Informer

Number 20, Autumn 2006
ISSN 0950-4974

BCS INFORMATION RETRIEVAL

In This Issue

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.

Those of us who are members of the BCS can't have failed to notice the attention currently being given to the subject of professionalism within IT - it features prominently in much of the new BCS promotional material, and is also given widespread coverage on the Society website. And this has got me thinking: what could this mean for the IRSG?

As many of you will know, the IRSG is one of a number of BCS specialist groups, and it is perfectly acceptable (and quite commonplace) to be a member of more than one specialist group - I, for example, am also very active within the HCI (human-computer interaction) group. And being a member of two groups, I often find it illuminating to draw parallels between the two communities. In many ways, they have much in common - both are underpinned by an established body of academic knowledge, both have a largely academic committee but count a significant number of practitioners in their membership, and both hold a major annual conference.

But what makes the HCI group different is that its practitioner members can also choose to join one of a number of professional societies specifically constituted to represent their interests - in their case the Usability Professionals Association and the Ergonomics Society (to name but two). So what then, is the equivalent professional body for people employed in the information retrieval industry? Will we ever see a "Search Professionals Association"?

To answer this question, I suppose we must first define exactly what we mean by the "information retrieval industry". First, there are those who *use* search tools and techniques as part of their job. For information professionals such as these, there are number of relevant bodies, such as ASLIB or CILIP. And then there are those who "adapt" search tools for some commercial advantage, for example by manipulating search engine rankings so that their clients' websites appear as near to the top as possible for given keywords. For these and related marketing professionals, SEMPO is probably the most suitable trade association.

But what of those who actually *design* or *develop* search tools and techniques? Which professional body is best placed to represent their interests, and the hundreds of associated professionals who either work for the major search companies, or related intermediaries, startups and VARs?

To a degree, you could argue that that is the role of the IRSG itself - but if that is the case, then I think we could (and should) be doing far more to reach out to such people. But if it is not the IRSG, then who? Perhaps we could fall back to saying that the BCS as a whole is the logical candidate. But to me, that doesn't feel quite right: firstly, there are many people in the search industry who aren't IT professionals, and secondly, the BCS is, well, exactly that: a computer society, not a search society.

Of course, another interpretation is that the search industry simply doesn't need a professional association (yet) - after all, it is currently dominated by a number of proprietary players, all with different (often patented) approaches and techniques, and that is a very different environment to (say) the usability industry, which relies on shared goals such as the development of a common skills framework to facilitate professionalism and mobility across organisations.

So, back to the question: will we ever see a "Search Professionals Association"? The short answer, it seems, is not for a little while yet - but I'd like to think that if ever we do get close to that point, then the IRSG will be leading the way in helping to shape its identity and purpose, rather than watching from a distance.

In the meantime, if you have other ideas, or would like to contribute to the debate, drop us a line at irsg@bcs.org.

All the best,
Tony Rose
Informer Editor and Vice chair, IRSG
Email: irsg@bcs.org.uk

Feature Article:

The Fall and Rise of Collaborative Filtering

By Paul Matthews

"When men exercise their reason coolly and freely on a variety of distinct questions, they inevitably fall into different opinions on some of them. When they are governed by a common passion, their opinions, if they are to be called, will be the same."
Alexander Hamilton (1755 - 1804)

"My Tivo Thinks I'm Gay"
Wall Street Journal, November 2002

Collaborative filtering (CF) is the mechanism behind recommendation web sites and the "you might also like.." features in e-commerce. IN CF, the behaviour and preferences of real people are used to predict your own taste and select books, films, music, or any other resource for you. CF has had something of a chequered history. After an initial surge of interest in the late nineties and early 00s and some promising (and some famously off-track) pilots, CF is now being remoulded in a web 2.0 guise, and is getting the backing of some of the big players in the community and search business. This article gives you a whirlwind tour of the principles and theory of CF, looks at some of its highs and lows and visits some snazzy web sites that are just waiting to recommend your next reading, viewing or listening material.

Some IR people might be wondering at this point just what CF has to do with IR. After all, IR is all about machine indexed content and retrieval based on a known search string isn't it? Movie recommendations are surely another animal? Yes, this must be partly true. IR is about "something I would like to know" whereas CF is about "something I would like". IR does rely more on automatic indexing and characterisation of information, whereas human preferences are central to CFs choices. But it doesn't take long to realise that some of our most successful search engines and relevant results have an element of CF in them. Central to the much celebrated PageRank itself is link density, which is a reflection of what a web master has judged to be a useful site. This type of relevance boosting and the continued rise of communities with shared interest mean that today, CF and IR are ever more interconnected. Moreover, CF also straddles the fields of AI, HCI and psychology.

CF encompasses a range of techniques that focus on various aspects of the user and their known preferences. The "classic" technique is to look at the user – item graph and identify similar users based on items that have been similarly ranked. Recommendations are then generated based on rankings from those similar users (for items that you don't have) . Another approach is to use a clustering algorithm to group users according to various interests and then recommend based on this identification of "like minded" peers. A third approach, popularised by Amazon, is to only look at the item level and build proximity lists based on items that have appeared together in shopping baskets in the past.

Each method has pros and cons, and many require fairly weighty computations on often quite sparse data. One advantage of the Amazon method was that it enabled more dynamic recommendations without the need to wait for the recommendation "job" to run. A further advantage of this type of technique is that it records implicit preferences based on user behaviour and did not require "training" by asking to rank some specimen data. We will see some more about how implicit data is being used later, but for now we will dip into CF's closet to view the skeletons..

The relative ease of collecting base data and applying the algorithms perhaps led to overconfidence amongst touters of recommendation technology. This led to some quite infamous examples of mis-recommendations. The "my Tivo thinks I'm gay" example was based on the digital video recorder that automatically records shows "you might also like". This led in one users experience to the machine recording gay porn, a topic that he wasn't actually that keen on. Interestingly, one theory on this is that Tivo users are skewed towards homosexual males, being just the type of affluent gadget fans that might own a machine. Or maybe the Tivo algorithm just wasn't that good.

Was the "my Tivo" moment when the research community started to lose interest in CF? There certainly does seem to have been a dwindling of interest in the topic at around this time (see graph). Perhaps it was felt that CF was failing to live up to its early promise.

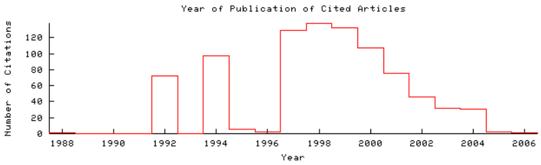
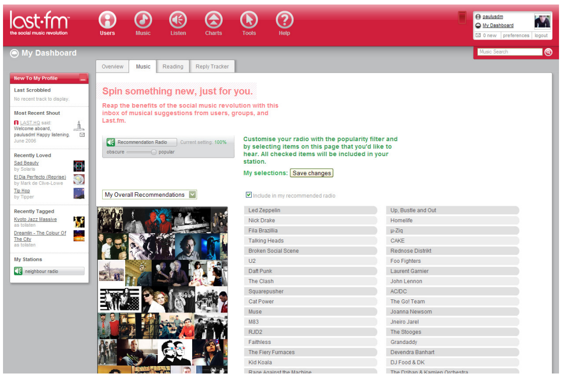


Figure 1: Citations for "collaborative filtering" on Citeseer

But like many things in technology, CF was only biding its time and awaiting resurgence. It has certainly found this in the phenomenon of Web 2.0, a rather loose term that brackets a range of new "social" web applications. In the typical Web 2.0 scenario, a range of like minded users share resources, tag and label the content and rate it in some way. This has provided a perfect platform for CF. Perhaps a slight difference this time around is a preference for simpler ranking methods and for more transparency over the method and confidence of prediction. This glasnost is typical of Web 2.0 and means that some very nice tools are available for the researcher as web APIs have become a "must have" on Web 2.0 sites.

The collection of implicit preference data has enabled some of the more successful and easier to use modern recommender engines, such as last.fm. This music community and recommendation site uses a plug-in for your music player – the brilliantly titled "audioscrobbler" – which sends your listening preference to the site as you enjoy it. This wealth of data enables the generation of recommendations, which you can then listen to as a custom made radio station. Perhaps of equal importance are the community aspects which enable you to identify and communicate with like minded users and groups. The purely manual recommendations of these sources, once identified, can be as valuable as the automatic ones.



My Dashboard

New To My Profile

Last Scrobbled
No recent track to display.

Most Recent Shout
LAST.HQ said:
Welcome aboard, paulusdm! Happy listening.
June 2006

Recently Loved
Sad Beauty
by Solaris
El Dia Perfecto (Reprise)
by Mark de Clive-Lowe
Tip Hop
by Tipper

Recently Tagged
Kyoto Jazz Massive
as tolisten
Dreamlin - The Colour Of The City
as tolisten

My Stations
neighbour radio

Spin something new, just for you.

Reap the benefits of the social music revolution with this inbox of musical suggestions from users, groups, and Last.fm.

Recommendation Radio Current setting: 100%
obscure popular

Customise your radio with the popularity filter and by selecting items on this page that you'd like to hear. All checked items will be included in your station.

My selections: Save changes

My Overall Recommendations

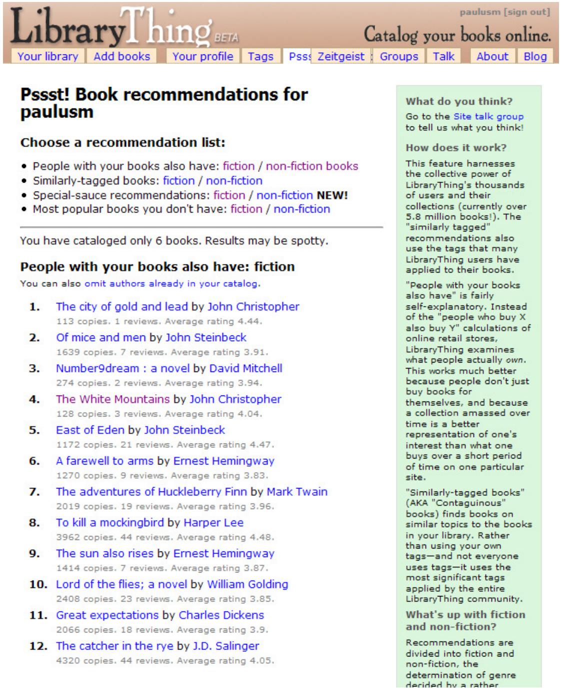
☑ Include in my recommended radio

Led Zeppelin
Nick Drake
Fila Brazillia
Talking Heads
Broken Social Scene
U2
Daft Punk
The Clash
Squarepusher
Cat Power
Muse
M83
RJD2
Faithless
The Fiery Furnaces
Kid Koala
Rage Against the Machine

Up, Bustle and Out
Homelife
µ-Ziq
CAKE
Rednose Distrikt
Foo Fighters
Laurent Garnier
John Lennon
AC/DC
The Go! Team
Joanna Newsom
Jneiro Jarel
The Stooges
Grandaddy
Devendra Banhart
DJ Food & DK
The Dzihan & Kamien Orchestra

Figure 2: Last FM: The more you scrobble the better it gets! (Hmm.. AC/DC?)

Perhaps a more traditional CF approach is taken my Librarything.com for books. Based on the user building a catalogue of favourite books, the site generates recommendations for you. What strengthens this site, and doubtless the quality of its recommendations is the large user base and catalogue (nearly 6 million books)



Pssst! Book recommendations for paulusm

Choose a recommendation list:

- People with your books also have: fiction / non-fiction books
- Similarly-tagged books: fiction / non-fiction
- Special-sauce recommendations: fiction / non-fiction **NEW!**
- Most popular books you don't have: fiction / non-fiction

You have cataloged only 6 books. Results may be spotty.

People with your books also have: fiction

You can also omit authors already in your catalog.

1. The city of gold and lead by John Christopher
113 copies. 1 reviews. Average rating 4.44.
2. Of mice and men by John Steinbeck
1639 copies. 7 reviews. Average rating 3.91.
3. Number9dream : a novel by David Mitchell
274 copies. 2 reviews. Average rating 3.94.
4. The White Mountains by John Christopher
128 copies. 3 reviews. Average rating 4.04.
5. East of Eden by John Steinbeck
1172 copies. 21 reviews. Average rating 4.47.
6. A farewell to arms by Ernest Hemingway
1270 copies. 9 reviews. Average rating 3.83.
7. The adventures of Huckleberry Finn by Mark Twain
2019 copies. 19 reviews. Average rating 3.96.
8. To kill a mockingbird by Harper Lee
3962 copies. 44 reviews. Average rating 4.48.
9. The sun also rises by Ernest Hemingway
1414 copies. 7 reviews. Average rating 3.87.
10. Lord of the flies; a novel by William Golding
2408 copies. 23 reviews. Average rating 3.85.
11. Great expectations by Charles Dickens
2066 copies. 18 reviews. Average rating 3.9.
12. The catcher in the rye by J.D. Salinger
4320 copies. 44 reviews. Average rating 4.05.

What do you think?

Go to the Site talk group to tell us what you think!

How does it work?

This feature harnesses the collective power of LibraryThing's thousands of users and their collections (currently over 5.8 million books!). The "similarly tagged" recommendations also use the tags that many LibraryThing users have applied to their books.

"People with your books also have" is fairly self-explanatory. Instead of the "people who buy X also buy Y" calculations of online retail stores, LibraryThing examines what people actually *own*. This works much better because people don't just buy books for themselves, and because a collection amassed over time is a better representation of one's interest than what one buys over a short period of time on one particular site.

"Similarly-tagged books" (AKA "Contaguinous" books) finds books on similar topics to the books in your library. Rather than using your own tags—and not everyone uses tags—it uses the most significant tags applied by the entire LibraryThing community.

What's up with fiction and non-fiction?

Recommendations are divided into fiction and non-fiction, the determination of genre decided by a rather

Figure 3: Library Thing will supply your holiday reading list for the next 500 years

We are also now seeing some innovation with interfaces. Liveplasma.com's flash display lets you visualise the interrelated resources and navigate to new discoveries.

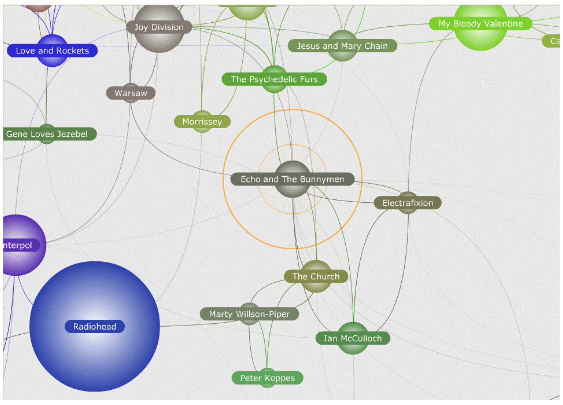


Figure 4: Live plasma's graphical view of connectedness

While these examples are mainly using more "traditional" CF methods, rather more bog-standard ranking is behind popular news sites such as digg.com, reddit.com and now even netscape.com. Users simply vote on a story to bump it up the ranking (though in actual fact users suspect that rather darker arts can be behind the eventual order of stories!) . An interesting twist is that reddit.com awards "karma" points for users who post the most highly valued stories.

The fact that Netscape have launched a CF-based news portal is indicative of the interest currently shown by some of the large players in the search and portal business. Yahoo and Microsoft are also showing great interest in incorporating CF into their products. New algorithms are being sought which combine the strengths of traditional IR with preference-based search. User behaviour such as click-through as well as explicit recommendations by fellow members of a community will be mined to improve the search experience.

"we really haven't had another breakthrough [in search technology] for some time now, until social search."
Bradley Horowitz, vice-president of advanced products at Yahoo, April 2006

But to dwell on search is to forget the pure unadulterated joy of a good recommendation.

The quest in the CF field had been to emulate the quality of recommendation you'd get from a trusted friend. It has taken time, but we are starting to see this. Importantly, Web 2.0 has shown us that building a community of trusted strangers (and enabling them to become friends too) is a great way to do it.

Paul Matthews is currently Knowledge Management IT Specialist at the Overseas Development Institute, with interests including information management, collaboration and ICT for development. Contact: p.matthews@odi.org.uk

Get Involved!

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, jobs ads, etc.

Right now we are running a series of Product Reviews, so if you are interested in reviewing any of the following:

- Copernic
- Ask Jeeves Desktop Search
- Blinkx
- MSN Search Toolbar

Then please get in touch with us via irsg@bcs.org.uk. All of the above are freely available as software downloads.

BCS-IRSG Announcements from the Chair

by Leif Azzopardi

Great news for IRSG members: the BCS-IRSG has formalized its relationship with the ACM Special Interest Group on Information Retrieval (ACM-SIGIR) with an 'in cooperation' agreement for the International Conference SIGIR. This agreement acknowledges the long-standing relationship between the BCS-IRSG and ACM-SIGIR, which began when both groups jointly ran the initial SIGIR conferences almost thirty years ago now. Due to this agreement, IRSG members will now be able to register to subsequent SIGIR conferences at significantly reduced rates. The SIGIR Mentoring programme deadline is 15th of November, 2006 for those who would like help with their submission, while the full paper deadline for SIGIR 2007 is the 28th of January, 2007.

The BCS-IRSG has also formed links with the AIRS Steering Committee, who run the Asia Information Retrieval Symposium (AIRS). AIRS aims to bring together international researchers and developers to exchange new ideas and the latest achievements in information retrieval. The scope of the symposium covers applications, systems, technologies, and theoretical aspects of information retrieval in text, audio, image, video, and multimedia data. The third **Asia Information Retrieval Symposium (AIRS2006)** is to be held this month in Singapore. The event is organized by **Institute for Infocomm Research** and co-organized by **National University of Singapore**. The BCS-IRSG is happy to announce its support for the symposium and look forward to continuing our support in the future.

Another BCS-IRSG supported event will be held in Glasgow this month, the **Symposium on String Processing and Information Retrieval** (SPIRE2006), which is organized by the **University of Strathclyde**. Whilst in

Copenhagen, Denmark, the **First Symposium on Information Interaction in Context** (IIiX2006) will be held. Recently, Mounia Lalmas and Anastasios Tombros from Queen Mary University of London in conjunction with the BCS-IRSG submitted a bid to host the next IIiX Symposium in 2008. We are happy to announce that this bid was successful and shall be held in London in 2008.

JOINT OSSG-IRSG EVENT: Next Month, there will be a joint event between the **Open Source and Information Retrieval Specialist Groups,** which will be held on the 21st of November, 2006 running from 6pm-9pm at the BCS-HQ in London. Andy MacFarlane from the IRSG and Richard Boulton from Lemur Consulting will present their views on Open Source software development for Information Retrieval software. **Refreshments and buffet will be provided and registration is free**. To register for the event, call the BCS-HQ on 01793 417417 or email Mark Elkins (mark_elkins at bcs.org). The event will be held at BCS Central London Offices, First Floor, The Davidson Building, 5 Southampton Street, London WC2E 7HA.

TLIR 2007 WORKSHOP: Also, the BCS-IRSG is running the **first international workshop on Teaching and Learning Information Retrieval** (TLIR2007) on the 10th of January, 2007 in London also at the BCS-HQ. The aim of this workshop is to create a common space where IR lecturers and researchers can share their experiences and opinions in the field of IR teaching at different levels of educational.

ECIR 2007 CONFERENCE: The next European Conference in Information Retrieval will be held in Rome during April, 2007 (ECIR2007). Just a reminder, the deadline for submissions is the 30th of October, 2006 for full papers and the 6th of November for poster papers. Also, the ECIR Workshop report is now available on line, which provides helpful guidelines for authors and reviewers (see the ECIR report).

Finally, congratulations to Keith van Risjbergen, the winner of the 2006 Gerard Salton Award presented during this year's SIGIR, for his significant, sustained and continuing contributions to research in the field of information retrieval.

Leif Azzopardi is a Research Fellow at the University of Strathclyde in Glasgow, UK. His research interests include: formal models for information retrieval, distributed information retrieval and evaluation of information access systems. He can be contacted by email via: leif.azzopardi@cis.strath.ac.uk.

Research Update:

Searching for people in the personal workspace

By Krisztian Balog

Some of the most valuable knowledge in an organization resides in the minds of its employees. Enterprises must combine digital information with the knowledge and experience of employees. Organizations may have many valuable experts who are dispersed geographically. Sharing knowledge can prevent them from reinventing the wheel, help them deliver resources, and support collaboration no matter where their people are located. The most effective way to exchange knowledge is human contact. Still, finding the right person to get in contact with is something where information technology can add value.

Computer systems that augment the process of finding the right expert for a given problem in an organization are becoming more feasible, thanks to the role of computer-based collaborative systems, an area which has seen significant growth recently.

The organization's internal and external websites, e-mail and database records, agendas, memos, logs, blogs, and address books are all sources of information to which people are connected in their work space. This *"personal work space"* covers the electronic data held by the organization. Within this setting it is natural to look not only for documents, but for entities: answers, services, objects, … **people**!

The main focus of my PhD research is to investigate methods, and techniques for these kinds of *"people search"* tasks.

People Search Tasks

We assume an organization to have a sufficiently large amount of textual content available in electronic form. This comprises heterogeneous document repositories containing a mixture of document types. People are a critical organizing element in workspace information, and an important retrieval cue for searching in this environment. There is much interest in people search both from a practical point of view and from the research community. This fact is witnessed by numerous recent publications on finding experts, and on the recent introduction of an expert finding task at Text Retrieval Conference (TREC).

Although TREC has provided a common platform to empirically assess methods devised for expert finding, evaluation is still a partially resolved issue. There is no collection that contains personal workspace data from any organization available for research purposes. Obviously, privacy is a big concern and has to be properly dealt with.

I. Expert finding

Expert finding addresses the task of finding the right person with the appropriate skills and knowledge: *"Who are the experts on topic X?"* For instance, an employee wants to ascertain who worked on a particular project to find out why particular decisions were made without having to trawl through documentation (if there is any). Or, they may require a highly trained specialist to consult about a specific problem. Identifying experts may reduce costs and facilitate a better solution than could be achieved otherwise.

"The most effective way to exchange knowledge is human contact"

We propose two models, based on probabilistic Language Modeling techniques, for accomplishing this task. Each model ranks candidates according to the probability of a candidate being an expert given the query topic, but the models differ in how this is performed.

In *Model 1* we start from the candidate and consider the documents with which he or she is associated. That is, we build a textual representation of individuals, based on documents that a candidate is associated with.

In *Model 2* we look at the documents that best describe the topic, and then look at people that are most strongly associated with these documents. Under this model we can think of the process of finding an expert as follows. Given a collection of documents ranked according to the topic (e.g. results of a search engine), we examine each document and, if it is relevant to our problem, we subsequently see who is or are associated with that document.

II. Expert profiling

The next natural task is to turn the expert finding task around: *"What does expert X know?"* Such profiling of an individual involves the identification of types and areas of skills and knowledge, and an evaluation of levels of proficiency in each. That is the candidate's *topical profile*.

In most cases, industrial practice still employs a database-type structure of skills and knowledge for each individual in the organization. Our aim with ongoing work is to put automatic methods into practice, and to reduce the human effort associated with maintenance. The knowledge of individuals is represented as a Language Model (LM) for each person, which is based on extracted and merged information from various sources. Then the person's competence is estimated by determining the distance between a candidate's and the knowledge area's LMs. The knowledge area's LM is estimated from the documents relevant to that area. Relevance is obtained by standard Information Retrieval methods.

"People are a critical organizing element in workspace information"

When people search for expertise, they are often looking for experts, but not in isolation. Context and evidence are needed to help users of expertise finding systems to decide whom to contact when seeking expertise in some area. *Who does she work with? What are her contact details? Is she well-connected, just in case she is not able to help us herself?* Collaborators, colleagues, co-authors, affiliations, etc. are all part of the person's *"social profile"*, and can serve as a background, or context, in which the system's recommendations should be interpreted. This collaboration network can also help us to look beyond individuals, and to explore the connections, spheres of influence, and roles of people within an organization.

The techniques we have developed so far have proven efficient for both the expert finding and topical profiling tasks. Initial results have confirmed that these tasks are two different views on the same data. While on the one hand, the same methods can be applied to both tasks, the intuition behind the users' information need differs in the two cases. Therefore both the expert finding and the profiling tasks should be addressed separately.

The visualisation of profiles and responses to queries is important in the final solution, and a prototype of such a user interface has been developed (for screenshots see http://staff.science.uva.nl/~kbalog/).

Krisztian Balog is a PhD student at the Information and Language Processing Systems group of the University of Amsterdam, under the supervision of Prof. Dr. Maarten de Rijke. He holds M.Sc. degrees in Computer Science from Vrije Universiteit, Amsterdam and from Eötvös Loránd University, Budapest. His research interests include: Intelligent Information Access, Information Retrieval, and Language Modeling. He can be contacted by e-mail via: kbalog@science.uva.nl

Book Review:

Intelligent Document Retrieval, by Udo Kruschwitz

Reviewed by Andrew Neill MBCS

This book is an extension of PhD work by Udo Kruschwitz whilst at the University of Essex (UK). Split into two main sections, it covers existing theory and work on search enhancement and document categorisation, and then leads into a discussion of the author's work on creating new tools for categorising collections of partially-structured documents such as can be found in many academic and corporate systems. It is aimed not at the web in general, but at smaller document collections with a more limited range of topics. The goal is to create an enhanced search tool that automatically categorises document collections, which is of interest to those who work in the Natural Language Processing and Information Search and Retrieval areas.

Kruschwitz begins by defining the problems with existing domain model analysis and classifications, looking over the areas of web search, automated clustering, and manual classification, and including ontologies (Kruschwitz quotes researcher Spark Jones as describing ontolology as a fashionable term for "structured classifications and thesauri") and similar attempts to build a pre-defined hierarchy for the semantic web.

The author quickly differentiates from others by stating that the goal of the research is to "ideally, [create] a domain model on the fly in an automated fashion without assumptions about the documents' content". This is a bold aim, and has several advantages where are described and explored in the review of the existing literature.

Whilst accepting that the goal of automatically acquiring domain models from collections of documents is not new, the main focus of research has been on word co-occurrences and/or linguistic information. Indeed, your reviewer can recall homework assignments to calculate the information value of particular terms, given their frequency within a document collection. Kruschwitz points out that less work has been published on extracting meaning (i.e. the semantic content) by exploiting the markup of the documents. The key advantage of this approach is that the process of extracting semantic concepts from documents can be performed without any a priori knowledge of the domain, or even an understanding of the language in which the documents are written!

"The goal is to create an enhanced search tool that automatically categorises document collections"

The author proposes and develops a system that will automatically process a document collection for concepts, and use these concepts to restrict or relax a search query based on a dialogue with the user. Kruschwitz defines a concept in a document as being noun-based terms that appears in N or more separate markup contexts within the document. For example, where N = 2, a term that appears both in the title and the document metatags (two different markup contexts) is a concept within that document.

Note that no judgement or estimation of the value of that term is required – it just exists, for good or ill. There is also no need for manual input to define the collection of concepts, and yet the concepts are relevant within the context for the document collection. This is in contrast to carefully created hierarchies of domain-independent knowledge, where relevancy cannot be guaranteed.

Two documents are related if they both contain the same concepts, or related concepts. More specifically, two documents are explicitly related if they contain the same concepts. Documents may also be implicitly related if they contain concepts that are related, for example by both existing in a third document within the collection.

By defining concepts and relatedness in this way, the author has abstracted a powerful tool for assembling a collection of concepts, and defining relationships between these concepts, within a set of documents, without requiring domain knowledge of the documents (beyond some ability to parse the formatting structure – i.e. the "partial structure" of the documents). The system can be supplemented by any additional information that is available – for example, if the document collection contains a classification already (such as in the author's example of a classified directory of advertisements) then this can be incorporated where appropriate. However, the goal is always to keep the system domain agnostic, and able to cope with any new documents and concepts as they are added.

Kruschwitz also describes how the dialogue with the user is used to improve the search experience. Whilst noting that 85% of users on web search systems never pass the first results page, and 77% users only use a single search query, Kruschwitz hopes that, by prompting user with related concepts that exist within the results, the user can be encouraged to add or remove concepts in order to improve the accuracy and usefulness of the information returned.

"this is an interesting, intelligent book… accessible and relatively easy reading"

The latter part of the book looks at three different prototype practical applications of the system, and examines the indexes, domain models, and user interfaces for each. The three examples use the algorithms to index the University of Essex web site, the BBC News web site, and YPA, a classified directory web site. Comparisons are made between using the site-specific function of Google, and a concept-enhanced search engine built by the author. Often, there was no significant difference between the two engines in terms of speed, accuracy and the like – but feedback from users was strongly in favour of the search interface which suggested refinements and relaxation options, as they found it much easier to use and better overall compared to a simple Google-type interface.

Overall, this is an interesting, intelligent book, which, despite its academic roots, is accessible and relatively easy reading. Some knowledge of logic expressions is useful, but these are adequately explained in the accompanying longhand text so the reader can grasp the meanings. The editing and production is of high quality, and references and indexes are plentiful. Future research areas would include more investigation of the user interface design – and there is probably a PhD (and a vast fortune!) for someone who perfects this sort of system for an enterprise search product. Highly recommended!

Andrew Neill is the Business Analyst for international city-based law firm Norton Rose. Prior to this, he worked in the technology and integration division of Deloitte & Touche Business Consulting, working on a wide variety of technology projects for blue-chip companies. Andrew also recently completed an MSc in Computing for Industry at Imperial College, and has been a member of the British Computing Society for 2 years. He specialises in search, knowledge and content management systems, enterprise architecture, and business process improvement. Outside of work, his interests include languages and travelling, and he and his new wife like to spend as much time as possible enjoying friends' hospitality, particularly in Spain, Italy and Greece.

ECIR Workshop Report

by Leif Azzopardi, Andy MacFarlane and Iadh Ounis

On the 20th of June, the BCS-IRSG held a workshop to discuss the organization and reviewing process and procedures for our annual European Conference in Information Retrieval (ECIR). Given the success of ECIR 2006 over previous instances of the conference (177 paper submissions to ECIR 2006 compared with 124 submissions to ECIR 2005), it was felt that the management of the reviewing process should be revised, and a decision needed to be made on which method to use – status quo, meta-reviewing or sub programme committees (Sub-PCs). About 20 participants attended the workshop.

With the quantity of submissions also comes the problem of ensuring quality; in terms of the submissions themselves, the reviews, and ultimately the final programme. To ensure and maintain the high quality of the conference, it was felt that establishing a set of guidelines for ECIR would be helpful to both authors wishing to submit and referees reviewing papers. We dedicated a session on discussing what would make a good ECIR paper given some of the different paper genres.

"With the quantity of submissions also comes the problem of ensuring quality"

The aims of the workshop were two fold: (1) to review current practice of reviewing and organization at ECIR and (2) develop a set of guidelines to help authors wishing to submit to ECIR and also to aid referees reviewing for ECIR.

With respect to the organization of the conference, the workshop participants discussed whether ECIR should consider changing the reviewing process and its structure. In particular, whether we should consider the introduction of meta reviewing or the use of sub programme committees (sub-PCs). The reasons to consider changing the current structure (i.e. pc-chair and reviewers) and opt for a more sophisticated approach is to be able to cater for the higher volume of submissions that ECIR now receives and ensure that the process maintains the quality of reviews, whilst being open and transparent.

After much discussion and debate, it was decided that at ECIR 2007 the use of sub programme committees for reviewing would be trialed. Sub-PCs are smaller groups of PC members within the PC, which review a set of papers on a particular topic. Each sub-PC member provides a ranking of the papers to the PC-Chair, who ranks the papers reviewed by the sub-PC and then calibrates the disparate sub-PCs, to select the final programme. It is anticipated that a better rating of papers will be obtained because all the reviewers of a sub-PC have reviewed the same set of papers – and will judge them relative to each other. Another benefit of using sub-PCs is that it can be employed without requiring any extra time in the reviewing process (unlike meta reviewing) and can be employed seamlessly within current practices. The usage of sub-PCs will be reviewed after ECIR 2007 to consider whether future ECIRs should also use this process.

However, most of the workshop was dedicated to developing guidelines for authors and reviewers of ECIR. We invited five speakers to provide their thoughts on what makes a good IR paper in a particular area. These presentations formed the core content for the guidelines and have been combined into a set of draft guidelines. These guidelines focus on several different categories of papers: theoretical (presented by Keith van Rijsbergen), conceptual (contributed by Mounia Lalmas), User studies / People in IR (presented by Ian Ruthven), applications and prototypes (presented by Ayse Goker) and experimental and system comparison (presented by Iadh Ounis). A definition of each category and its requirements was formulated based on the presentations and the ensuing discussion. An example of the requirements for a theoretical paper is shown below.

Example: Theoretical Papers

What is a theoretical paper?
A theoretical paper proposes a theory for Information Retrieval (or some phenomena within the domain). A theoretical paper should

present a supposition or system of ideas intended to explain some phenomena within IR. It should be based on general principles independent of the phenomena to be explained (i.e. Darwin's Theory of Evolution). It could provide a set of principles on which the practice of an activity is based (i.e. a theory of information seeking behaviour). Or it could present an idea used to account for a situation or justify a course of action. Consequently, this does not necessarily imply that the theory is grounded in mathematics or some other formalism, which is a common misconception about theoretical papers in IR. Instead, a theoretical paper may also be discursive in nature, providing arguments and reasoning through the discourse.

What makes a good theoretical paper?

First, the paper must go beyond the existing theory already present in the literature – and thus fulfill the originality criteria. In order to convince the reader that this is the case there should be links to older theory to provide the context of the paper. The relationship between the old and the new should be related and explained.

Second, it is important for a theoretical paper in IR to provide the necessary contextualisation of the theory within IR. That is, what is the relevance of this theory to IR? Consequently, the generic application of a machine learning approach, for example, is not relevant. The burden is on the writer to example the link between the theory and the practice, given the domain.

Third, the clarity of the presentation is very important because the emphasis of the paper is to present an account for a phenomenon. Consequently, the arguments presented need to be clear and justified. One way of ensuring clarity is to provide illustrative and practical examples to aid the reader's understanding.

Fourth, a theoretical paper aims to link theory with practice; once a theory is presented, the inevitable question arises; does it work in practice? However, "proof" that a theory holds is not a necessary requirement for a theoretical paper to be acceptable, as it is not always possible for a theory to be put forward and for it to be tested to the nth degree. There are various reasons for this; the work is in its early stages; the machinery doesn't exist for it to be tested; etc. In such cases when experimental work can not be provided to ensure that the paper is acceptable, there are other criteria that the paper should meet. A discussion should be included about the testability of the theory present, comments on whether it can be falsified, how the theory could be tested in practice, its tractability, its relationship with experimentation, and whether it is possible to implement it or not. Addressing such issues is paramount to papers, which present novel/new theory.

However, there are cases when the theory presented is an extension to the existing theory. In this case, where the theory has been tested previously, it is necessary to provide some experimental work in order to show that this extension is actually significant, useful, successful etc. Re-stated, delta theory papers should provide some empirical testing. On the point of significance, a theoretical paper should also discuss what would constitute a significant result and how to quantify this.

These guidelines are aimed at explaining some of the main criteria or properties a good ECIR paper in a particular genre should possess and what an author should aim to achieve. However, it should be pointed out, that these guidelines should be interpreted in the spirit in which they were written, to provide a helpful and informative list of properties that a good paper should possess. However, these guidelines are not entirely complete, or comprehensive, nor are they cast in stone. Consequently, good ECIR papers may possess other attributes or qualities that are out with these descriptions and it is up to the referees to identify such instances. Through providing such guidelines, we hope to ensure and increase the quality of the submissions and the conference; specifically helping students and first time submitters to ECIR to know some of the expectations and qualities required for a good ECIR paper. Also, the guidelines serve as way to show what types of papers are appropriate and acceptable for the conference.

The workshop proceedings are available for download from the BCS-IRSG website, along with the draft guidelines for writing a good ECIR paper. The draft guidelines, we hope, will provide a useful resource to authors wishing to submit to ECIR2008 and onwards, and referees reviewing for ECIR2007 and onwards. Unfortunately, the guidelines were not finalized for release in time for this year's ECIR, but this gives referees the chance to provide their feedback on the guidelines before ECIR 2008. If you have any comments or questions about the draft guidelines please forward any enquires to myself (IRSG Chair), Iadh Ounis (IRSG ECIR Coordinator) or Andy MacFarlane (IRSG Secretary). We would be happy to try and incorporate any suggestions within the next version of the guidelines.

Finally, the BCS-IRSG would like to thank the Department of Computing Science, University of Glasgow, for hosting the meeting and all the survey and workshop participants for their comments and contributions.

Forthcoming Events

Edited By Andy MacFarlane

Intelligent processing and the information lifecycle - lessons for Data Protection Act compliance
Monday 11 December 2006, 5.30pm for a 6.15pm start, London. This lecture will consider how achieving compliance with the DPA requires an 'intelligent' processing environment where the information lifecycle is properly understood. Booking is essential. £10 for BCS members and £15 for others.
http://www.bcs.org/events/dataprotection

5th International Conference on Natural Language Processing (ICON 2007)
IIIT, Hyderabad, India, 4-6 January 2007. An NLP conference which will be of interest to many IR researchers and practitioners working in the area of natural language.
http://ltrc.iiit.net/icon2007/showfile.php?filename=icon2007.php

IJCAI 2007 Workshop on Analytics for Noisy Unstructured Text Data. International Workshop on Artificial Intelligence and Music (MUSIC-AI 2007).
Both are part of IJCAI2007. Hyderabad, India, 6th – 12th January 2007. What do we do about retrieving messages, wiki's items, blog postings etc. IJCAI 2007 workshop deals with such issues. MUSIC-AI deals with issues in music IR.
http://research.ihost.com/and2007/ and
http://www.iua.upf.es/mtg/MusAI/

First International Workshop on Teaching and Learning of Information Retrieval (TLIR'07)
Covent Garden, London, 10th January 2007. The first IRSG workshop on the issue of teaching and learning the subject of information retrieval and search.
http://tlir07.soi.city.ac.uk/

The 11th International Conference on Database Theory (ICDT 2007)
Barcelona, Spain, 10-12 January 2007. A general theoretical database conference with a theme on IR.
http://www.lsi.upc.edu/~icdt2007/

International Conference on Computing: Theory and Applications
Kolkata, India, 5-7 March 2007. A conference with many themes of interest including information retrieval, question answering and digital libraries.
http://www.isical.ac.in/~iccta/

Special Track on: INFORMATION ACCESS AND RETRIEVAL, 2006 ACM Symposium on Applied Computing (SAC 2007)
Seoul, Korea, 11th – 15th March 2007. A big ACM conference, with a track on IR.
http://www.cis.strath.ac.uk/external/SAC2007/

Collaborative Knowledge Management (CoKM2007)
Potsdam, Germany, 28-30 March 2007. A general Knowledge Management conference.
http://www.wm-tagung.de/CoKM2007/

7th Dutch-Belgian Information Retrieval Workshop (DIR 2007)
Katholieke Universiteit Leuven, Belgium, March 28-29, 2007. This workshop provides an excellent meeting place for information retrieval researchers to exchange and present innovative research developments.
http://law.kuleuven.be/icri/liir/dir2007/

29th European Conference on Information Retrieval (ECIR 2007)
Rome, Italy, 2-4 April 2007. The groups annual conference held at Fondazione Ugo Bordoni.
http://ecir2007.fub.it/

IEEE 3rd International Workshop on Web Personalisation, Recommender Systems and Intelligent User Interfaces (WPRSIUI'07)

Istanbul, Turkey, 16-20 April 2007. A conference on recommender Systems.
http://www.doc.ic.ac.uk/~gu1/WPRSIUI/WPRSIUI07/index.html

ITNG 2007 Web Technologies Track, 4th INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: NEW GENERATIONS
Las Vegas, Nevada, USA, 16-19 April 2007. A web conference with a theme on web search.
http://www.itng.info/

23rd IEEE International Conference on Data Engineering (ICDE 2007)
The Marmara Hotel, Istanbul, Turkey, 17-20 April 2007. A 'data engineering' conference with a theme on IR.
http://www.srdc.metu.edu.tr/webpage/icde/

Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2007)
Rochester, New York, U.S.A. 22-28 April 2007. A major linguistics conference with a theme on NLP for information retrieval.
http://www.cs.rochester.edu/meetings/hlt-naacl07/

The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
Amsterdam, The Netherlands, 23-27 July 2007. This is the premier IR conference, which is proudly supported by the BCS-IRSG. Now BCS-IRSG will receive discounts to this event, so it is not to be missed!http://www.sigir2007.org

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.