SIGIR 2012 Portland, Oregon, USA August 12–16, 2012

Industry Track: Schedule & Abstracts

Welcome to the Industry Track of SIGIR 2012!

The Industry Track's objectives are three-fold. The first objective is to present the state of the art in search and search-related areas, delivered as keynote talks by influential technical leaders from the search industry. The second objective of the Industry Track is the presentation of interesting, novel and innovative ideas related to information retrieval. Finally, a highly-interactive panel session will conclude the track.

Your Industry Track co-chairs:

Udo Kruschwitz, University of Essex, UK (udo@essex.ac.uk)
Su-Lin Wu, Google, USA (sulin.wu@gmail.com)

Session 1: NLP and IR

09:00-09:15	**Udo Kruschwitz and Su-Lin Wu** *Welcome and Introduction*
09:15-09:55	**Eric W. Brown (IBM Research)** *The Jeopardy! Challenge and Beyond*
09:55-10:15	**John O'Neil (Attivio)** *Entity Sentiment Extraction Using Text Ranking*
10:15-10:45	Coffee Break

Session 2: Emerging Trends in IR

10:45-11:25	**Andrei Broder (Google)** *IR Paradigms in Computational Advertising*
11:25-12:05	**Dan Rose (A9.com)** *CloudSearch and the Democratization of Information Retrieval*
12:05-14:00	Lunch Break / Business Meeting

Session 3: Personalization and Context

14:00-14:40	**Susan Dumais (Microsoft Research)** *Putting Context into Search and Search into Context*
14:40-15:00	**Ilya Segalovich (Yandex)** *Making Web Search More User-Centric: the State of Play and the Way Ahead*
15:00-15:20	**Mitul Tiwari, Linkedin (Joint work with Azarias Reda, Yubin Park, Christian Posse, and Sam Shah** *Related Searches at LinkedIn*
15:20-15:45	Coffee Break

Session 4: Social Networking and Panel Discussion

15:45-16:25	**Thomas Hofmann, Google Research, Zurich (Joint work with Enrique Alfonseca, Yasemin Altun, Katja Filippova, Massimiliano Ciaramita, Ioannis Tsochantaridis)** *Towards Summarizing the Web around Entities*
16:25-17:25	**Panel Session** chaired by David Hawking (Funnelback) Panelists: Trystan Upstill (Google), Jerome Pesenti (Vivisimo/IBM), Krishna Gade (Twitter), Stephen Robertson (Microsoft Research and City University London), Diane Kelly (University of North Carolina)

Eric W. Brown, IBM Research

Title: The Jeopardy! Challenge and Beyond

Abstract: Watson, named after IBM founder Thomas J. Watson, was built by a team of IBM researchers who set out to accomplish a grand challenge build a computing system that rivals a human's ability to answer questions posed in natural language with speed, accuracy and confidence. The quiz show Jeopardy! provided the ultimate test of this technology because the game's clues involve analyzing subtle meaning, irony, riddles and other complexities of natural language in which humans excel and computers traditionally fail. Watson passed its first test on Jeopardy!, beating the show's two greatest champions in a televised exhibition match, but the real test will be in applying the underlying natural language processing and analytics technology in business and across industries. In this talk I will introduce the Jeopardy! grand challenge, present an overview of Watson and the DeepQA technology upon which Watson is built, and explore future applications of this technology.

Bio: Eric Brown is a Research Staff Member and Manager at the IBM T. J. Watson Research Center. Eric earned his B.S. at the University of Vermont and M.S. and Ph.D. at the University of Massachusetts, all in Computer Science. Eric joined IBM in 1995 and has conducted research in information retrieval, document categorization, text analysis, question answering, bio-informatics, and applications of automatic speech recognition. Since 2007 Eric has been a technical lead on the DeepQA project at IBM and the application of automatic, open domain question answering to build the Watson Question Answering system. The goal of Watson is to achieve human-level question answering performance. This goal was realized in February of 2011 when Watson beat Ken Jennings and Brad Rutter in a televised Jeopardy! exhibition match. Eric's role on the project has spanned architecture development, special question processing, and hardware planning, and he is currently focused on applying Watson to clinical decision support in Healthcare.

John O‘Neil, Attivio

Title: Entity Sentiment Extraction Using Text Ranking

Abstract: Entity extraction and sentiment extraction are among the most common types of information extracted from documents. Both have been studied extensively, with different approaches. However, the problem of directly associating entities and sentiment has received less attention.

We assume that entities of interest have already been extracted in the document. Entity extraction is well-understood, and can be done in numerous ways. Also, we assume that appropriate word sentiment weights have been calculated. We use a training set of documents, where each document is associated with a positive or negative sentiment label, and use standard discriminative supervised machine learning to calculate sentiment weights for each word appearing in the corpus. We also use a stopword list and an absolute value weight cutoff to derive a relatively small set of word-sentiment values.

Then, we use Text Ranking (Mihalcea 2004), adapting it to use a bipartite graph of entities and sentiment-laden words and phrases, where entities and sentiment words are connected if within a configurable distance. We calculate the dominant eigenvector of the matrix corresponding to the graph; we then extract final sentiment weights to the entities, given all the contexts each entity appears in. Also, we can more accurately determine the sentiment weights of the words in the document given their contexts, as well as discovering the positive or negative sentiment “hot spots” in a document.

We then explore a number of related issues. First, we explore if (and how much) this improves the accuracy of the sentiment associated with entities in a document, compared to assigning the document's overall sentiment to all the entities in the document. We also explore if (and how much) of an accuracy improvement there is if we track the entity sentiment over time. Then, we explore whether faceting on entity sentiment is improved using this technique in the context of search applications. Finally, we consider performance.

Bio: John O‘Neil is the Chief Scientist at Attivio. He has written and designed software for search, natural language processing and machine learning for more then a decade. After receiving a Ph.D. in linguistics from Harvard University, he worked for LingoMotors, where he designed their main commercial search product, and at other search and IR companies. He also worked for over five years at Basis Technology, Inc., where he was the designer and lead developer for the Rosette Linguistics Platform, their language processing and entity extraction suite of products. He is the author of more than twenty papers in Computer Science, Linguistics and associated fields, and has given talks at numerous professional and academic conferences.

Andrei Broder, Google

Title: IR Paradigms in Computational Advertising

Abstract: The central problem in the emerging discipline of computational advertising is to find the “best match” between a given user in a given context and a suitable advertisement. The context could be a user entering a query in a search engine (“sponsored search”), a user reading a web page (“content match” and “display ads”), a user streaming a movie, and so on. In some situations, it is desirable to solve the “dual” optimization problem: rather then find the best ad given a user in a context, the goal is to identify the “best audience”, i.e. the most receptive set of users and/or the most suitable contexts for a given advertising campaign. The information about the user can vary from scarily detailed to practically nil. The number of potential advertisements might be in the billions. Thus, depending on the definition of “best match” and “best audience” these problems lead to a variety of massive optimization problems, with complicated constraints, and challenging data representation and access issues.

In general, the direct problem is solved in two stages: first a rough filtering is used to determine a relatively small set of ads that should be considered as potential matches, followed by a more sophisticated secondary ranking where economics considerations take center stage. Historically, the filtering has been conceived as a database selection problem, and was done using simple Boolean formulae, for instance, in sponsored search the filter could be "all ads that provide a specific bid for the present query string or a subset of it". Similarly for the dual problem (audience definition) for, say, a sports car ad, the filter could be "all males in California, aged 40 or less".

This "database approach" for the direct problem has been recently supplanted by an "IR approach" based on a similarity search between a carefully constructed query that captures the advertising opportunity and an annotated document corpus that represents the potential ads. Similarly, in the dual problem, the newer approach is to devise an efficient and effective representation of the users, then form a query that represents a prototypical ideal user, and finally find the users most similar to the prototype. The aim of this talk is to discuss the penetration of the IR paradigms in computational advertising and present some research challenges and opportunities in this area of enormous economic importance.

Bio: Andrei Broder has recently joined Google as a Distinguished Scientist. Previously, he was a Fellow and Vice President for Computational Advertising in Yahoo!. Prior to this, he worked at IBM as a Distinguished Engineer and the CTO of the Institute for Search and Text Analysis and at AltaVista as Vice President for Research and Chief Scientist. He was graduated Summa cum Laude from Technion, the Israeli Institute of Technology, and obtained his M.Sc. and Ph.D. in Computer Science at Stanford University. His current research interests are centered on computational advertising, user understanding, context-driven information supply, and randomized algorithms. Broder has authored more than a hundred papers and was awarded thirty-six patents. He is a member of the US National Academy of Engineering, a fellow of ACM and of IEEE, and past chair of the IEEE Technical Committee on Mathematical Foundations of Computing.

Dan Rose, A9.com

Title: CloudSearch and the Democratization of Information Retrieval

Abstract: Amazon CloudSearch is a new hosted search service, built on top of many cloud-based AWS services, and based on the same technology that powers search on Amazon's retail sites. Because of its ease of configuration and scalability, CloudSearch represents the next step in the democratization of information retrieval. This democratization process, increasing access to search for both end users and potential search providers, has continued over several decades, through technologies like early online metered search services, enterprise search software, web search, and open source search tools. CloudSearch further reduces barriers to entry, allowing a person or organization to basically say "make my content searchable" and have it happen automatically. CloudSearch may also offer an opportunity to overcome the stagnation that has occurred in search user experiences over the past 15 years. When you no longer need to be a search expert to make your content available, you're not stuck with ten blue links. Instead, you can focus on providing the kind of interaction that makes sense for your application and your users. CloudSearch enables a flowering of search applications that need not be tied to the web, and an opportunity to explore new ways of interacting with information retrieval technology.

Bio: Daniel E. Rose is the Chief Scientist for Search at A9.com, a subsidiary of Amazon.com. Prior to A9, he held key positions at Yahoo! and AltaVista. Earlier, Dan worked at Apple's Advanced Technology Group where he led the Information Access Research team, which created desktop search for the Macintosh. He has worked on many aspects of information retrieval, from relevance ranking to posting list compression to improving the search user experience. Dan holds a Ph.D. in Cognitive Science and Computer Science and an M.S. in Computer Science, both from UC San Diego, as well as a B.A. in Philosophy from Harvard University. He has published a variety of technical articles, focusing on IR and human-computer interaction, and holds over 20 issued U.S. patents.

Susan Dumais, Microsoft Research

Title: Putting Context into Search and Search into Context

Abstract: It is very challenging task to understand a short query, especially if that query is considered in isolation. Luckily, queries do magically appear in a search box - rather, they are issued by real people, trying to accomplish a task, at a given point in time and space, and this "context" can be used to aid query understanding. Traditionally search engines have returned the same results to everyone who asks the same question. However, using a single ranking for everyone, in every context limits how well a search engine can do. In this talk I outline a framework to quantify the "potential for personalization", that can be used to characterize the extent to which different people have the same (or different) intents for a query. I will then describe several examples of how we represent and use different kinds of context to improve search quality. Finally I conclude by highlighting some important challenges in developing such systems at Web scale including system optimization, evaluation, transparency and serendipity.

Bio: Susan Dumais is a Principal Researcher and manager of the Context, Learning and User Experience for Search (CLUES) Group at Microsoft Research. Prior to joining Microsoft Research, she was at Bellcore and Bell Labs for many years, where she worked on Latent Semantic Indexing (a statistical method for concept-based retrieval), interfaces for combining search and navigation, and organizational impacts of new technology. Her current research focuses on user modeling and personalization, context and information retrieval, temporal dynamics of information systems, interactive retrieval, and novel evaluation methods. She has worked closely with several Microsoft groups (Bing, Windows Desktop Search, SharePoint Portal Server, and Office Online Help) on search-related innovations. Susan has published more than 250 articles in the fields of information science, human-computer interaction, and cognitive science, and holds several patents on novel retrieval algorithms and interfaces. Susan is also an adjunct professor in the Information School at the University of Washington. She is Past-Chair of ACM's Special Interest Group in Information Retrieval (SIGIR), and serves on several editorial boards, technical program committees, and government panels. She was elected to the CHI Academy in 2005, an ACM Fellow in 2006, received the Gerard Salton Award from SIGIR for Lifetime Achievement in 2009, and was elected to the National Academy of Engineering (NAE) in 2011.

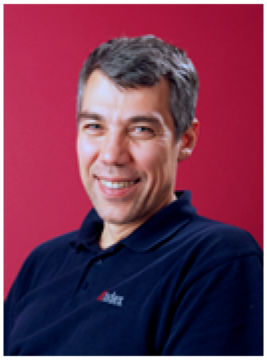
Ilya Segalovich, Yandex

Title: Making Web Search More User-Centric: the State of Play and the Way Ahead

Abstract: In this talk I will cover some critical challenges in web search, arising from the need to make it more personalized and user behavior driven. Some of the problems are more specific to non-English markets and users with certain cultural and linguistic background. I will start from the overview of several user-focused cutting-edge Yandex technologies, such as personalization of search by inferring the level of foreign language knowledge of our users and their socio-demographic features. Then, I will demonstrate how search can be improved using short-term and long-term user search context.

In the second part of my talk, I will explain how cross-lingual IR techniques can be applied to the problem of query understanding and will continue with such topics as balancing relevancy and freshness in search results and query suggestions, and adapting off-line and on-line retrieval quality measures to the real-world demands.

I will also overview various recent initiatives started by Yandex, aimed to support public research and engage the IR community in exploration of advanced topics in web search in order to consolidate and scrutinize the work started at industrial labs.



Bio: Ilya Segalovich is one of Yandex‘s co-founders and has been Yandex‘s Chief Technology Officer and a director since 2003. He began his career working on information retrieval technologies in 1990 at Arcadia, where he headed its software team. From 1993 to 2000, he led the retrieval systems department for CompTek International. Mr. Segalovich received a degree in geophysics from the S. Ordzhonikidze Moscow Geologic Exploration Institute in 1986. Ilya also took an active role in starting Russian research and scientific initiatives in information retrieval and computational linguistics. He gave invited talks at SIGIR 2010 and CIKM 2011 Industry days.

Mitul Tiwari, Linkedin (Joint work with Azarias Reda, Yubin Park, Christian Posse, and Sam Shah

Title: Related Searches at LinkedIn

Abstract: Search plays an important role in online social networks as it provides an essential mechanism for discovering members and content on the network. Related search recommendation is one of several mechanisms used for improving members' search experience in finding relevant results to their queries. This talk describes the design, implementation and deployment of Metaphor, the related search recommendation system on LinkedIn, a professional social networking site with over 160 million members worldwide. Metaphor builds on a number of signals and filters that capture several dimensions of relatedness across member search activity. The system, which has been in live operation for close to two years, has gone through multiple iterations and evaluation cycles. We will describe the signals used in Metaphor, their scalability, and the practical concerns in deploying related search recommendations.

Bio: Mitul Tiwari is a Senior Research Engineer at Search, Network, and Analytics group at LinkedIn. Previously, he worked at Kosmix (now Walmart Labs) as a Member of Technical Staff. He completed his PhD in Computer Science from University of Texas at Austin in 2007. Earlier he received his under graduation from Indian Institute of Technology, Bombay. At LinkedIn, he is working on data driven products such as “People You May Know” and “Related Searches”. His interests include information retrieval, large-scale data mining, machine learning, and distributed systems.

Thomas Hofmann, Google

Title: Towards Summarizing the Web around Entities

Abstract: We want to approach the challenge of how to best summarize content published on the Web. Our ultimate goal is to be able to answer questions like: “what does the Web know or say about X?”, where X can be some entity of interest. It is our belief that pivoting multi-authored and transient (e.g. user-generated) content around concrete entities such as persons, organizations, locations, etc., but possibly also around abstract concepts, provides a sensible organization principle that offers a value-add to users relative to the views of the Web created by search engines, news / blog sites, or social sites.

The above challenge naturally decomposes into the following subproblems: (i) entity base curation, i.e. semi-automatic generation of a catalog of entities alongside with information around relations and naming, (ii) entity linking, i.e. automatically detecting and linking entities in Web documents, (iii) document or entity mention clustering, i.e. grouping related (redundant as well as complimentary) documents or passages together, (iv) multi-source summarization of a given cluster relative to a query entity, (v) summary display, where we investigate timeline views. Obviously, each of these subproblems is a significant challenge on its own. The purpose of this talk is thus not to develop near-perfect solutions to any one of these, but to present an end-to-end prototype system, which is a first step towards the ambitious overall goal.

Bio: Thomas Hofmann received a Ph.D. in Computer Science from the University of Bonn in 1997 and subsequently held postdoctoral positions at MIT and at UC Berkeley and the International Computer Science Institute. From 1999 until 2004 he was Assistant & Associate Professor in the Computer Science Department at Brown University. Between 2004 and 2006, he held a position as a Full Professor of Computer Science at the Technical University of Darmstadt, while also serving as the Director of the Fraunhofer Institute for Integrated Publication and Information Systems. He is also co-founder and former CEO & Chief Scientist of Recommind Inc, a privately owned, global company focusing on enterprise search and predictive technologies & text analytics. Since July 2006, Thomas is Director of Engineering at Google and one of the site leads of Google's engineering center in Zurich, Switzerland. He has lead projects in various areas, including Web search, e-commerce, and internet advertising. His scientific interests are in machine learning, natural language understanding, and information retrieval. He has published 60+ scientific papers in these areas. Since 2012, he is also an adjunct professor in Computer Science at ETH Zurich.

Panel Information

Panel Chair: David Hawking, Funnelback

Abstract: The early pioneers of IR would probably be flummoxed (or even gobsmacked) by the scale, diversity, pervasiveness and economic importance of current industrial applications of IR research. They would no doubt also be delighted at the extent to which industrial applications have extended research frontiers and spurred further discoveries. The 2012 Industry Panel will canvass the diversity of IR in modern practice and will explore differences and similarities in viewpoints between stakeholder groups. Each of our distinguished panelists has been asked to attempt to represent the likely viewpoint of a particular group or “vertical” while responding to a series of questions notified in advance. These questions will cover perspectives, perceived challenges, opportunities and roadblocks. Trystan Upstill of the Search Quality Team at Google will represent “large scale web search”; Jerome Pesenti, Chief Scientist at Vivisimo / IBM, will represent “enterprise search”; Krishna Gade, Engineering Manager at Twitter will represent “real-time and social search”. Stephen Robertson, Emeritus Professor at City University, London and MSR will represent “academic research” and Diane Kelly, author of an influential Foundations and Trends monograph on user-involved evaluation will take the all-important perspective of “users”. The audience, too, will have their say!

Bio: David Hawking was a Coordinator of the TREC Web Track from 1997-2004 and SIGIR Program Chair in 2003 and 2006. He bridges the academic-industry divide, working as Chief Scientist at the Funnelback internet and enterprise search company (of which he is a founder), and supervising doctoral students at the Australian National University.