll take effort & commitment, but as is often the case, you get out of it what you are prepared to put in.

So in the meantime, I'd like to welcome you once more to the new Informer, particularly to those for whom this is a new experience. I hope you like it, and find something in here of value. If you do, tell your friends. If you don't, tell me.

Best regards,
Tony Rose
Editor, Informer
Email: irsg@bcs.org.uk

Events

Edited by Andy MacFarlane

SAC'05: 20th ACM Symposium on Applied Computing

Santa Fe, New Mexico, March 13 - 17, 2005. Special tracks on Information Access and Retrieval, concerned with the theory, implementation and evaluation of information access technologies to novel application areas and novel contexts. Also tracks on Data Mining, Document Engineering & lots more. http://www.acm.org/conferences/sac/sac2005/

ECIR'05: 27th European Conference on Information Retrieval

Santiago de Compostela, Spain, 21st - 23rd March 2005. The premier European conference on IR. Covers all major areas of information retrieval. http://www-gsi.dec.usc.es/ecir05/.

AVIVDiLib'05: 7th International Workshop on Audio-Visual Content & Information Visualization

Cortona, Italy, May 4-6, 2005. Focus on new solutions for content organization, access and interaction in audiovisual digital libraries. http://delos.dis.uniroma1.it/Workshops/default.aspx

ISI-2005 IEEE International Conference: Intelligence and Security Informatics

Atlanta, Georgia, USA, 19th – 20th May 2005. This conference focuses on security issues as pertaining to the intelligence community. Of particular interest to the IR community is the issue of Information Sharing and Data Mining theme of the conference, particularly the use of Digital Libraries as applied to archiving, preserving and searching intelligence information. Deadline for submissions is 20th January 2005. http://ecom.arizona.edu/ISI/.

COLIS5: 5th International Conference on Conceptions of Library and Information Science

Glasgow, UK, June 4-9 2005. Focus on exploring and understanding information-centred disciplines. It will examine how the issue of context influences the design and use of information access technologies. http://www.cis.strath.ac.uk/external/colis5/

I-Know'05: 5th International Conference on Knowledge Management
Graz, Austria, 29th June - 1st July 2005. The focus is on application-oriented research for Knowledge Management – submissions from both academia and industry/commerce are solicited. The conference has a number of tracks which include: Knowledge and Information Visualization (KIV'05), Knowledge Discovery and Semantic Technologies (KDaST'05) and Knowledge Sharing in Research and Higher Education (KSR'05). The deadline for submission on each of these tracks is 24th January 2005. http://www.i-know.tugraz.at/conference/i-know05/iknow05_home.htm.

SIGIR'05: 28th ACM Conference on Information Retrieval
Salvador, Brazil, 15th - 19th August 2005. The premier world conference on research in Information Retrieval. Covers all major areas of information retrieval. Deadline for submissions is 28th January 2005 for Full Papers, 4th February 2005 for Workshop proposals, and 25th February 2005 for Poster, Tutorial and Doctoral Consortium submissions. http://www.dcc.ufmg.br/eventos/sigir2005/.

ECDL 2005: 9th European Conference on Research and Advanced Technology for Digital Libraries
September 18-23, 2005 Vienna, Austria. The major European conference on digital libraries, and associated technical, practical, and social issues, bringing together researchers, developers, content providers and users in the field. http://www.ecdl2005.org/

Feature Article

"Beyond documents – a new wave of technology solutions"
By Alex Bailey

As a research student not so long ago, my experience of libraries, not surprisingly, has been dominated by shelves of books, journals, and magazines. Whether you are looking for a book itself, an article, or just an excerpt, the first thing you look for is the physical object on the shelf. The information I needed would usually be spread around the library, possibly on different floors. The size of the building and the size of the bindings, dictated how far I had to walk and how much I had to carry to the photocopier. My prize would be the text, which upon reading would turn out to be relevant, or irrelevant, or just relevant enough to make a mental note of its whereabouts and put it back on the shelf.

These days, all that heaving of large bound volumes from shelf to photocopier seems a thing of the past. And so it should be - after all, the information itself doesn't weigh anything, does it? We have moved some way beyond physical books; individual articles are available from the publisher's digital library, or from preprint archives; web-sites can be searched & browsed. The electronic document is now the primary information container, a homogenous, self-contained body of text. Web-pages, journal articles, emails, magazine or newspaper articles, memos, invoices - these are the targets of our search. If it's been indexed, then it will turn up in a list of search results on a computer screen near you.

Personal search

By now we are all used to finding documents on the web thanks to the likes of Google, but the latest wave of commercialisation in information retrieval that is making the news is 'personal search', or 'desktop' search engines. It has long been the case that you could find a document on a web-server on the other side of the world faster than you could find last week's email on your own machine. But now new products are appearing that index the documents on your hard drive. There are offerings from new hopefuls such as Blinkx (http://www.blinkx.com/), and from the big guys like Google and Microsoft. However the current trend is still document-centric. The search engines will help you find a particular document, and then it's up to you to find the information you want from the document.

In the library I had to carry around whole bound volumes of journals just for the sake of a single paper, likewise these days I have to plough through a list of documents to find the information I need. When I need a name, an address, or confirmation of a particular fact or opinion, then I don't need the whole document. What I really need is a tool that lets me search 'beyond documents'.

Analysis and visualisation

Of course, this is nothing new for the information retrieval community. There is a mature and growing body of research into question answering, XML retrieval, and the semantic web that is attempting to break down documents and index their constituent parts in an information-centric manner. This technology has yet to break into the commercial mainstream, but there are some innovative companies that are forging ahead. This article looks at some of the software available that attempts to put the information first, regardless of the original information container.

What is interesting about information-centric retrieval is that the presentation of the information is no longer constrained by the layout of the document, or the set of documents. Purpose-built visualisations and user interfaces can be used to display the information and allow the user to interact with them. It is this combination of information analysis and information visualisation that is key. And in the corporate information arena there are a number of applications that are driving this kind of technology into the market. Three application areas are reviewed here, namely Reputation Management, Business Intelligence, and Litigation Support Software.

Watch your reputation

Reputation management, also known as Sentiment Analysis, is a technological solution to the question ‘What do people really think of this product?’ The general idea is that the impact or reputation of a particular company or product can be extracted automatically from on-line news-sites, web reviews, blogs, and enthusiast’s websites etc. This can then be viewed and analysed in a way that would not be possible at a document-level view.

Developed and situated at IBM’s Almaden research centre, Web Fountain (http://www.almaden.ibm.com/webfountain/) is not a product as such, but a huge linux cluster that continuously crawls millions of web-pages each day. Information is extracted using natural language processing, statistics, pattern recognition, and machine learning resulting in a wealth of metadata stored in XML format. Sitting on top of this is a project which goes by the name of the ‘Sentiment Analyser’ that attempts to extract opinions, and hence reputations, from the underlying data.

Biz360 (http://www.biz360.com/) is a newer company undertaking a similar endeavour, without applying quite the same level of brute force as IBM. They offer products and services concentrating on automated market analysis.

Good solid intelligence

ClearForest’s (http://www.clearforest.com/) offering is a more traditional product that a company might buy and install in-house. The products are designed to meet the needs of ‘Business Intelligence’. Using the same concept of information extraction, entities and relationships are identified and counted, and the resulting information is displayed using a number of visualisation tools.

ClearForest apply these technologies to a number of business intelligence solutions, ranging from law enforcement, and federal intelligence, to ‘quality and warranty’ solutions. The latter is an attempt to further automate the discovery of issues surrounding product failure. It involves the analysis of customer support comments and service technician notes, which tend to be written in free text, and evade traditional systems. The goal is to capture the elusive information that just doesn't get seen by traditional analysis.

Legal eagles

Staunchly traditional, even the legal profession is undergoing a change in the way they handle documents. When a request for relevant documents is made in judicial or regulatory proceedings, it really matters that the important documents are found. In the past, when the bulk of documents were paper, this would be done by hand. Nowadays a significant portion of those documents is electronic, with no paper counterpart. This in turn has forced the legal profession to embrace electronic document management. A particularly interesting application of visualisation technology has evolved from this need, having been developed in-house at Preston, Gates, and Ellis legal practices, and spun-off in 2001 as the Attenex Corporation (http://www.attenex.com/). And yes, that is Bill Gates' dad.

The Attenex software uses noun-phrase extraction, document clustering, information retrieval and visualisation to present the user with an intuitive picture of the document collection, allowing them to review documents efficiently, and mark 'responsive' documents. What is interesting is that this application has been driven by a real need, which results in a genuinely useful application of the technology.

So when can we all get hold of this?

Now, the holy grail of a general-purpose information analyser, one that can extract and visualise information to satisfy any non-specific query is always a long way off. But it is good to see that today's research and technology is making its way into real-world applications. There are many more companies fighting for this new market space. Not all these enterprises will be profitable in the future, but no matter which of them survives, the next step up the technology ladder will be a little closer.

Uptake can be slow, however, and there are always barriers. When working in research it is tempting to look at a new technology in awe of the possibilities and ask why on earth isn't everyone using this today. A significant barrier is that most companies, large or small, are only starting to invest in company-wide document management solutions. There are more down to earth problems such as security, workflow, and systems integration that need to be tackled first. Only once this foundation is laid can more radical technologies be applied effectively. This is echoed in the personal computing space where the desktop search engines are only just breaking the surface.

Another important barrier to information extraction and visualisation technology today is a problem of human computer interaction. It's easy to get people excited about visualisation technology in the first instance because it looks cool. But then is it really useful? Many systems are not as intuitive as they first look, and people will need to be trained how to use them. Other systems just haven't been designed with the user in mind.

The technology will continue to develop through research, and corporate investment. When every corporate and home user is accustomed to the bread and butter of document management and information retrieval technology, when the document-centric view starts to become tired, that is when the real pull will happen. When general users realise that they need the information-centric view then it will be ready. At that point we can all be in control of the information without needing to plough through each document in turn.

Alex Bailey leads the Document Analysis team at Canon Research Centre Europe, where he has been working for the last four years. His interests focus on the use of document clustering, information extraction, and information retrieval for corporate document management systems. Prior to working at Canon he gained his PhD in statistical pattern recognition at the Image, Speech, and intelligent Systems research group at Southampton University. Alex is also the One-Day Event Co-ordinator for the IRSG and regularly attends ECIR and SIGIR conferences. He can be contacted via: alexb@cre.canon.co.uk.

My PhD

"Context Based Retrieval using Ontologies with Probabilistic Retrieval"

By Jagdev Bhogal

Background

A classic problem in information retrieval is that ambiguous queries produce a long set of results, some of which are not relevant to the user's needs. This problem exists on a much larger scale in web-based information retrieval where the user is not retrieving from a static document collection but a rapidly growing information resource [http://www.nature.com/cgi-taf/dynapage.taf?file=/nature/journal/v400/n6740/abs/400107a0_fs.html].

The early search engines returned only those documents whose index terms matched the terms in the user query. Third generation search engines known as meta-search engines work towards determining query context thus allowing the user to obtain more relevant results.

Context is such a broad term; it has been interpreted in different ways [http://www.cs.technion.ac.il/~lev/publications/context_search.pdf]. Personalisation involves learning short term and long-term user models i.e. user interest and preferences over time. The user models provide context for queries so two users with the same query could end up with different results if they have different profiles [http://portal.acm.org/citation.cfm?id=502693]. Another approach is to derive context from the information space surrounding documents such as different types of hyperlinks [http://jodi.ecs.soton.ac.uk/Articles/v02/i03/El-Beltagy/]. Context is also important in a ubiquitous computing environment whereby environmental information such as physical location changes the context.

Another way of determining context is to use concept hierarchies or ontologies [http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/IR99.pdf]. A node in the ontology hierarchy, also known as a concept can represent many different terms, and the hierarchy can be traversed to broaden the query or make it more specific.

Project Aims

The aim of my research is to investigate the use of ontologies for query expansion on ambiguous queries and to establish whether the use of semantic information can be used in the probabilistic retrieval model to improve its effectiveness.

Why ontologies, why probabilistic retrieval model?

The probabilistic retrieval model retrieves documents and then ranks them by calculating the probability of them being relevant to the query. The model combines term weighting, term frequency and document length information in a statistical manner. The research will be conducted using OKAPI. This is an information retrieval system that was developed at City University, London and is based on an improved version of the standard probabilistic model (BM25). The BM25 model has shown a considerable amount of success for large document collections [http://citeseer.ist.psu.edu/jones98probabilistic.html].

Ontologies are rich in semantics and can be used for deriving the context of an ambiguous term by examining its different semantics. Hypernymy/Hyponymy (parent-child relationship), antonymy (opposite relationship), meronymy/holonymy (whole-part relationship) are examples of types of semantic information modelled by ontologies.

A domain specific ontology such as Medical Subject Headings (MESH) will be used with the OHSUMED document collection [http://medir.ohsu.edu/~hersh/sigir-94-ohsumed.pdf]. In addition to this a general ontology such as WORDNET [ftp://ftp.cogsci.princeton.edu/pub/wordnet/5p apers.ps] will be used with a large general document collection such as NEWSWIRE. Ambiguous query terms will result in users being presented with possible choices of additional related terms in order to assist with query refinement. This research project will look at the impact of ontological information on the user's information seeking behaviour and on the nature of the information seeking tasks. The research project will attempt to introduce additional ontological term weightings to extend the probabilistic retrieval model into a new direction.

Conclusion

The probabilistic retrieval model has proved to be effective and widely used in information retrieval. More recently, ontologies have also proved to be useful in the area of query disambiguation. This project will explore the feasibility of combining the benefits of these two approaches that are already successful in their own right and thus producing a more effective meta search engine.

Jagdev Bhogal started her academic career at Wolverhampton University where she obtained her MPhil. In 1989, she joined the School of Computing, University of Central England, Birmingham as Senior Lecturer in Database Systems. She is currently studying for her PhD at City University, London. She can be contacted via ad893@soi.city.ac.uk.

Feature Article

"The 10 Commandments of Information Architecture"

By Jonathan Engel

Moses and Metadata

Last week information architect Jonathan Engel fell asleep after reading the Old Testament and a book on information architecture in the same sitting. As you do. While he dozed, the following dreamy conversation occurred:

Engel - - So, Moses, still railing against graven images?

Moses -- The people bow before other false gods. They worship the bells and whistles of information technology without reverence for information structure.

Engel -- What can you do? Most IT directors are gadget freaks. They'd replace the London Symphony Orchestra with a golden I-Pod if they could round-up enough earphones.

Moses - - I have new commandments to put the people on the true path of information architecture.

Engel - - I hope you accentuate the positive this time. "Thou shalt not" was too downbeat. And between you and me, commandment number seven isn't winning many converts.

Moses -- Ah, that was a transcription error. It should have read "Thou shalt not commit idolatry."

Engel—So let's hear commandment number one for structuring information.

Moses -- ***Business requirements should determine IT solutions, not vice-versa.*** You wouldn't employ a building contractor before you had your blueprints, would you?

Engel -- You haven't seen my house. How about commandment number two?

Moses -- ***Requirements for classifying content should be drawn from users and subject specialists.*** You need that bottom-up detail to balance your top-down business goal of information consolidation.

Engel -- Speaking of bottoms up, can you turn your mineral water into something more interesting?

Moses -- That wasn't my schtick. But I'll split it with you.

Engel -- OK, then how about number three?

Moses -- ***Information architecture should be component-based.*** An organization's resources can be described with a multitude of metadata tags -- by location, by subject area, by audience, by document type. That approach makes the structure infinitely expandable, yet consistent.

Engel -- So it's analogous to buying compact hi-fi components rather than a bulky music centre?

Moses -- If you say so. Ready for number four? ***Information architecture should use and support international and industry standards.***

Engel -- Say, while we're on that subject, whatever happened to the cubit?

Moses -- OK, sometimes you need to map between standards, especially with legacy systems.

Engel -- Where do taxonomies fit in?

Moses -- That's commandment number five. ***Taxonomy should be the core information structure.*** That way, you get the benefit of hierarchical inheritance. You can then refine the structure as a thesaurus with synonyms and an ontology that shows other relationships and shared properties.

Engel -- Like the Red Sea and red wine?

Moses -- You're catching on fast. Now number six is one of my favourites. ***Your information structure will evolve but top levels should remain stable.***
There should always be an obvious, existing location for future sub-headings.

Engel -- So you can have a big vision, but realize it in small steps, perhaps over several years.

Moses -- Tell me about it. Next time I journey to the Promised Land, I'm getting a good project manager.

Engel -- OK, so how about number seven?

Moses -- ***Your standard structure should be the unifying core of all knowledge management initiatives.***

Engel -- Sorry, when you said "knowledge management" I must have dozed off.

Moses – In other words, the same information structure should link content management systems, customer relationship management projects, expert location software, plus any information displays on the intranet, extranet and Internet.

Engel -- Doesn't that mean the structure has to be flexible?

Moses -- You bet your burning bush. To keep it future proof, ***the structure should be platform-neutral and display-neutral.***

Engel -- For those who are counting, that was number eight. Hey, we're on a roll. Hit us with number nine.

Moses -- **The same structure should support seamless search, hierarchical navigation and push delivery.**

Engel -- Can the same structure help integrate third-party information sources with internal content?

Moses -- You have to ask?

Engel -- So who are the chosen people to apply these metadata tags to content?

Moses -- Ideally, the content authors themselves. Which brings us to commandment number 10. **The structure should be simple enough to be applied manually.** However, a well-structured classification scheme will make automated tagging much easier to implement, and more precise. Do you want to talk about Bayesian algorithms?

Engel -- Maybe another time. So that's it? The 10 commandments of information architecture?

Moses -- Well, we had another five, but the man upstairs forgot to tag them as "commandments." We lost them in a batch delete.

Jonathan Engel is the director and chief information architect at consultants InfoArk Limited. He has built content classification schemes for Reuters, Syngenta, the NHS and other UK government bodies. He can be reached at j.engel@infoark.co.uk.

Authors

Informer welcomes contributions on any aspect of information retrieval. We are particularly interested in feature articles and opinion pieces, but are also pleased to receive news articles, book reviews, job ads, etc.

If you have an idea for an article, please contact the editors at: irsg@bcs.org.uk.

Terrier

Terrier (Terabyte Retriever) is Glasgow University's modular platform for the rapid development of Web, intranet and desktop search engines, providing indexing and retrieval functionalities. Terrier is based on the Divergence from Randomness (DFR) framework. It can index various document collections, including the standard TREC collections, such as AP, WSJ, WT10G, .GOV and .GOV2. It also provides a wide range of parameter-free weighting approaches and full-text search algorithms, aiming to offer a public test-bed for performing large-scale Information Retrieval experiments.

Terrier is written in Java. Terrier has been successfully used for ad hoc retrieval, Web search and cross-language retrieval, in a centralised or distributed setting. Currently, it is also being used for running various intranet search facilities, such as the departmental search engine at the University of Glasgow. A version of Terrier is available for download as open source software with the aim to facilitate experimentation and research in the IR community. For further details see http://ir.dcs.gla.ac.uk/terrier.

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 (http://www.bcs.org/bcs).
To automatically receive your own copy of Informer, simply join the IRSG via the BCS website (http://irsg.bcs.org/join_form.php).