Information Retrieval
Specialist Group

Search Solutions 2011

16th November 2011
BCS London Office, UK

Udo Kruschwitz, Andrew MacFarlane,
Tony Russell-Rose, Nilesh Thatte, Jun Wang and
Murat Yakici (Eds),
BCS Information Retrieval Specialist Group

Programme

15:15 - 15:45 Coffee Break

Session 4: Beyond Search: Content Analysis

15:45	Jarred McGinnis, **Press Association**, *The Press Association Approach to Search and News*	13
16:10	Ian Kegel, **British Telecom**, *Broadband TV and Recommendation: Improving the customer experience*	14
16:35	Kristian Norling, **Findwise**, *Information flow on the intranet at Region Västra Götaland*	15
17:00 to 17.30	Fishbowl Session *Hot Topics in Search* Want to discuss your own hot topic? Join in, it's a fishbowl session!	16

18:00 to 19:00 **IRSG AGM**

17:30 to 20.00 Drinks Reception

Introduction

Udo Kruschwitz, Andrew MacFarlane, Tony Russell-Rose, Nilesh Thatte, Jun Wang and Murat Yakici
BCS IRSG

Search Solutions is a special one-day event dedicated to the latest innovations in web & enterprise search. In contrast to other major industry events, Search Solutions aims to be highly interactive, with attendance strictly limited.

The programme includes presentations, panels and keynote talks by influential industry leaders on novel and emerging applications in search and information retrieval.

John Tait, Principal,
johntait.net Ltd,
"Using Physical Quantities in Finding Similar Documents"

A common problem in a wide variety of technical and scientific searching is to find documents which express a similar idea to a query but do so using different words or terms. A particular form of this deals with physical quantities, where for example an American-originated query might look for plastics with a melting point over 400 degrees Fahrenheit, but product literature documents originated in Europe may only specify melting points in Celsius. There are two issues here: one is the issue of the units in which the physical quantity is expressed, the other is recognizing that here 200℃ is similar to 400℉. This may be compounded b y ranges - 180℃ to 250℃ is similar to 345℉ to 450℉. Current technology searc h systems pretty much ignore this problem. Google (on the day of writing) in response to the query "elements which melt above 400 degrees F" only returned documents referring to Fahrenheit (interestingly it did return "Fahrenheit"). Wolfram Alpha, which one might expect to do better returned nothing relevant. Recently max-recall Information Systems of Vienna, Austria have developed a suite of plug-ins to Apache Lucene/Solr, a significant step forward for this important form of search.

Lewis Crawford,

Web Archiving Programme Technical Lead,

British Library,

"Analytics and Access to the UK Web Archive"

The World Wide Web is an information system which has witnessed unprecedented growth in the last 20 years since its birth in 1991. It plays an undisputed important role in modern society, fundamentally changing the way we live and communicate. Its impact has been felt in how we publish, learn, teach and research, and many other areas of human activities. The current and transient nature of the Web means that new information replaces older information constantly without any records of the previous state (or versions) of the same information. While new information is being added, existing information also disappears from the web, leave a significant gap in our knowledge of the historical web and potentially in social history. It is therefore not surprising that memory institutions around the world quickly realised the need and value of collecting the content on the Web and started the epic journey of archiving and preserving it since the mid-1990s. The Internet Archive's Wayback Machine is the earliest and most comprehensive web archive to date, containing over 150 billion web pages archived from 1996. The British Library began Web Archiving activities in earnest in 2004 and has collected through explicit permission based archiving 38,000 instances of web sites and more than 240 million web documents. A traditional view is that researchers access previous states of individual web pages and sites in a web archive. There has been a shift of focus in web archiving, from human access to machine access and from the level of single webpages or websites to the entire web archive collection. There is a realisation that rather than looking for 'needles' there may be significant value in the "haystacks" themselves. Using visualisation and data analytic techniques, there are opportunities to provide access to different views of a web archive, unlocking embedded patterns and trends, relationships and contexts. This talk will consider existing access solutions for search to such a collection both visual and full text search based approaches. 'Big data' analytics of what can be derived from treating the corpus as a data set looking at aggregates rather than pages and an insight into the next generation of search tools considering the impact of exponential increases in archived web content."

Toby Mostyn, CTO,

Polecat,

“Search for Public Conversations”

As on-line data grows, and as that data takes increasingly disparate forms (news, social media, blogs etc), getting an overview of this data becomes very difficult. Users (particularly companies) often need to get an understanding of the whole on-line conversation around a particular issue in order to make strategic decision. The traditional search paradigm fails here; the user needs to understand the main themes and signals in the data, not read the constituent documents. What is needed instead is an alternative approach, one where the user is shown the "landscape" of the data, and allowed to explore this landscape to discover detailed information. This talk will focus on Polecat's attempt to provide a solution to this issue, and highlight some of the problems they have encountered in doing so.

Ricardo Baeza-Yates, VP Research,
Yahoo!,
“Beyond the Ten Blue Links”

The classic Web search experience, consisting of returning "ten blue links" in response to a short user query, is powered today by a mature technology where progress has become incremental and expensive. Furthermore, the "ten blue links" represent only a fractional part of the total Web search experience: today, what users expect and receive in response to a "web query" is a plethora of multimedia information extracted and synthesized from numerous sources on and off the Web. In consequence, we argue that the major technical challenges in Web search are now driven by the quest to satisfy the implicit and explicit needs of users, continuing a long evolutionary trend in commercial Web search engines going back more than fifteen years, moving from relevant document selection towards satisfactory task completion. We identify seven of these challenges and discuss them in some detail. This is joint work with Andrei Broder and Yoelle Maarek, from Yahoo! Research in California and Israel, respectively.

Gabriella Kazai, Research Consultant,
Microsoft Research Cambridge,
"Crowdsourcing Search Relevance"

Crowdsourcing of relevance judgments is increasingly used to address the scale and the cost of search engine evaluations which, traditionally, relied on relevance labels from trained judges. However, the benefits of crowdsourcing come with risks that need to be mediated. In crowdsourcing, engagement with workers is defined indirectly, through a Human Intelligence Task (HIT). Once HITs are posted on a crowdsourcing site, a self-selected group of individuals - the crowd, motivated by different incentives, completes the tasks with varying levels of attention and success. This increases the uncertainty about the quality of the outcome and the need for a careful design of HITs to attract the right crowd for a given task. Indeed, research has shown that the design of HITs influences the quality and the utility of the collected data.

In this talk I will detail some recent results from a series of crowdsourcing experiments that aim to contribute to our understanding of the factors that are involved in the self-selection and successful engagement of crowd workers in a given task. Our investigation compares data collected in the context of a relevance labelling task using two different HIT designs: with and without design features that control against random behaviour. In addition to the relevance labels, the HITs capture various task-related, e.g., feedback on the experience with the performed tasks, and task independent worker characteristics, e.g., demographics and personality profiles of individual workers. We use the collected data to analyze the correlation of the output quality with the characteristics of the workers attracted by the HITs. The findings reveal a clear segmentation of the crowd based on the HIT design. For example, we see that a design rich in quality control features attracts more conscientious workers, mostly from the US, while a simpler design attracts younger and less serious workers, mostly from Asia.

Marianne Sweeney, Principal,
Daedalus Information Systems,
Successful Enterprise Search by Design

When your colleagues say they want Google, they don't mean the Google Search Appliance. They mean the Google Search user experience: pervasive, expedient and delivering the information that they need. Successful enterprise search does not start with the application features, is not part of the information architecture, does not come from a controlled vocabulary and does not emerge on its own from the developers. It requires enterprise-specific data mining, enterprise-specific user-centered design and fine tuning to turn "search sucks" into search success within the firewall. This presentation looks at action items, tools and deliverables for Discovery, Planning, Design and Post Launch phases of an enterprise search deployment.

Matt Taylor, General Manager,
Funnelback,
“Search Experience Management - Techniques for Search Success”

Website search is a powerful contributor to user experience. Search applications can be used to deliver dynamic, personalised content and effective website search can dramatically increase conversions. This talk will present a number of techniques for successful website search, illustrated with case study demonstrations from City University, Skype, Dyson and more.

Iain Fletcher, VP Marketing,
Search Technologies,
"A New Content Processing Framework for Search Applications"

Almost all of the desirable search features offered by today's leading search engines depend on good data structure. Indeed, data issues are the most common cause of poor relevancy in existing search systems. Data sizes, especially unstructured data, continue to balloon, and enterprise search surveys continue to show user dissatisfaction. This presentation will propose a new approach to these data-driven problems which combines an implementation methodology and a purpose-built content processing framework based on open standards. Use of the methodology and framework will be illustrated by case examples, and key technical features of the framework will be explained and discussed.

Jarred McGinnis, Research Manager,

Press Association,

“The Press Association Approach to Search and News”

On a daily basis, the Press Association creates, curates and distributes vast amounts of content and data as well as providing a range of services to all major media organisations and many global corporations. Recently, the Press Association was appointed as both the host national news and data agency for the London 2012 Olympics. In order to address the growing demand for 'fast, fair and accurate' content and data, both from our customers and journalists, the Press Association recently committed to a large-scale project to redesign its IT infrastructure by putting semantic web technologies and principles at its core. These technologies allow the organisation to represent in a more human-friendly manner the diverse content and data spread across disparate systems. The use of ontologies, Linked Data and XML-based databases enables computer systems to understand the context as well as the explicit search criteria from a user. The Press Association Group is a global content operation with specific focus on news, sport and entertainment. The group operates a diverse range of businesses across the UK, Ireland, Europe, Canada and Asia as well separate brands specialising in areas such as Marketing (TNR) and Weather (Meteo Group).

Ian Kegel,

Technical Group Leader, British Telecom,

"Broadband TV and Recommendation: Improving the customer experience"

In recent years we have seen the arrival of a bewildering number of content providers, delivery platforms and devices capable of displaying TV content – mostly enabled by the ubiquity of broadband networks. However, the vast majority of TV viewing in the UK remains on 'live' broadcast channels, and when people do watch TV on demand, it is usually to catch up with a programme they recently missed on a broadcast channel. Helping their customers to find and watch programmes on demand from their extensive catalogues has therefore become a high priority for broadband TV content providers. Between 2008 and 2010, the MyMedia EU collaborative project developed an open framework for content recommendation and tested it in both IPTV and Internet-based scenarios. Its trials identified several challenges inherent in delivering effective content recommendations to the TV screen, such as the sparseness of on demand viewing and the inability to identify who is watching at any one moment. This presentation will review these challenges and how they are being addressed by content and service providers in current and future products. It will discuss BT's ongoing research into techniques for optimising recommender system performance in order to improve its customers' TV experience, and will also consider the impact of companion devices such as smartphones and tablets as more and more people interact with social media while watching TV. The presentation will conclude with an open discussion of the challenges discussed, and will encourage delegates to contribute their opinions and ideas.

Kristian Norling, Consultant

Findwise

"Information flow on the intranet at Region Västra Götaland"

At Region Västra Götaland the goal to give the right person, the right information at the right time, place and context on the intranet, led to a holistic view of information flow including the life cycle of information. Even though the search platform plays a very important part of the information flow, it exists in an ecosystem of systems supporting the intranet. Kristian will talk about how that ecosystem works in practice and show some examples.

Fishbowl session: "Hot topics in Search”

Want do discuss your own hot topic?

It’s a fishbowl session!