Informer

Number 28, Autumn 2008
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.

About a year ago I made the point in this column how ironic it was that a group so concerned with finding, managing and using information should employ such primitive methods for managing the information and opinions of its own members. We speak of grand visions for knowledge sharing and re-use - but how much of this do we do for ourselves? Well, we have a mailing list (used for conference announcements), a mass mailer (used for broadcast messages), and… err, that's about it. No online forums, no members directory, no … community, as such. (Note that I'm referring solely to our online presence – our physical events programme is, by contrast, very healthy indeed – but more of that later.)

It's not as if it can't be done – with fewer than a hundred members and a budget far smaller than ours Conrad Taylor and the KIDMM group have already shown what can be achieved with the initiative and enthusiasm of a few members: a thriving discussion list, regular podcasts, online opinion pieces and tutorials, and the innovative use of social software platforms such as Know.Ware to create an online 'knowledge community'.

Clearly, we have a long way to go. Well, in the spirit of long roads and tiny steps I'm hoping you'll sign up with the recently formed BCS IRSG group on LinkedIn. Evidently, it's not the answer to all our needs, but it's a start – it acts both as a members' directory and, more importantly, provides the discussion forum that we currently lack. New to LinkedIn? Not a problem – signing up takes just half a minute. Think of it as Facebook for grown ups.

Anyway, I mentioned above our events programme, and how much we have to be proud of. So in that respect I'd like to thank Udo Kruschwitz and Alvaro Huertas for their review of Search Solutions and FDIA. If you were one of the lucky few who made it SS

2008, then I'm sure you'll agree with me that it was the best yet – not just in terms of the programme (which for the first time brought everyone together in a highly topical panel session) but the whole community feel – facilitated no doubt by the drinks reception afterwards. I really think we're really onto a winning formula here, and am looking forward to next year's event already (which, incidentally, we will probably move to October – trying to promote an event at the height of the summer holiday season was an experience few of us on the organisation team would wish to repeat).

We should show our appreciation also to Springer, who continue to provide us with an excellent selection of books for review. This time around we have "Social Computing, Behavioural Modeling and Prediction", reviewed by Paul Matthews, and "Visualization for Information Retrieval", reviewed by Andrew Neill. Our thanks go them both for their great contributions. And special thanks go to Bob Bater, who has kick-started our product review series with a detailed analysis of desktop search tool Aduna Autofocus.

Of course, if you'd like to join us as a reviewer, just drop us a line at informer@bcs.org. In the meantime, don't forget to point your browser at: http://tinyurl.com/5t2n2g.

Until next time,
All the best,
Tony

Tony Rose, PhD MBCS CEng
Editor, Informer
Vice chair, BCS IRSG
Email: irsg@bcs.org.uk
LinkedIn: http://www.linkedin.com/pub/0/5b/959

Product Review: Aduna Autofocus 5.0 Desktop Search

By Bob Bater

Try as you may, there's no knocking Google off its pedestal as *the* pre-eminent Web search engine. Its 'three Rs' – *Reach* (some 30B pages), *Reliance* (on full-text indexing) and *Relevance ranking* (through citations) - have yet to be bettered in providing access to an unwieldy collection of 50B plus pages on all topics under the sun. But will it scale as the Web grows? There's no reason to think not.

A less frequently asked question is 'How does it scale *down*?' Because there are many scenarios where topics are narrower and deeper, where page-count amounts to mere millions, and where the ability to retrieve reliably in depth (i.e. with *precision*) within a known scope is essential. Google is not good at that. Added to which, it can cope with barely more than the commonest Web document formats - HTML and PDF – and restricts itself to rendering only the most basic types of metadata.

The scenarios which challenge Google's broad-and-shallow ethos are those fields of endeavour where the resolution of both queries and results needs to be much higher than the average Web search. They include the Humanities and the Natural Sciences, Medicine, organizational activities and most areas of pure and applied research – in fact any field where precision of focus and clarity of context are paramount.

There are a number of search applications on the market which recognize these needs. Enterprise search applications, both stand-alone and bundled within ECM systems, offer a combination of free-text and metadata-based searching across a variety of sources and file types as the industry standard solution. Others enhance this standard offering with techniques like faceted search-and-browse, pioneered

primarily in the Humanities and Heritage communities (Flamenco, AquaBrowser), but there are strong adherents also in the Semantic Web community (Longwell, mSpace, /facet) and in the commercial world of enterprise search, Endeca's 'Guided Navigation' being a prime example.

Aduna AutoFocus, although not exactly a new kid on the block, is the latest addition to this list of enhanced search solutions offering faceted search-and-browse combined with an innovative graphical interface for presenting search results, a comprehensive file source repertoire and a robust server-based back-end management console. Aduna call their approach to faceted search-and-browse 'Guided Exploration'.

AutoFocus in Brief

AutoFocus is a desktop and enterprise search solution built with Semantic Web technologies. It comprises two separate but integrated applications, both Java-based and therefore platform-independent. *AutoFocus* is a serious alternative to other free desktop search applications such as those from Yahoo and Google (not to mention Windows' effort) and Copernic Home version. Out-of-the-box, it offers fast, friendly retrieval of a variety of file types from local Sources (file systems and email stores on internal and attached storage), Web sites, and/or Sources defined via its companion back-end *AutoFocus Server*.

Sources defined through *AutoFocus* itself are workstation-specific and cannot be shared. However, *AutoFocus Server* extends the searchable Sources beyond the desktop workstation to the corporate LAN or WAN. Sources on a LAN or WAN (or the Web) can be specified and organized into Profiles appropriate to different user communities' interests. Profiles can be shared network-wide under controlled conditions, and refreshing of Sources can be initiated either manually by individual users or automatically via refresh and rescan schedules set within *AutoFocus Server*.

Semantic Web Pedigree

The origins of AutoFocus are interesting. It is a partial offshoot of the OntoKnowledge project, a wide-ranging research project funded by the European Union under Framework Programme 5 from 2000-2002 in which Aduna were a participant. As a research project, OntoKnowledge wasn't designed to produce commercial products directly. It did however lay a firm foundation for a number of technologies which underpin products now entering the market, in particular Sesame - an RDF (Resource Description Framework) database – which provides the storage for AutoFocus.

Following the OntoKnowledge project, Aduna developed the semantic infrastructure and a visual front-end for exploring a large database of medical papers (EMBASE) in the DOPE (Drug Ontology Project for Elsevier) project, funded by the Advanced Technology Group of Elsevier Science. This incorporated Aduna's unique cluster map visualization technology. AutoFocus evolved out of the DOPE deliverable by adding the Aperture crawler and additional code for indexing and term-weighting.

Sesame

We need say little here about Sesame, except that its transparency in operation is deceptive. You don't need to interact with it at all, but because it is built on Semantic Web standards, you can extract data from it with a variety of third-party RDF query tools.

Aperture

Like Sesame, Aperture works invisibly and seamlessly behind the scenes on the Sources you specify. It extracts both content and metadata for AutoFocus to apply what appear to be frequency weightings to determine 'significant terms' in the content. Out-of-the-box, AutoFocus' implementation of Aperture indexes four key metadata attributes of source items in addition to text content:

- Text
- Title/Subject
- Summary & Description
- Path & File Name
- Authors & E-mail Names

Any one or any combination of these may be selected for inclusion in what's often called a 'fielded search'. Out-of-the-box the fields available are fixed, but Aduna offer a customization service which can add user-specified fields - Purchase Order Number for instance, or Customer Number where these occur in discrete fields in the document corpus.

In addition, Autofocus groups indexed items into nine 'facets' which may be used for retrieval in their own right, or as filters on existing search results:

- Keyword Suggestions
- Tags
- Source
- Location
- Date & Time
- Type
- People
- Language
- Size

Most of these are self-explanatory, but two require further comment. Keyword Suggestions presents a drop-down list of all 'significant terms' in the result set. This allows filtering of the result set to see which documents include which significant terms. The Tags facet is a feature new to Autofocus 5, and allows users to tag Source items – or result sets - with their own terms and search or filter on them.

Aperture can extract significant terms from the content of a wide variety of file types and also any metadata in Title, Author and Keyword elements. Supported file types include:

- plain text
- HTML, XHTML, XML
- PDF
- RTF
- MS Word, Excel, Powerpoint, Visio, Publisher, Works
- OpenOffice 1.x: Writer, Calc, Impress, Draw
- StarOffice 6.x - 7.x+: Writer, Calc, Impress, Draw
- OpenDocument (OpenOffice 2.x, StarOffice 8.x)
- Corel WordPerfect, Quattro, Presentations
- e-mails (.eml files)

In addition, it will index filenames and paths (but not content) of all other common file types, including MP3 audio and MP4 containers, executables (EXE, BAT, shell

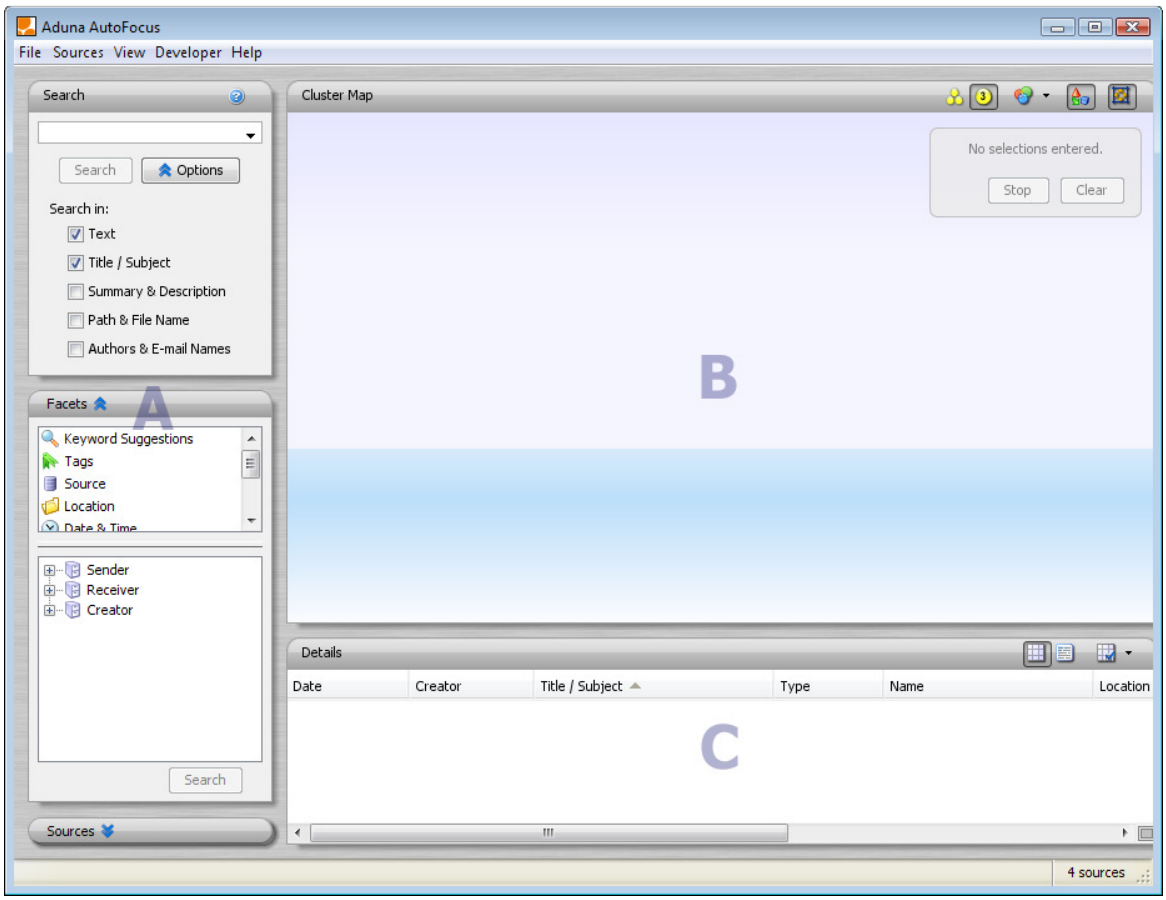

Figure 1

scripts), most image file formats (including ICO but not PNG), WMV, AVI, FLV, MOV, RM, CHM, SWF, and even REG.

That's an impressive range of file types, but there are some notable quirks and omissions. For instance, why include the minority WordPerfect WP document format when the only email format supported is .EML, which is not used directly as a storage format by any of the top email clients? Metadata embedded in multimedia files is not supported either. Aduna would greatly enhance the appeal of AutoFocus were they to bring Aperture up to date to include Outlook PST and MBOX email sources, archive files (ZIP etc.) and multimedia metadata such as that embedded in TIFF images, MP3 audio and MP4 containers. Aduna tell me that these capabilities are under development.

AutoFocus Interface

The AutoFocus interface illustrated here is that of the client application, which is an executable. This interface is however also replicated in AutoFocus Server, where the faceted navigation engine Spectacle provides an equivalent web-based interface. The interface is divided into three resizeable main panes (Figure 1). On the left is the Search and Navigation Pane (A) comprising three panels, Search, Facets and Sources. The Search panel is displayed at all times but the other two panels may only be displayed alternately.

On the right, the Cluster Map pane at the top (B) displays search and filter results as cluster maps – a variation on the Venn diagram - using Aduna's unique visualization technology. At the bottom, is the Details pane (C) which displays retrieved items either as a list or as a table when a result cluster is selected. The columns displayed in the Details pane are user selectable.

AutoFocus in Action

The Search pane contains a Search box and a list of search fields from which those to be searched can be selected. AutoFocus supports

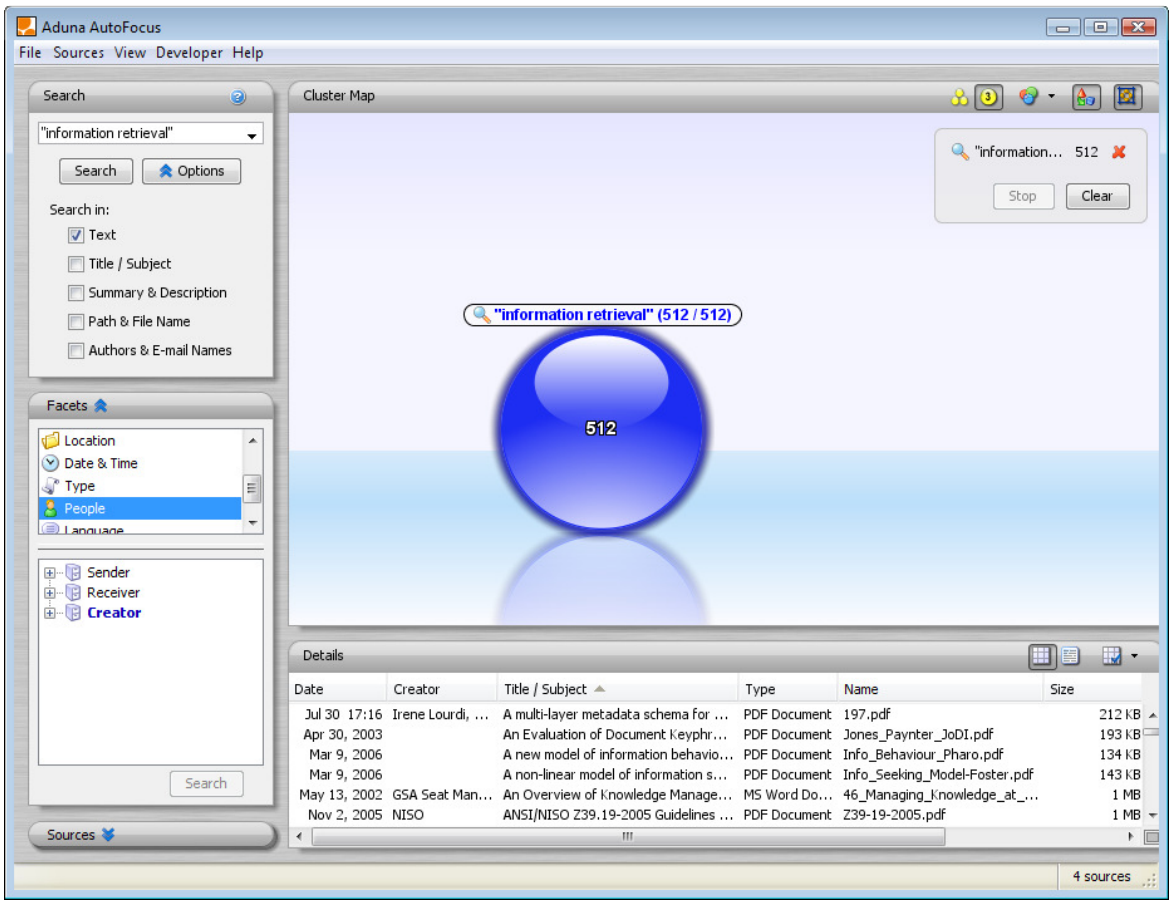

Figure 2

the search syntax of the Lucene open source search engine (http://lucene.apache.org/java/docs/queryparsersyntax.html) including the standard Boolean operators and the use of parentheses for subqueries. Because AutoFocus supports combined search and browse, iterative searches are possible, allowing an initial result set to be refined in a number of ways. This provides a high degree of flexibility, allowing AutoFocus to support quite sophisticated discovery strategies.

For instance, a search on the compound term "information retrieval" can be performed in the Text search field alone using the compound term in double quotes. In this case, a single cluster map is displayed in pane B. Figure 2 shows that 512 items have been retrieved. Any other search field or any combination could have been selected before performing the search.

Alternatively, a query can be performed using the term information alone (3479 items), followed by a second search using the term retrieval alone (1152 items). In this case, three clusters are displayed, one for information, one for retrieval, and a third cluster in-between them showing items containing both terms. Clicking any of the clusters displays the results for that cluster in the Details pane.

Query building need not stop there. It could be part of our discovery strategy for instance, to see how many items consider the term information retrieval important enough to include it in the Title/Subject field. Figure 3 shows the cluster map resulting from querying first on the Text field and then on the Title/Subject field. The answer to our question is that there are 20 items with the term in the Title/Subject field, 19 of which also have the term in the Text field, and only one with it solely in the Title/Subject field. Of the original 512 with the term in the Text field, 493 don't also have it in the Title/Subject field.

Result sets can be refined in various ways by selecting any of the facets in the Facets panel

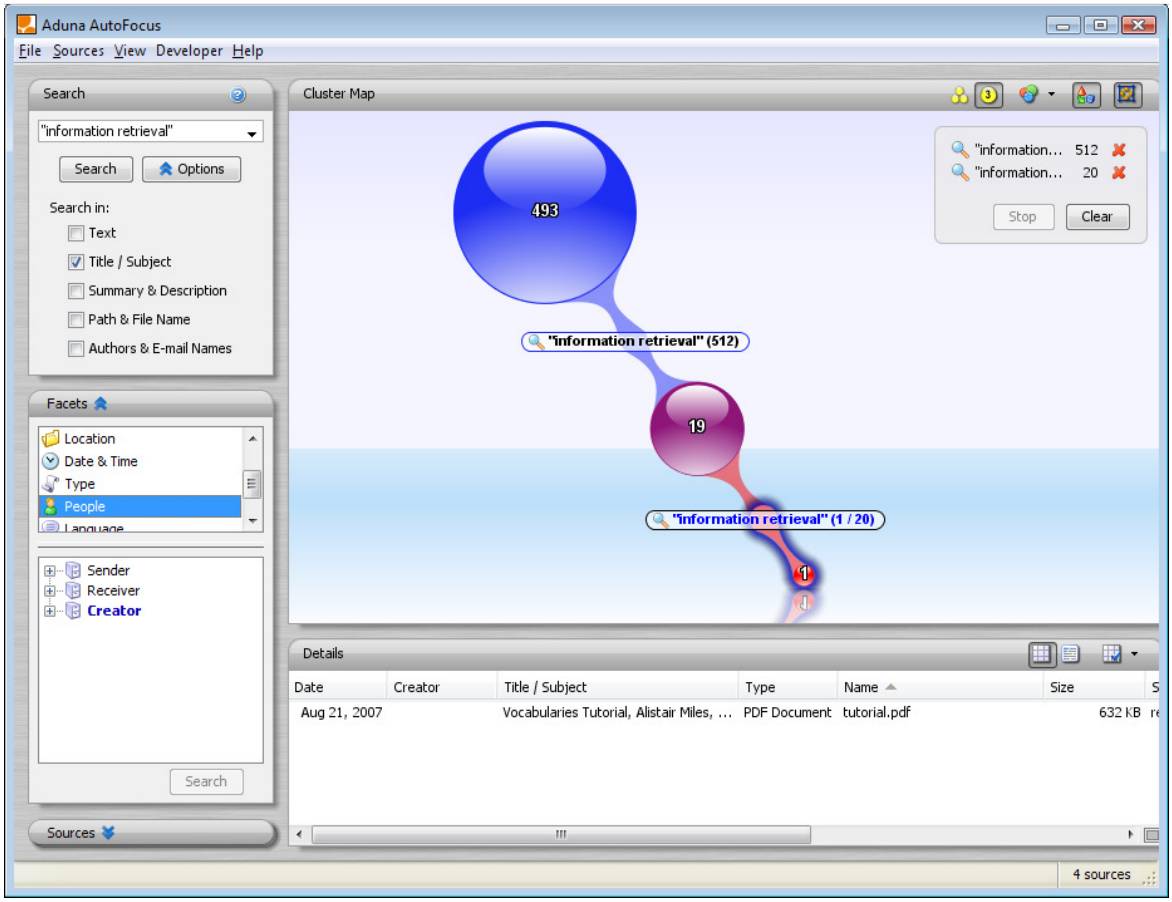

Figure 3

and then selecting one of the values presented. This overlays each cluster with legend showing the proportion of documents containing the filter value. For example, we might decide to see whether any of the result sets include documents by the **Creator** (author) Tony Rose, because we know that he is editor of the authoritative periodical *Informer*, published by the BCS. We know that some **Creator** values are available because in Figure 3, the **Creator** facet is emboldened.

Figure 4 shows what we get when we select **People** in the Facets panel, display all the values under **Creator**, scroll down and select 'Tony Rose'. Our result set contains 14 documents with the value 'Tony Rose' in the Creator/author field. All of these are in the cluster where the term *information retrieval* occurs in the Text field ('14/493') and none in the cluster representing the intersection of both queries ('19'). The filtered list of items is displayed in the Details pane, and we can open any of them by double-clicking.

The example above uses the facilities of the Facets pane more-or-less as filters on a result set obtained through keyword searching. In fact, you need not start with a keyword search at all, since all of the Facets may be searched independently, with the single exception of Keyword Suggestions. For instance, you might search on a People > Creator value first, then filter the results by date = 'Past Year' and finally by 'Keyword Suggestions'. I know of no other comparable product which provides such flexibility.

AutoFocus offers various other less central yet useful points of functionality of which we will mention only one here. The ability to save sophisticated queries is a useful facility, as is the ability to export the result sets generated as a list of items and their locations. This is only partly possible in AutoFocus. Clusters can be saved by tagging them and can then be retrieved at a later date by searching on that tag. However, this does not save the query parameters which generated the cluster, so the query is not re-run and the result sets are

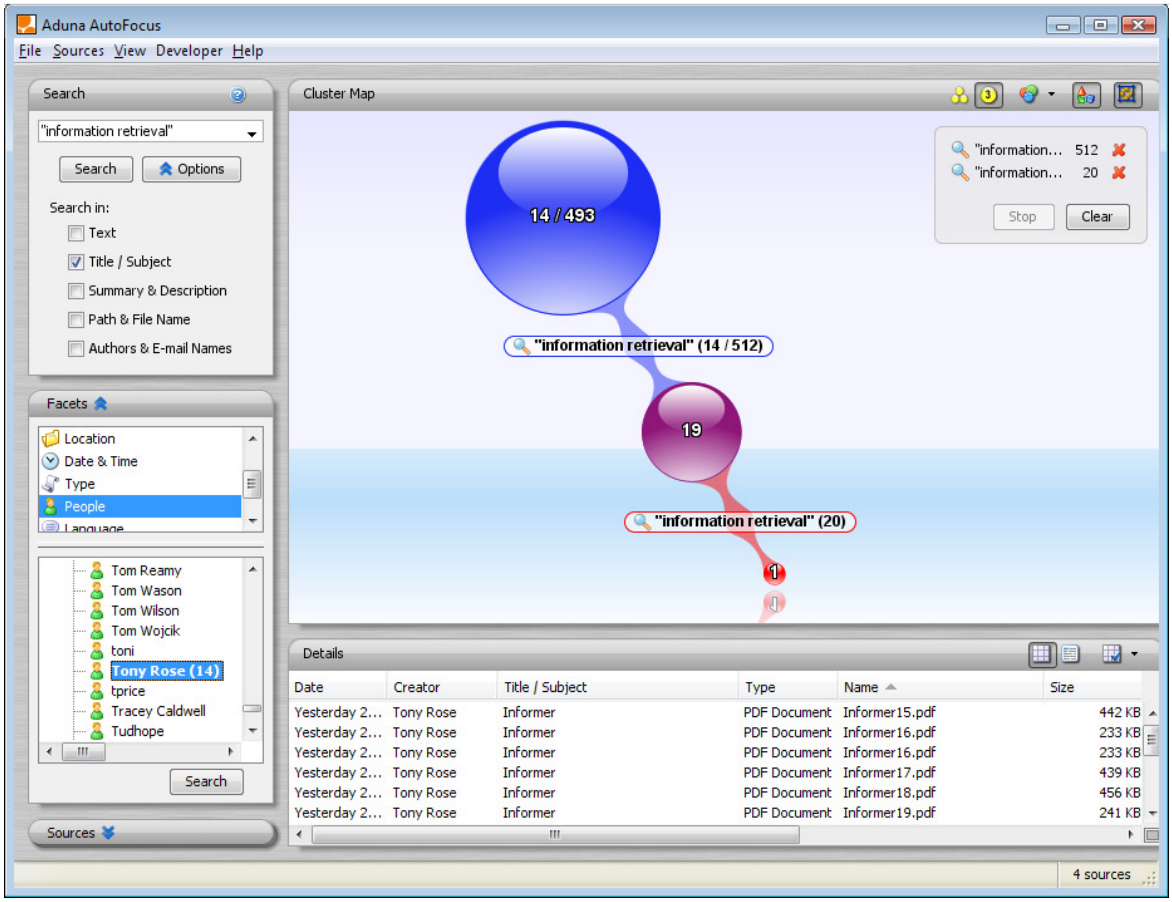

Figure 4

therefore static. Again, I am informed that the ability to save queries is a feature under development.

The 'Export Cluster Map' menu item does not export the parameters either. Nor does it export a list of items comprising a cluster, but does something rather clever instead. This menu item offers three options: PNG Image, HTML Image Map and XML Image Map. The first exports a simple PNG image, which is not a great deal of use on its own. The second option however, can export the cluster map image embedded in an HTML page together with all the item names and paths/URLs mapped to discrete areas of the cluster image in an HTML image map. Hovering over one of these areas displays the title and URL, while clicking generates a browser pop-up asking if you want to save or open the item.

"Autofocus desktop client combined with Autofocus Server provide a partnership which many of the bigger players will struggle to beat"

This is a nice piece of functionality which would allow result sets to be exported, say, to an intranet; if it worked properly that is, which it doesn't because the map is as dead as a Dodo. In AutoFocus 5, this feature generates an HTML 4.01 Transitional page which fails the W3C validation test dismally. It was only after the broken HTML was mended in several places that it worked. And even then, the on-screen locations of the image map areas were displaced some 2cm lower than where they should have been.

Conclusion

AutoFocus looks good, performs well and is as competent an implementation of the faceted search-and-browse paradigm as I've come across. It supports sophisticated iterative search strategies affording a retrieval precision I've seen achieved only by enterprise search products with a four-figure price tag or more. No doubt, the 'big boys' in enterprise search will play catch-up and will match that degree of precision – at a price – but few of them, if any, will find it easy to offer the two key features which set AutoFocus aside from the generality of desktop and enterprise search solutions.

Firstly, AutoFocus' cluster maps presentational idiom is no gimmick; it actually does enhance the search experience, not just because the graphics are attractive, but because it renders analytical depth to search results, providing a far higher resolution than the average Web search can provide. In real-world retrieval situations, that's an edge which can make a big difference. Secondly, AutoFocus is based on Java and Semantic Web technologies, which means it is Semantic Web-ready. At the enterprise level, Autofocus desktop client combined with Autofocus Server at the back-end provide a partnership which many of the bigger players will struggle to beat – not only in terms of value-for-money, but also in terms of being future-proof.

Sure, AutoFocus has some rough edges. It needs to extend the range of file types it can index, it needs some broken (though not core) functionality to be mended, and the documentation is currently poor. The fact that the desktop client is free does not excuse these flaws. Nevertheless, I've recently switched to it as my main desktop search application after using a competing product for over five years. And I've no regrets.

The Autofocus desktop client 5 is downloadable free from http://www.aduna-software.com/home/download/overview.view. AutoFocus Server version 4 is downloadable from the same URL, again free under an open source licence. Note however that AutoFocus desktop client 5 is incompatible with Server version 4. AutoFocus Server 5 is due to be released fourth quarter 2008, but will be available only under a commercial licence.

Bob is Principal Associate with InfoPlex Associates, formed in 1994 to provide advisory and development services in knowledge and information organization. He has written and presented regularly on these topics over the years. Bob is currently Vice Chair of ISKO UK and a member of the BCS, CILIP and the RMS-GB. He can be contacted via: bbater@infoplex-uk.com.

Book Review

"Social Computing, Behavioural Modeling and Prediction", edited by H. Liu, J. Salerno & M. Young

Reviewed by Paul Matthews

This publication is a collection of 27 papers presented at the First International Workshop on Social Computing, Behavioural Modeling and Prediction in Arizona, USA in April 2008. Although the conference, and these proceedings, are billed as being strongly multidisciplinary, the majority of the papers in the volume describe applications of computer modelling techniques to social network data and group behaviour patterns. With the conference sponsorship coming from the Air Force Research Laboratory, it was perhaps not surprising that many of the papers address the problem of detecting terrorist cells and networks and predicting violent or insurgent activity. Despite my slight disappointment at the emphasis on this theme, I was pleased to see that there was also recognition of the more benign and constructive reasons that people form groups and cooperate, and the need to better understand these.

Rather than mention all the papers, I will pick out several that I found to be particularly interesting. In the opening chapter "Rational Choice Theory: A Forum for Exchange of Ideas between the Hard and Social Sciences in Predictive Behavioural Modeling" Sun-Ki Chai presents a thoughtful review of rational choice theory, covering its origins in mathematics and computer science and its uptake in the economics and the social sciences. Whilst mindful of the theory's limitations across a range of contexts – such as situations where there are less tangible benefits to cooperation, Chai stresses its continuing predictive and explanatory power, and describes his attempts to accommodate cultural variety within the theory. He finishes by challenging the computer modelling community to apply and extend rational choice theory rather than operating in theoretical isolation. One paper which later on takes up this mantle is "Metagame Strategies of Nation-States, with Applications to Cross-Strait Relations". Here, the authors use the idea of strategic, conditional game strategies (the metagame) to explain political outcomes, and claim success in predicting the status quo between China and the US over Taiwan. I say "claim", as I for one would have needed more time to understand if their methods justified this.

"Such techniques have enormous potential for providing an unprecedented level of understanding of social networks"

Several of the papers undertake social network analysis on real or test datasets. As an entrée I found the chapter "Social Network Analysis: Tasks and Tools" by Steven Loscalzo and Lei Yu to be a very useful introduction to the area in introducing methods and available software for conducting social network analysis. Interested researchers could do worse than refer to this paper for a concise summary of available tools and methods. In "Mobile Phone Data for Inferring Social Network Structure", Nathan Eagle ad colleagues from MIT and Harvard study 330,000 hours of mobile phone data from 94 subjects, including proximity and location data. In an interesting analysis, they show how reciprocal friendship can be inferred from a number of simple factors such as amount of communication, proximity at home and proximity on Saturday nights. They conclude that such techniques and huge datasets have enormous potential for providing an unprecedented level of understanding of social networks. Using a similar dataset, Karsten Steinhaeuser and Nitesh Chawla show in "Community Detection in a Large Real-World Social Network" that weighting using node attributes (i.e. things people have in common) leads to much more accurate community identification than that based on topological features alone. This is not really a very surprising finding, however, I felt – even the authors themselves refer to the 16th Century aphorism "birds of a feather flock together"

Spatial data are also considered in the types of modelling and prediction addressed in the proceedings. In "Where are the slums? New approaches to urban regeneration" Beniamino Murgante and colleagues from the University of Basilicata in Italy describe how spatial statistics can be used to more accurately identify areas requiring regeneration interventions. They use spatial autocorrelation (Moran's Index) on census data to show that education levels have the highest spatial autocorrelation of a range of indicators in their study area (Bari in Italy) and that this information could be used to guide the planning of new schools.

Another type of software represented, albeit by a single paper, is group-based collaboration. Gregorio Convertino and colleagues from the Pennsylvania State University study the use of groupware in emergency management in "Designing Group Annotations and Process Visualizations for Role-Based Collaboration". The authors describe the development and testing of some interesting shared mapping and annotation prototypes which recognise that experts from a range of areas are called in to work on emergency relief but that it is critical that these experts can collaborate on key base level information. In addition to the addition of data to shared maps, the prototypes allowed the development of user-defined tags for the annotations added.

"Maybe I'll also provide a copy of this work to Fabio Capello"

While studies such as these have a clear social purpose, others in the volume are rather more exploratory, taking agent-based modelling algorithms and applying them to social behaviour and networking problems. "An Ant Colony Optimization Approach to Expert Identification in Social Networks" by Muhammad Aurangzeb and Jaideep Srivastava and "Particle Swarm Social Model for Group Social Learning in Adaptive Environment" by Xiaohui Cui and colleagues are two such examples. Aurangzeb and Srivastava liken specific expert knowledge to the ant's path to a food source - current for a while, then decaying. By applying an ant colony exploration strategy to an expert network, they claim success in being able to access a relevant expert faster than through a random walk. Cui et al use the particle swarm algorithm – where an individual particle's movement is influenced by its neighbours' – to draw parallels to the adaptive nature of social learning, where people hang on to a piece of knowledge until a "fitter" piece is encountered. Results of their analysis I found rather opaque, though the approach seems an interesting one.

"There was remarkably little attempt by some of the authors to explain or generalise the substance of their work"

One of the more bizarre – yet oddly compelling - papers is "Clustering of Trajectory Data obtained from Soccer Game Records – A First Step to Behavioural Modeling" by Shoji Hirano and Shusaku Tsumoto from Shimane University in Tokyo. They analysed 64 games from the 2002 World Cup (I hope the "soccer pro" who had to do this was suitably paid!) to identify contiguous passing combinations that led to goals. They then used multiscale matching and cluster analysis to group the events into recognisable patterns. While the potential practical applications of this work did not exactly leap out of the page at me, the authors claim the possible use of such spatio-temporal data mining in analysing other types of "goal"-driven behaviour. Maybe I'll also provide a copy of this work to Fabio Capello, just in case it is the missing piece of the jigsaw in the England match strategy!

Chai's opening warning is somewhat borne out by the range and disconnection between the theoretical approaches of the different authors in this book. To some extent, this may be expected in an emerging field. But equally, the transition alluded to in the title – from modelling a discrete domain to prediction across domains – is where much of this work falls short in both ambition and execution. Also, for an event that was billed as being multidisciplinary, I felt there was remarkably little attempt by some of the authors to explain or generalise the substance of their work for the benefit of experts in other subjects, which

might lead a sceptic to suspect that there wasn't enough substance to be explained in the first place.

Paul Matthews is a Senior Lecturer at the Bristol Institute of Technology in the University of the West of England. He teaches on various topics in web design and information and library management, specialising in information architecture. His research interests include ICT for Development and Social Technology adoption in voluntary and community organisations.

Conference Review: FDIA & Search Solutions 2008

By Udo Kruschwitz and Alvaro Huertas

Two more great IRSG events! Read on ...

The subject of Information Retrieval, and more encompassing ones like Information Access, are the playground of a thriving research community; a community that heavily relies on conferences and meetings for its development. But, how do you get involved, in the first place? FDIA is the perfect starting point for a PhD student to become familiar with these academic events, and learn to best approach and enjoy them.

FDIA 2008 took place in London, in the British Computer Society (BCS) headquarters, an excellent venue in the heart of the City. This event offers its participants the occasion to meet a good number of other researchers in different stages of their PhD, and an exciting variety of areas of research as well. Participants can also benefit from getting feedback by leading figures in the field such as Stephen Robertson.

"FDIA is the perfect starting point for a PhD student to become familiar with these academic events"

The presented research was grouped in three main categories: *Context and Language in IR, Applications and Distributed Systems, and New Domains of IR.* Students and researchers from all around the UK and Europe presented their projects of research, current results and experiences in an informal and friendly setting, and had an overview of the research of others. After the presentations, where interaction with the audience was omnipresent, discussion around the posters showed even more unexpected relations between different research objects and methods. Blackboards were supplied to complement the discussion around the posters, and were mercilessly filled with diagrams, citations of authors and formulae.

The first session was opened by Emanuelle di Buccio with a formal proposal for the

modelling of the evolution of semantic context. Deirdre Lungley presented her work on using implicit user feedback in an intranet environment to learn document/term associations. Maarten van der Heijden examined the possibility of parameter-free Language Model formulations for IR. The final speaker in the session was Yi Chen who approached the problem of re-finding personal information using our knowledge about human memory. She is part of the team at Dublin City University that explores "life logging" with images. Life logging refers to those people who carry that Sensecam around their necks - day and night! An exotic technological development, and a formidable research challenge.

"A lot of familiar faces around, which made it feel like a family reunion"

In the second session, Ahmad Abusukhon presented his research on the interplay between load balance and query throughput in a grid IR system, Gianmaria Silvello introduced the problem of managing metadata in systems of different archives, and Marina Santini presented the primary assessment of a Genre-Enabled Application, with all the challenges and subtleties of the very definition of Genre.

The third session was the most general of all, encompassing the excitingly multidisciplinary research of Hanna Jochmann-Mannak on IR systems for Children, and a proposal for a context-aware lexical measurement scheme for IR based on the concepts and mathematical machinery of Quantum Theory, presented by Alvaro Huertas-Rosero.

The variety of the research presented proved to be rich in points of contact from a formal point of view, just as much as the very experience of undertaking a research project in each field.

There was plenty of time to socialise afterwards. Discussions continued later in the bar; in fact, when the group left, there was neither more draught beer nor any bottled beer left (this is true!)

FDIA was followed by Search Solutions 2008, the second such event following the successful Search Solutions 2007. This is a special one-day event, similar to the Industry Day, dedicated to the latest innovations in information search and retrieval. The event aims to be interactive and collegial with a high quality technical programme. To achieve maximum interaction attendance was limited to about 50 people. The event proved to be highly interactive! All talks were given by experts from industry, some of them with an academic background. A lot of familiar faces around, which made it feel like a family reunion.

"The day started with talks by the heavy weights: Yahoo!, Microsoft and FAST (it's still spelled FAST but now pronounced "Microsoft)"

The day started with talks by the heavy weights: Yahoo!, Microsoft and FAST (it's still spelled FAST but now pronounced "Microsoft"). Yahoo! is opening up its search platform and invites people to use BOSS ("Build You Own Search Service"). Milad Shokouhi explained how Microsoft tries to address the inherently hard problem of assessing the user's intent and concluded that clickthrough patterns are extremely important.

The second session included talks by Teezir about sentiment analysis (with striking examples to illustrate why *every* company should make use of such a tool), an update on Trexy's work to collect and manage user search trails. In this update we learnt how Trexy is dealing with the 7 deadly sins of searching. Finally, Richard Boulton from Lemur Consulting showed more on real life examples: this time, about faceted search for query enhancement.

The lunch break lasted an hour but it felt much shorter. Why? Interaction! What makes Search Solutions so valuable for the visitor is the opportunity to meet people who face similar problems and who are happy to share ideas and experiences: it emerged, for example, that there is a very good reason why search engine providers move away from a simple interface to something that allows more interaction (e.g.

facets): internal studies show that people now actually make use of such features, in intranets as well as on the Web.

The afternoon sessions started with talks by Autonomy followed by WebOptimser, a search engine marketing company that contributed useful statistics. It turns out that 70% of all e-commerce transactions originate from search, a surprisingly high number. Ayse Göker (AmbieSense) finished the session with a talk on mobile search presenting a commercial product that emerged from an EU project. An interesting aspect of her presentation was the explanation of how mobile search differs notably from common Web search.

"What makes Search Solutions so valuable for the visitor is the opportunity to meet people who face similar problems and who are happy to share ideas and experiences"

Elias Pampalk of Last.fm started the last session with a talk on music recommendation. The company started only six years ago in a small place in East London. To get a feel for how big the company has grown you only need to look at the number of data points Last.fm has in its database (e.g. the pieces users are listening to). It's 20,000,000,000. In addition to that there are 50 million user tags. But we also learned that even if you are lucky to have that much data, you can still face the challenge of not having enough of it because new artists are constantly emerging, trends come and go. etc. The very last talk by Solcara proposed a way of harvesting structured data from the Web using the RDF query language SPARQL.

A panel discussion rounded up the day. Such panels can drag on forever without ever getting anywhere. However, Conrad Taylor must be congratulated for chairing this very concise and fruitful one which revolved around the question of how information can get organised, classified and categorised without an organisation to do it. Naturally, there was a lot of discussion about user-tagged content. A point Elias made about tagging was "If it's a pain to apply a tag, nobody is going to use it!" There was broad agreement that the Semantic Web community can learn a lot from folksonomies, but there was also a feeling that in some domains there is simply a need for formal taxonomies, e.g. in medicine. Generally speaking, it is very appealing to ask users to provide their input, but, in Andy MacFarlane's words, do we collect the "wisdom of crowds" or the "madness of the mob"?

"70% of all e-commerce transactions originate from search"

How could Search Solutions be summarized in a single sentence? Perhaps like this: It offers a look behind the scenes of the major search companies, provides an overview of what tools are being used in the real world, and gives a feel for the emerging trends in search and retrieval. Tony, well done!

Look out for more events on the BCS IRSG Web site.

Alvaro Huertas is a Colombian chemist (Universidad Nacional de Colombia, Bogotá, Colombia) who abandoned the exercise of Chemistry to undertake a Master in Phyisics (Universidad de los Andes, Bogotá, Colombia). In 2006, after some years teaching physics in Colombian universities, he abandoned Physics as well to undertake a Ph. D. in Computing Science at the University of Glasgow in the group of Information Retrieval (IR), where he currently developes his research. This apparently disconnected path from one discipline to another has been guided by the study of Quantum Theory as a constant axis, one that can be tracked up to his present research. He now explores methods and practical schemes for IR based on the concepts and mathematical machinery of Quantum Theory, following the groundbreaking ideas of his supervisor C. J. van Rijsbergen.

Udo Kruschwitz is a Lecturer in the Department of Computing and Electronic Systems at the University of Essex. He received a Diplom in Computer Science from Humboldt University Berlin and a PhD in Computer Science from University of Essex. His main research is in natural language processing, information retrieval and the implementation of such techniques in real applications. He is the author of the monograph "Intelligent Document Retrieval: Exploiting Markup Structure", published in Springer's Information Retrieval series. He is a member of the BCS-IRSG committee.

Book Review

"Visualization for Information Retrieval", by Jin Zhang

Reviewed by Andrew Neill MBCS

This book is another in the Springer information retrieval series, aimed at the academic market. Hardback, with 268 pages of dense text, regular equations and occasional diagrams, it is not immediately accessible in the way a "…*for dummies*" book might be. However, Zhang provides a guide across the field of visualisation that is comprehensive, extensive and knowledgeable to a target audience of researchers and specialists in the field. As a practitioner, I was pleasantly surprised that I could follow – and even, dare I say it, enjoy! – the ideas and models discussed, tribute no doubt to Zhang's command of the subject.

Overview of Book Subject

Visualisation and Information Retrieval are intimately linked because all retrieval must be followed by some form of presentation of what has been found. The simplest, most widespread (via the internet) and least powerful method for visualisation is the linear text list, found and used daily by various famous search engines. This one-dimensional visualisation model is only one method for providing feedback on the results, how these results came to be retrieved, and how they related to the wider data repository, and in fact provides very little information about these potentially important factors (a fact that is acknowledged in some web searches today – for example, "*tag clouds*" as used on the Times Online, Cluuz.com and others show relative popularity of tag terms that match the results, and the property search *GloBrix.com* shows how the results break down by using a 2D graph of results against price). More powerful visualisation techniques, based on and working alongside the underlying structure of the data repository, can enhance the user's ability to understand how the result set was arrived at, and allow interaction and

modification of the query to achieve a better result.

Description of Book Content

This book aims to be a one-stop-shop for the field of information visualisation. In his preface, Zhang outlines the structure of the book, and explains the approach to this work – to provide a good understanding of the status of theory and practice behind mainstream models, provide guides to the work of leading researchers, discuss current limitations and provide practical advice for those who are planning to implement. He also explains his approach for selecting which specific models to cover – they must be mainstream and mature, represent the major types of visualisation that other models are derived from, demonstrate fundamental characteristics of information retrieval and visualisation, and display deep semantic relationships amongst the data being visualised.

"Visualisation and IR are intimately linked because all retrieval must be followed by some form of presentation of what has been found"

The book begins an introduction to the subject – explicitly separating the fundamentally different paradigms of searching versus browsing, followed by a discussion of the background theory. These sections provide the reader with enough of the vocabulary, theory and research environment to profit from the rest of the book. I also found Chapter 3, which discusses visualisation models with multiple reference points, essential to my understanding of the later chapters.

Chapter 4 reviews Euclidian-based spatial visualisation models – projections against X and Y axes using distance/distance or distance/angle models. Chapter 5 discusses neural network-based models and Kohonen Self-Organising Maps, and includes a clear and comprehensible description of artificial neural networks.

Chapter 6 describes the Pathfinder Associative Network model, which is a visualisation technique that illustrates underlying semantic relationships whilst discarding insignificant links, and presenting a line-based network that maps the concepts.

Chapter 7 discusses multi-dimensional scaling, which represents elements geographically according to their empirically judged relatedness. This technique is useful where relationships are not known, and can reveal hidden patterns in data. Chapter 8 discusses internet information visualisation – how to browse and navigate, as well as how to visualise search queries and traffic.

Chapter 9 describes in some detail the problems faced by ambiguity in information visualisation, covering each of the models discussed previously and illustrating some of the weaknesses of each.

"Zhang provides a guide across the field of visualisation that is comprehensive, extensive and knowledgeable"

Chapter 10 is a fascinating overview of "the implications of metaphors in information visualisation" that describes the mental models of metaphor, the use of metaphor and mental models in human-computer interaction, and then provides discussion of applying metaphor to information visualisation. Use of a suitable metaphor can vastly improve the usability and effectiveness of a visualisation model, because the human mind appears to be innately tuned to metaphor (see Steven Pinker "The Stuff of Thought" for a discussion of how deeply metaphor penetrates thought).

Chapter 11 describes how to benchmark and evaluate different visualisation models, and chapter 12, "afterthoughts", is a summary of the book's essential messages – a comparison of the models, and issues and challenges still faced.

Positives

Despite the complexity and difficulty in the subject, Zhang has largely been successful – his book is expansive in scope, yet flows well, and the structure and pace are good. I found that it was pitched at the right level – difficult

and challenging, but not impossible and ultimately rewarding. Several elements are critical to this success – the illustrative diagrams are useful (as one would imagine with a book about visualisation!) but I found that the accompanying algorithms were essential to confirm understanding. Major research is covered, explained, and provided in context.

Criticisms

Despite all the positives – the fact that it is basically very good – there are a couple of problems that significantly damage the book, which can probably be traced to a common problem: insufficient care when editing. Firstly, the text is riddled with grammatical errors – I estimate two per page on average. Zhang is originally from China and not a native English speaker, so this can be forgiven. However, on occasion these distort the meaning of the text, and I was worried that other, more subtle but fundamental, errors were slipping through. As a result, I am reluctant to trust the content.

"The visualisation research field must create models that will be adopted by the majority"

For example, in chapter 4 on Euclidian space, Zhang states that one of the fundamental tenets is that the "*distance from a point X to another point Y is always equal to the distance from the point Y to* **another** *(sic) point X*". Surely, this is not *another* point X, but instead back to the original?

I would have appreciated a glossary, and a refresher on logic notation (for example, the symbols for "*is a member of*" and "*union*"). Similarly, emphasising important terms and formulae – in bold, or better yet in a call-out box – would have made cross-referencing terms and concepts easier. This would have enhanced the generally good use of numbering for formulae and diagrams.

For a book about visualisation, I would really have appreciated some more visual examples of the models being described. I appreciate that this book concentrates on the theory and research, which rarely comes with beautiful illustrations or interfaces, and I also realise that full-colour screenshots of 3D models (those from "Webstar" aside) would have made the book more costly to print. I just felt that more examples would have added to my enjoyment and hastened my understanding. For example, Tianamo Web Search (search.tianamo.com) has a 3D cluster map of related concepts to a search built on top of Yahoo! that really illustrate how visualisation techniques can enhance the user's experience.

Summary and Conclusion

Despite the flaws in the text, I found Zhang's book well structured, comprehensive, clear and interesting. The first chapters provided the grounding required to properly appreciate the later models discussed, and the balance between text, equations and diagrams was good. However, emphasising important text and adding more illustrations and examples would have enhanced comprehension further. Consider, by contrast, "Information Dashboard Design" by Stephen Few, which illustrates the psychology of vision and the corresponding design implications for information presentation. Although Few's book is more practitioner-oriented, it covers academic research, and benefits from heavy use of visual tools such as examples, diagrams and illustrations.

"The academic projects discussed here are just not intuitive enough"

Another concern I have relates to the entire field discussed by Zhang. Early on in his book, I asked myself "would this make things easier for users?", and I am not convinced that it would.

I am all in favour of challenging the hegemony of the '10 blue links' model epitomised by Google, but the visualisation research field must create models that will be adopted by the majority (see "Don't make me think" by Steve Krug for usability examples). I suspect that the current models are too complex for general users. One gap in Zhang's book is an assessment of usability or user adoption of the models.

One reason that alternatives like these have not caught on – despite their obvious power to illustrate, illuminate and extend the ability of the user to manipulate information – is that users need skill and time to learn each of these models to be able to benefit from them. *10 blue links* extends both the "web browsing" and "book index" mental models, whereas the academic projects discussed here are just not intuitive enough. Like the "advanced search" option, they enhance the power of 1% of users who are information retrieval professionals. Most users want effective simplicity and I'm yet to be convinced that current research is delivering on that need.

Despite my irritation with the errors, I enjoyed this book and would recommend it. It is a useful summary of the field, and is broad and comprehensive. Hopefully, a second edition will iron out the problems and add some more screenshots.

Andrew Neill MEng MSc MBCS is the Business Analysis Team Leader in the Information Systems department at Herbert Smith LLP, an international law firm based in the City of London. He specialises in web technologies, information retrieval and knowledge management, and has experience of implementing FAST Search and Autonomy. Previously, Andrew worked at law firm Norton Rose, and before that as a senior consultant at Deloitte & Touche. He is a graduate of both Imperial College and Strathclyde University, and lives with his wife and baby son in North London.

Forthcoming Events

Edited By Andy MacFarlane

Tenth 42nd Annual Hawai'i International Conference on System Sciences: Minitrack on Classification of Digital Documents (HICSS-42)
Of interest to members working in the area of digital libraries. Waikoloa, Big Island, Hawai, 5th-9th January 2009. http://www.hicss.hawaii.edu

2009 International Conference on Communications and Mobile Computing (CMC 2009)
Conference of interest to members working in the area of Mobile Search. Kunming, Yunnan, China, 6th – 8th January 2009.
http://world-research-institutes.org/conferences/CMC/2009/

First International Conference on Human Computer Interaction (HCI 2009)
An HCI conference of interest to members working on the user side of IR. Allahabad, India, 20th to 23rd January 2009. http://hci.iiita.ac.in/hci2009/

Document Recognition and Retrieval XVI, Part of the IS&T/SPIE International Symposium on Electronic Imaging.
A conference which looks at OCR issues and beyond, also in terms of retrieval. California, USA, 21st -22nd January 2009.
http://fens.sabanciuniv.edu/drr/

DigitalWorld 2009.
A collection of conferences in Health, HCI, machine learning, information society of interest to members who work in those specialist areas of search. Cancun, Mexico, 1st – 6th February 2009.
http://www.iaria.org/conferences2009/DigitalWorld09.html

The 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)
A general IR conference with many themes. Enschede, The Netherlands, 2nd – 3rd February 2009.
http://dir2009.cs.utwente.nl/

Second ACM International Conference on Web Search and Data Mining (WSDM) 2009
WSDM (pronounced "wisdom") is a young ACM conference intended to be the publication venue for research in the areas of search and data mining. Barcelona, Spain, 9th – 12th February 2009.
http://www.wsdm2009.org/

AES 35th International Conference - Audio for Games
A Confererence of interest to members working on audio retrieval (either speech or music) for the Games industry. London, U.K., 11th – 13th February 2009. http://www.aes.org/events/35/

IADIS INTERNATIONAL CONFERENCE MOBILE LEARNING 2009
Conference of interest to members working in the area of Mobile Search. Barcelona, Spain, 26th – 28th February 2009. http://www.mlearning-conf.org/

24th ACM Symposium on Applied Computing (SAC 2009).
A general conference on computer science with a special track on IR. Waikiki Beach, Honolulu, Hawaii, USA, 8th – 12th March, 2009. http://www.disco.unimib.it/go/1780137097

9th Conference of the ISKO Spanish Chapter
A knowledge representation conference of interest to members working in that area of search. Valencia, Spain, 11th-13th March 2009. http://www.iskoix.org/

Data Compression Conference (DCC 2009)
Of interest to members working in the area of compression and IR. Snowbird, Utah, U.S.A., 16th – 18th March 2009. http://www.cs.brandeis.edu/~dcc/

AAAI 2009 Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0
Of interest to members working in the area of social search. As part of an overall AI Symposia. Stanford, California, USA, 23rd -25th March 2008. http://www.aaai.org/Symposia/Spring/sss09symposia.php#ss08

3rd International Conference on Adaptive Business Information Systems
Of interest to members who work in the area of IR and business. Leipzig, Germany, 23rd -25th March 2009. http://siwn.org.uk/2009leipzig/ABIS09.htm

Third International Quantum Interaction Symposium QI-2009
A general conference on quantum computing of interest to members working in quantum IR. Barcelona, Spain, 25th -27th March 2009. http://www-ags.dfki.uni-sb.de/~klusch/qi2009/

5th Conference on Professional Knowledge Management: Experiences and Visions (KM 2009)
A knowledge management conference for practitioners. Solothurn, Switzerland, 25-27th of March 2009. http://www.km-conference2009.org/

31st European Conference on Information Retrieval (ECIR 2009)
The IRSG's annual conference focused on all aspects of IR. Toulouse, France, 6th to 9th April 2009. http://ecir09.irit.fr

32nd Annual International ACM SIGIR Conference
The big IR conference, with all themes on the subject of interest to members. Boston, MA, U.S.A, 19th – 23rd July 2009. http://sigir2009.org/

Summer Schools

European Summer School in Information Retrieval
A Bi-Annual summer School on IR. Will also contain satellite events including FDIA 2009 and a panel on Information Retrieval Evaluation. University of Padua, Italy, August 31 - September 4, 2009. http://essir2009.dei.unipd.it/

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.