Informer

Number 16, Autumn 2005
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.

Unless you've been hibernating for the past couple of months you can't have failed to notice the coverage for Industry Day 2006, which, as you may know, is a bit of a first for the IRSG (although the idea itself is quite well established among other BCS groups, e.g. HCI in particular). We're currently at the "call for abstracts" stage, which, as you might imagine, is the point where we invite IR practitioners to send us an abstract with a view to giving a 30-minute talk at the event. Of course, part of the motivation behind this event is to maximise the synergy with ECIR itself, and to that extent we welcome submissions in the form of full papers too. If you'd like to be part of this event, all you need do at this stage is send us an abstract by December 2nd. Further details on the submission process and the event itself can be found on p11 and on the ECIR website.

In the meantime, I'm pleased to introduce Alex Bailey as our product reviewer for this issue, with his thoughts on Yahoo Desktop Search. And if, like me, you're one of the many who have yet to install any type of desktop search product, you'll find this article quite timely – I for one will no longer be assuming that Google is the automatic first choice.

Also examined in this issue is "Web Dynamics" by Levene & Poulovassilis. As I mentioned in the last editorial, book reviews are set to become a regular feature in Informer, so our appreciation goes to Paul Matthews for kicking off the series with such a well-written piece.

Incidentally, does anyone actually read editorials anymore? Having announced in the last issue that various new IR books would be sent free of charge to guest reviewers, my first concern was how to deal with the inevitable deluge of applicants. Polite replies along the lines of "*I'm sorry but due to overwhelming demand …*" were half drafted in my head. However, it soon became apparent that the number of people queuing up for the literary

equivalent of a free lunch was somewhat less than anticipated.

So, as a plan B, I picked 50 random email addresses from the membership database and sent out a short reminder, thinking that I'd probably need to repeat the exercise each day until I'd gone through the whole list. How wrong I was. Next morning, I had 24 eager volunteers, all wanting the same 3 books. Oh dear. Turns out those half drafted polite replies came in handy after all.

So, for the time being, we probably won't be making a habit of advertising books by email - at least not until I get a bigger email inbox. Instead, you'll find a listing of available titles advertised within these pages (p6, to be precise). Alternatively, if you don't fancy what you see this time around, but would like to be considered for future book reviews, just drop us a line and we'll keep you posted with upcoming titles.

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

Product Review: Yahoo Desktop Search

By Alex Bailey

Desktop search is big business these days. You know it's big business because all the big players are getting involved. Google released a beta version of a desktop search engine last year, along with Microsoft, and a few hopeful contenders. And late last year saw the release of the Yahoo Desktop Search Beta as Yahoo steps up to meet its main rivals.

But the one thing that is odd about this kind of business, is that all the big players are giving this technology away for free. Whether you call it a beta, a toolbar, or whatever, if it's free and it performs a useful task then that's got to be good for the consumer. And what's particularly good about the Yahoo Desktop Search (YDS) software is that Yahoo have kindly bought a fully working application from a very competent software solution provider, just so they can give it to you for free.

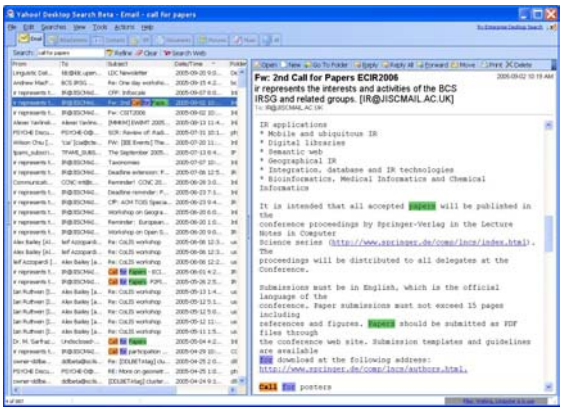
The YDS is a reworking of the X1 Desktop Search from X1 Technologies Inc. Looking at the X1 website it seems that Yahoo have not changed YDS significantly from it's original form. Let's have a look at what we get.

Installation and first-time indexing

I downloaded the software from the Yahoo site. There are two installers that you will need, one for the main application, and another for extra supported file types. I downloaded them both just for good measure but began by installing the main application.

The main application installed swiftly, and after selecting a few simple options, the application started up. One of the options was to synchronise with an existing Yahoo account, which I don't have. But this wasn't a problem and it was straightforward to cancel this option and get on with the indexing. And that's

exactly what the application did. It's a native window's application that runs in a full-screen window, unlike Google's offering, which runs through a browser. And once this application fires up, it starts indexing straight away: Outlook emails and attachments first and then application files. And remarkably, you can start searching immediately. The interface is very responsive, returning search results as you type.



The indexing was also very swift. It finished indexing several thousand emails and files in under an hour. I've had this laptop for a couple of years now and there's plenty to index. The only slight problem was that when you get round to installing the expansion pack for the extra files type it then has to update the main index, which involves some re-indexing. Had it not been in such a rush to start indexing I would have run the second installer immediately after the first, and saved the repetition. In any case the indexing again was fast, and I've yet to notice any inconvenience from any index updates.

The search interface

It should be no surprise to many readers of The Informer that you can have the best indexer and the best retrieval engine in the world, but if the user interface is not up to scratch then the final result will not do the job properly. This is where the standard Microsoft indexing service falls down, and even the Google Desktop Search Beta fails here. Before writing this article I was using the Google Desktop Search, which runs through a browser, and mimics Google's web search interface. However the end result is an inefficient use of space, and too much clicking here and there to get what you want, which at the end of the day is the relevant document or it's content.

In this respect Yahoo wins hands down. The interface itself is very intuitive and startlingly fast. The main window consists of a left panel, which displays the search results with a number of relevant fields, such as 'From', 'To', 'Subject' for emails, and a preview panel on the right, which shows the currently selected document as it would appear in the original application. Along the top there is the query text box, and a number of tabs, which filter the results to one of several file types. The available file types are emails, attachments, contacts, instant messages, documents, pictures, music, and an all inclusive tab for everything. Clicking on a file type refreshes the results panel and the preview panel to reflect the specific file type. There is also a button on the toolbar to launch the current query on the Yahoo web search in a separate browser.

"The YDS has won me over completely"

What is particularly useful about the interface is that it has been specifically designed to help the user actually act on what they have found. The preview pane is again responsive and accurately renders the document, whatever the file type, letting the user immediately recognise a relevant document and access the content. But the interface goes much further than that, allowing the user to reply to emails, print documents, send instant messages, and listen to audio tracks straight from the preview pane itself. Also in the email preview you can open the email in outlook, and more usefully, locate the email in whichever outlook folder you managed to hide it in the first place.

Some flaws, though not serious

One downside, although Yahoo have declared this to be a feature, is that YDS does not index your browser history. This is something that the Google beta does do, but over which Google has been criticised, due to security and privacy concerns on shared machines. It would certainly be nice to be able to search for that website you were looking at last week that you just can't seem to find again, but for me 95%

of my desktop searches are for emails and documents.

Another problem, though no fault of Yahoo, is that with the burgeoning competition and the availability with so many free indexing programs then you might find that you have more than one indexer running at any one time, competing for disk space, and precious CPU time. I uninstalled Google Desktop Search before installing Yahoo just to make sure it doesn't clog up my laptop, but now I'm not sure if I turned off the Microsoft indexer, in fact right now I'm not sure how to turn it off.

Conclusions

The YDS has won me over completely. It has pretty much everything I need from a desktop search application and it works well. I still use MS Hotmail for my personal email, and Google for my web search, so I'm not sure what Yahoo get for all this apart from a happy non-paying customer and a lot of kudos. But then it seems big business is being fought in ever more subtle ways, and happy customers and kudos must be worth a lot of money in the long run.

Alex Bailey leads the Document Analysis team at Canon Technology Europe. His interests focus on the use of document clustering, information extraction, and information retrieval for corporate document management systems. He can be contacted via:
alexb@cte.canon-europe.com

Forthcoming Events

Edited By Andy MacFarlane

14th ACM Conference on Information and Knowledge Management (CIKM)
Bremen, Germany
31 October – 5 November 2005. A general conference on knowledge management with themes on information retrieval, including a workshop on Geographical IR and peer to peer networks for IR.
http://www.tzi.de/CIKM2005/

3rd Latin American Web Congress
Buenos Aires, Argentina. 31 October – 2 November 2005. A general web conference, with a theme on search. This is co-located with SPIRE 2005 for one day (see below).
http://www.ing.unlpam.edu.ar/laweb05/

SPIRE'2005: String Processing and Information Retrieval
Buenos Aires, Argentina. 2-4 November 2005. A well regarded annual conference which focused on String Processing of all kinds including Information Retrieval. It has a strong South American focus.
http://www.la-web.org/spire2005

DocEng 2005: ACM Symposium on Document Engineering 2005
Bristol, England, UK. 2-4 November 2005. A symposium which is devoted to the dissemination of research on models, tools and processes that improve our ability to create, manage and maintain documents.
http://www.hpl.hp.com/conferences/DocEng2005/

Online Information 2005 / Content Management 2005
Olympia Grand Hall, London, UK. 29 November - 1 December 2005. An exhibition and conference concentrating on information and content management, including such issues as enterprise search.
http://www.online-information.co.uk/

2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies
London, England, UK. 30 November – 1 December 2005. A multimedia workshop with a theme on multimedia information retrieval, using ideas such as relevance feedback and ontologies.
http://www.acemedia.org/ewimt2005/index.html

ASIAN'05 Tenth Asian Computing Science Conference Data management on the Web

Kunming, China. 7-9 December 2005. Focus of this particular conference is data management on the web, including issues such as search and information retrieval, Semantic web etc.
http://www.ynu.edu.cn/asian05/

Information Retrieval and Digital Library Applications Mini-Track at the Thirty-ninth Annual Hawai'i International Conference on System Sciences
Kauai, Hawaii. 4-7 January 2006. This mini track focuses on search in Digital Libraries.
http://www.hicss.hawaii.edu/

The Tenth Australasian Document Computing Symposium (ADCS 2005)
University of Sydney. Monday 12 December, 2005. A symposium on document management that has a theme on retrieval issues.
http://goanna.cs.rmit.edu.au/~aht/adcs2005/

ASIAN'05 Tenth Asian Computing Science Conference Data Management on the Web
Kunming, China. 7-9 December 2005. Focus of this particular conference is data management on the web, including issues such as search and information retrieval, Semantic web etc.
http://www.ynu.edu.cn/asian05/

Seventeenth Australasian Database Conference (ADC 2006)
Hobart, Australia. 16-19 January 2006. A general database conference with a theme on information retrieval.
https://www.se.auckland.ac.nz/~adc06/

Book Review:

"Web Dynamics", by M. Levene & A. Poulovassilis

Reviewed by Paul Matthews

Drinking through a straw at the end of the garden hose? This is how one author in Web Dynamics describes the inadequacies of the traditional browsing model for sucking useful information from the web. In the face of problems with the size and searchability of the web, this book represents the latest research aimed at describing its dimensions, improving our navigation through it and enabling systems to help human users by responding to content changes and user needs more flexibly and automatically.

Organised into four main sections, chapters by different authors present literature reviews and original research in the areas of Evolution of Web Structure and Content; Searching and Navigating; Events and Change and Personalised Access.

Inadvertently illustrating the exponential growth of the web, the first chapter looks at estimations of its size using search engine sampling and quotes figures of 7-800 million pages in 1997/8. This seems quite quaint when we see Google alone reporting to have indexed 8 billion pages today!

Following chapters cover the algorithms used by search engines to rank sites (with some interesting recommended adjustments to Google's PageRank) and also to detect communities within the larger network. The actual network structure of the web is also described, which is reportedly something like a bowtie, having a central heavily connected area, one section connecting in, and one section connecting out (As well as many small spots floating alone out in no mans land!). Search engines and browsers should thus also take account of the popularity - connectedness

- of pages as well as being mindful of search semantics and the most efficient paths between sites.

Several chapters deal with the need for an ECA (Event, Condition, Action) model on the web, particularly in the light of the increasing importance of XML and web services in business transactions and in maintaining data repositories. Methods are described for developing a standard equivalent to database triggers, where an application can react conditionally to updates and carry out actions such as maintaining data integrity or sending notifications. What is clear, however, is that an ECA model for XML on the web is a hugely different animal to a traditional database system, given the former's distributed, heterogeneous and often unreliable nature. Incidentally, this section of the book also gets the award for tenuous use of the recursive acronym (REBECA, or Rebeca Event-Based Electronic Commerce Architecture)

"This is not a book for the faint-hearted"

Chapters toward the end of the book deal with the emerging – and promising – field of adaptive hypermedia. These researchers are concerned with developing a framework for applications that can model not only the subject domain itself, but also the skills, needs and environment of the user. Content can then be delivered which is customised for the user, saving frustration in navigation and providing interfaces appropriate to the task. Here, as elsewhere in the book, there is a plea for the development of open standards, which can do much to prevent reinvention and enable work to build on pre-existing frameworks. Several promising applications for adaptive hypermedia in electronic learning and information systems are listed.

Much of the work described in Web Dynamics is at an early or prototype stage, which would perhaps explain why there may be less interest in it for solution providers than for other researchers (though some of the improved search algorithms seem rather more mature). Certainly, some of the screenshots of geeky looking java interfaces do not smack of polished user interfaces ready for beta testing!

This is not a book for the faint hearted – knowledge of mathematical concepts such as graph, set and probability theory is required to make the most of the ideas being presented. For those seeking a lighter overview of these topics this may not be the book. But that said, the extra effort may be worthwhile. Understanding the algorithms underlying web search and retrieval is the first step in making life on the web easier, slowing the water pressure on that hose so we can actually get to drink.

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

Book Reviews

The following titles are currently available for IRSG members to review:

- Web Content Delivery, ISBN: 0-387-24356-9
- Information Retrieval, ISBN: 1-4020-3003-7
- Wiki Web Collaboration, ISBN: 3-540-25995-3
- Intelligent Data Mining, ISBN: 3-540-26256-3

To obtain a copy, all you need do is write a review for publication in Informer. Simply email us at irsg@bcs.org.uk with your contact details, including full postal address. This offer is available to IRSG members only, and is on a "first come first served" basis - so be prepared to be disappointed if you're not quick!

Feature Article:

INEX: Evaluating content-oriented XML retrieval effectiveness

By Mounia Lalmas

Content-oriented XML (eXtensible Mark-up Language) retrieval systems aim to exploit the logical structure of documents, marked-up in XML, to retrieve document components, the so-called XML elements, instead of whole documents in response to a user's query. Implementing this more focused retrieval paradigm means that an XML retrieval system needs not only to find relevant information in the XML documents, but also determine the appropriate level of component granularity to return to the user. Evaluating how good these systems are, hence, requires test-beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

In March 2002, the Initiative for the Evaluation of XML Retrieval (INEX) started to address these issues. The aim of INEX, now in its fourth year, is to establish an infrastructure and to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. This article describes the INEX initiative and its challenges.

Test collection methodology

Evaluating retrieval effectiveness is done by using test collections assembled specifically for evaluating particular retrieval tasks. A test collection usually consists of a document collection, a set of user requests (i.e. topics) and relevance assessments. Most of existing test collections treat documents as atomic units and make assumptions that become invalid in the context of XML retrieval:

- Documents are independent units, i.e. the relevance of a document is independent of the relevance of any other document. In XML retrieval, since we allow for elements to be retrieved, multiple elements from the same document can hardly be viewed as independent units.
- A document is a well-distinguishable separate unit. When allowing for retrieval of arbitrary elements, we must consider overlap of components; e.g. retrieving a complete section consisting of several paragraphs as one component and then a paragraph within the section as a second component. This means that retrieved elements cannot always be regarded as separate units.

The above require a re-visit of the standard test collection methodology so that to appropriately evaluate content-oriented XML retrieval effectiveness.

Document collection

The INEX document collection is made up of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995-2002, and totalling 494 megabytes in size. The collection contains scientific articles of varying length. On average an article contains 1,532 XML elements, where the average depth of an element is 6.9. Overall, the collection contains over eight millions XML elements of varying granularity (from table entries to paragraphs, sub-sections, sections and articles, each representing a potential answer to a query). The collection has just been extended with articles up to year mid-2004.

Topics

To consider the additional functionality introduced by the use of XML query languages, which allows the specification of structural query conditions, INEX defined two types of topics:

- Content-only (CO) queries are standard in information retrieval similar to those used in TREC. Given such a query, the goal of an XML retrieval system is to retrieve the most specific XML element(s) answering the query in a satisfying way. Thus, a system should e.g. not return a complete article where a section or

even a paragraph of the same document may also be sufficient.
- Content and structure (CAS) queries contain conditions referring both to content and structure of the requested answer elements. A query condition may refer to the content of specific elements (e.g. the elements to be returned must contain a section about a particular topic). Furthermore, the query may specify the type of the requested answer elements (e.g. sections should be retrieved).

Relevance

XML elements forming a document can be nested. Some elements will be large (e.g. sections) and others small (e.g. paragraphs). Since retrieved elements can be at any level of granularity, an element (the larger element) and one of its child elements (the smaller element) can both be relevant to a given query, but the child element may be more focussed to that given query than its parent element. In this case, the child element is a better element to retrieve than its parent element, because not only it is relevant to the query, it is also specific to the query.

To capture the above, INEX uses two dimensions to express relevance: exhaustivity and specificity. Exhaustivity is defined as a measure of how exhaustively an XML element discusses the topic of request, while specificity is defined as a measure of how focused the element is on the topic of request (i.e. discusses no other, irrelevant topics). Exhaustivity refers to the standard relevance criterion used in IR, whereas specificity provides a measure with respect to the size of a component as it measures the ratio of relevant to non-relevant content within an element. The combination of the two dimensions is used to identify those relevant XML elements, which are both exhaustive and specific to the topic of request and hence represent the most appropriate unit of information to return to the user.

Retrieval tasks

The retrieval task to be performed in INEX is the ad hoc retrieval of XML documents. In IR literature, ad hoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of documents using a new set of topics. While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library. INEX therefore defined the following two sub-tasks:

- Using CO queries, the retrieval system aims to identify the most appropriate XML elements to return to the user. The assumption is that these elements are components that are most specific and most exhaustive with respect to the topic of request.

- Using CAS queries, the structural constraints of a query can be interpreted as strict or vague conditions.

Relevance assessments

For each topic, the retrieval runs from the participants' submissions have been collected into pools using the pooling method. The assessment pools were assigned then back to participants. To ensure complete assessments, assessors had the use of an on-line assessment system and the task of assessing every relevant element, and their ascendant and descendant elements within the articles of the result pool. This is because if a child element is relevant so must be its parent element, although to a different extent. Therefore, it is not enough to assess the relevance of retrieved elements, but also the relevance of structurally related elements. The participants were given detailed information about the evaluation criteria and about how to perform the assessments. In addition, rules were implemented to ensure consistent assessments (e.g. exhaustivity cannot decrease when going from an element to its parent element).

Metrics

To measure effectiveness, up to 2004 INEX used a metric based on the measure of precall, which computes the probability that an XML element viewed by the user is relevant. To apply the metric, the two relevance dimensions are mapped to one dimension

using various quantisation functions. For example, a strict quantisation function is used to evaluate retrieval methods with respect to their capability of retrieving highly relevant elements. A generalised function is used to credit retrieved elements according to their degree of relevance, thus also allowing to reward less relevant elements.

The latter is important as it allows considering "near-misses" when calculating effectiveness performance. When users access a result element, they have access for example through browsing to structurally related elements (i.e. child elements). Near-misses are elements, which may be themselves not "exactly" relevant to the user's query, but from where users can access relevant content. The idea is that XML retrieval approaches are partially rewarded for finding such elements, as it is still better to return near-misses than irrelevant elements.

The metric, in its current version, does not consider overlap. In INEX, the recall-base consists of a large proportion of overlapping elements (if an element is relevant, so is its parent element). This so-called "overpopulated" recall-base can lead to misleading effectiveness results because the recall-base contains more relevant elements than an ideal retrieval system should in fact return. In fact, highest effectiveness can only be reached by XML retrieval approaches that return all the relevant elements of the recall-base, including all the overlapping elements. Such retrieval behaviour, however, contradicts the definition of an effective XML retrieval system.

Future work

A crucial issue now is to investigate how users interact with XML repositories. This is for two reasons. First, approaches for XML retrieval must be effective in user-based environments. Second, the evaluation methodology is based on certain assumptions of user behaviours that have not yet been empirically validated. The interactive track at INEX, which started in 2004, has started to look at this. A second crucial issue is to adopt a metric that overcomes the problem of the over-populated recall-based. In 2005, a new metric will be used, based on the concept of cumulative gain, which was shown to overcome the above mentioned problem.

INEX 2005 has now four tracks: interactive track; heterogeneous collection track comprising various XML collections from different digital libraries; XML multimedia track investigating access to multimedia content embedded in XML document; and document mining track concerned with clustering and categorising tasks for XML documents. The latter two started this year.

Mounia Lalmas is a Professor of Information Retrieval in the Queen Mary Information Retrieval (QMIR) research group. Prior to this, she was a Research Scientist at the University of Dortmund in 1998, a Lecturer from 1995 to 1997 and a Research Fellow from 1997 to 1998 at the University of Glasgow, where she received her PhD in 1996. Her research focuses on the development and evaluation of intelligent access to interactive heterogeneous and complex information repositories.

Acknowledgements
INEX is led by Norbert Fuhr and Mounia Lalmas, and is partly funded by the DELOS Network of Excellence on Digital Libraries.

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.

Research Update: Collaborative Community-Based Web Search

By Oisin Boydell

Search engines continue to struggle with the challenges presented by Web search: vague queries, impatient users and an enormous and rapidly expanding collection of poorly structured documents all contribute to a hostile search environment. The I-Spy group at University College Dublin is looking at exploiting past search behaviour within communities of like-minded searchers as a means to improve the Web search experience. Our collaborative, community-based personalization approach to Web search, preserves individual user's privacy while providing benefits in search performance for groups of searchers with similar information needs, and for those new to a particular topic.

The I-Spy Web search engine

We implemented our approach to collaborative, community-based web search in the I-Spy search engine (http://ispy.ucd.ie). I-Spy is a meta-search engine, in that it uses the search services of a number of underlying web search engines such as Google and HotBot to provide a basic list of results for a query. I-Spy then uses information from previous search behaviour to enhance the results list by re-ranking and promoting results according to the interests of the community within which the search is being performed.

A community may be defined explicitly, for example somebody may set up a "mountain biking" community and invite people who are interested in mountain biking to join and search within the community. Alternatively, a community may be implicit as in the case of searchers using a search box located on a mountain biking themed website. These searchers form an ad-hoc community that shares a common interest in mountain biking.

The Advantages of Community-Based Web Search

Collaborative, community-based Web search is able to use the experience and knowledge of a group of searchers with similar information needs to decrease query term ambiguity and bring context into the search. For example, searching for the query term 'jaguar' in a conventional web search engine would return result pages about cats and cars. The meaning of the search term is ambiguous: there is no associated context supplied to indicate which meaning of word 'jaguar' the user is interested in. Community-based search uses context learned from a group of searchers to disambiguate the query. A community defined by the users of a search box located on a motoring web site would receive results pertaining to cars above those about cats since for previous searches for queries related to the query term 'jaguar', pages relating to cars were selected.

Our community-based search approach also allows the knowledge of expert searchers to be available to novice searchers. The result selections of an experienced searcher are promoted in future searches for similar queries, so that others, and particularly novice searchers benefit from this expertise. This knowledge is effectively shared within a community.

Many commercial web search companies have been jumping on the personalization band wagon recently (http://news.bbc.co.uk/1/hi/technology/4488927.stm). Most of these offerings provide personalization at the level of the individual. As well as missing out on the collaborative and community benefits mentioned above, these services introduce privacy concerns. Many personalized search engines require that the user log in, and so search behaviour which may include sensitive personal information is recorded for individuals. In community-based web search, the privacy of the individual is maintained. Information about search behaviour is only recorded on a community level and so individual users remain anonymous.

Handling 'Noisy' Result Selections

The benefit of any system that learns from past behaviour in order to improve future performance is only as good as the quality of the captured past behaviour. In community-based Web search, the behaviour of users cannot always be relied upon in terms of their ability to consistently select relevant results. Page titles and snippets of text shown in the results list may be confusing, or a page may be selected not because it was relevant to the search query or community but because it appeared interesting to the searcher for some other unrelated reason.

As part of my research I am looking into methods for identifying these 'noisy' results which are not relevant to the community, with a view to improving the performance of collaborative web search. Analyzing the past search behaviour with respect to result pages that are repeatedly not selected for related queries within a community, and also comparing the page content of selected results are just two possible approaches that I am interested in. Community-based web search can use this information to improve the quality of results that are promoted and re-ranked in future search sessions to improve the overall web search experience.

Oisin Boydell is a Ph.D. candidate in the School of Computer Science and Informatics at UCD, Dublin. He received his B.Sc. from UCD in 2002 and after working in industry he returned to take up a research position in the I-SPY group. Oisin's research focuses on personalization in Web search.

CALL FOR ABSTRACTS

BCS IRSG Industry Day

in association with
28th European Conference on Information Retrieval (ECIR 2006)

13th April 2006, British Computer Society Headquarters, London, UK

For the first time in its history, the IRSG's annual conference (ECIR) will be followed by a special day devoted to the interests and needs of IR practitioners. This forum presents an opportunity for commercial organisations and individuals to share their experiences with a wider audience, and for researchers to learn more about the issues and problems faced by IR practitioners in developing practical solutions for the information search and retrieval industry.

Abstracts for 30-minute presentations should be submitted in plain ASCII by e-mail to mailto:irsgevents@bcs.org. Submissions will be reviewed by IR practitioners on the basis of the originality and authority of the work, and the contribution to the IR profession. For further details please refer to the conference web site.

Abstracts due: 2 December 2005

Industry Day 2006 will be held at BCS HQ in central London (10 mins by Tube from the main ECIR conference venue). A separate one-day registration rate will be available.

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).