

NGS08

Third Workshop of the HCSNet Next-Generation Search Technology Priority Area

Program, Information and Abstracts

**State Library of Victoria
Conference Centre
Melbourne
13th November 2008**



Contents

Welcome.....	1
About HCSNet.....	2
Registration and General Information	3
Venue.....	3
Public Transport to the Library	3
Parking	3
Registration Desk.....	3
Delegate Materials	3
Travel Support.....	4
Refreshments and Meals	4
Program.....	5
Abstracts.....	7
Long Presentation 1	7
Brett Poole	7
Short presentations - Session 1	7
Paul Watters.....	7
Alexandra Uitdenbogerd	7
Tom Rowlands	7
Student presentations - Session 1	8
Sharmin Choudhury	8
Luiz Augusto Pizzato.....	8
Timothy Jones	8
Andrew Lampert.....	9
Long Presentation 2	10
David Hawking	10
Short Presentations - Session 2.....	10
Alexander Krumpholz.....	10
Cecile Paris	10
David Martinez	11
Jai Gupta.....	11
Short presentations - Session 3.....	11
Jim Brander.....	11
Tuomo Kakkonen	12
Peter Eklund.....	12
Student presentations - Session 2	12
Jack Tseng.....	12
Kenneth Treharne	13
Tina Du.....	14
Workshop Contact List	16

Welcome

Welcome to the NGS08 - Third Workshop of the HCSNet Next-Generation Search Technology Priority Area.. Existing information retrieval systems effectively treat documents as unstructured bags of words. As current Web search engines demonstrate, this approach works surprisingly well. At the same time, it is clear that human processors of information make use of a much deeper understanding of text than these systems exhibit. Humans cannot compete with machines in terms of quantity, but their abilities far exceed those of machines when it comes to quality.

Linguistics treats texts as richly structured objects that obey complex and interacting rules about language use, and natural language processing attempts to implement computational models that embody these ideas: how do we add this sophistication to information retrieval in a way that scales and which delivers better results? Are there insights from the cognitive sciences that can tell us how to build better tools for finding information? How do we extend these technologies when the data we are concerned with includes audio and video as well as text?

The aim of HCSNet's Priority Research Area in Next Generation Search is to bring together researchers from a wide range of disciplines to address these questions.

Dr Diego Molla-Aliod
Prof Robert Dale

Macquarie University
Macquarie University

About HCSNet

The ARC Research Network in Human Communication Science – HCSNet - was awarded five years' funding by the Australian Research Council in late 2004. The aim of HCSNet is to promote and facilitate interdisciplinary research in human communication science by connecting leading researchers in language, speech and sonics.

Priority Research Areas in HCSNet are:

- Speech
- Effective Interfaces
- Next-Generation Search Technology
- Human Communication Disorders
- Perception and Action

By generating an explosion of new approaches and knowledge, the network aims to build Australia's reputation as a leader in communication science and technology via advances in areas as diverse as automatic speech recognition, distress call monitoring, hearing prostheses, web interfaces, and data retrieval and data mining systems.

Getting involved in HCSNet is easy: visit www.hcsnet.edu.au to sign up as a member of the network. You'll be added to our online profile database, and automatically receive our weekly electronic newsletter, HCSNet Update, which will keep you informed of HCSNet activities, including the annual SummerFest, and events in the range of HCSNet disciplines. Australian-based HCSNet members can apply for funding under our various programs.

Registration and General Information

Venue

State Library of Victoria, Conference Centre
Entry 3, La Trobe Street
Level 2A

Public Transport to the Library

The Library's location in the centre of Melbourne means that it is very easy to visit using public transport. Whether you are travelling by train, tram, bus or taxi, the area is well served by all forms of public transport. Maps, timetables, route and fare information is available at Metlink, the official online guide to public transport in the Melbourne metropolitan area.

Trains - Directly in front of the main entrance to the Library is Melbourne Central Station.

Trams - Tram stops are located on each corner of the La Trobe and Swanston Streets intersection.

Buses - Buses are located in nearby Russell and Exhibition Streets.

Taxis - The nearest taxi rank is located in Elizabeth Street outside the Melbourne Central shopping complex. However, if you want to try your luck hailing a taxi, there are usually plenty of taxis roaming the CBD.

Parking

While the Library does not have any onsite parking, there is ample parking nearby.

Street parking - There is no parking in Swanston Street and limited spaces around the corner in La Trobe Street throughout the week. Offering two-hour meter parking, the south and north sides of La Trobe Street are designated tow-away zones during the morning and evening peak hours.

Disabled parking - There are three disabled parking spaces in La Trobe Street, near the corner of Swanston Street. Available from 9.30am to 10pm, from Monday to Friday and any time on Saturday and Sunday, these metered spaces have a two-hour time limit. Vehicles using these spaces must have a disabled parking permit clearly displayed.

Commercial car parks - The area is well-served by three commercial car parks located within walking distance of the Library. Two are located in La Trobe Street and another in Lonsdale Street.

Registration Desk

The registration desk is the place for enquiries related to registration, travel support reimbursement claims or the local area.

Delegate Materials

Delegate materials include the following: Program, Information and Abstracts Book, name tag, pen and notepad, Workshop evaluation form, HCSNet Survey, and HCSNet brochure.

Travel Support

Workshop participants who have been awarded travel support to attend the workshop are asked to submit travel and accommodation receipts at the workshop to Dr Diego Molla-Aliod if possible to enable reimbursements to be made before Xmas. Alternatively the postal address to mail them is:

Chris Cassidy
HCSNet Administrative Coordinator
Department of Computing
Division of Information and Communication Sciences
Macquarie University NSW 2109
Email: ccassidy@ics.mq.edu.au

Refreshments and Meals

Lunch and Afternoon Tea are provided for all registrants, note that lunch will be provided between 11.30am – 1.00pm.

Program

9.00am – 9.15am	Opening
9.15am – 9.45am	Long Presentation 1 Brett Poole <i>Yahoo! and Next Generation Search Experience</i>
9.45am – 10.30am	Short Presentation 1 Paul Watters, Alexandra Uitdenbogerd and Tom Rowlands
10.30am - 11.30am	Student Presentations 1 Sharmin Choudhury, Luiz Augusto Pizzato, Timothy Jones and Andrew Lampert
11.30am – 13.00pm	Lunch Break
13.00pm – 13.30pm	Long Presentation 2 David Hawking <i>Promoting Diversity in Search</i>
13.30pm – 14.30pm	Short Presentations 2 Alexander Krumpholz, Cecile Paris, David Martinez and Jai Gupta
14.30pm – 15.00pm	Afternoon Break
15.00pm – 15.45pm	Short Presentation 3 Jim Brander, Tuomo Kakkonen and Peter Eklund
15.45pm – 17.00pm	Student Presentations 2 Jack Tseng, Kenneth Treharne and Tina Du
17.00pm – 17.15pm	Closing

Abstracts

Long Presentation 1

Brett Poole

Yahoo! and Next Generation Search Experience

Find out what Yahoo! is doing to define the next generation search user experience. Yahoo! has opened up its developer platform and made available tools to interact directly with its search results page. Also, we will cover new development environments accessible to developers and start-ups to build next generation search solutions that can compete head-to-head with the principals in the search industry.

Short presentations - Session 1

Paul Watters

The Data Access Project at the UK Medical Research Council

In 2007, the Data Access Project at the UK Medical Research Council faced a dilemma - to choose a highly structured, ontological search function to allow external users to identify metadata of interest, or to choose a relatively unstructured "bag of words" approach. The longitudinal dataset comprised some 20,000 variables stored in a variety of cleaned, normalised and derived formats, the structure of which had fluctuated over the past 60 years. I will present some ideas on the usability of the system, drawn from the first round of expert user feedback, and indicate where I feel more structured search could be useful to similar projects.

Alexandra Uitdenbogerd

Query-by-singing Music Retrieval

We are currently embarking on the development of robust techniques for query-by-singing music retrieval systems. The challenges that are posed by matching melodies sung by users against music in an on-line collection include not only the potential inaccuracies that are likely with singing by non-habitual singers, but variations in the acoustic environment, the identification of note onset times, and the acoustic properties of different voices.

Tom Rowlands

Direct query to annotation matching for search

Textual annotations present a valuable additional source of external evidence. Anchortext, click associated queries and folksonomy tags have all been shown to be valuable. Concatenated, as surrogate documents, annotations can deliver good information retrieval results for popular searches and for finding key resources.

However, this technique may, in some cases, lead to false positives by not honoring annotation boundaries, or false negatives by not facilitating the lack of word boundaries. We wish to investigate opportunities in direct query to annotation matching. Simple methods include exact, substring, and intersecting term set matching. More advanced techniques include adding stemming, multi-annotation to query matching, using past query logs, and introducing translation.

Student presentations - Session 1

Sharmin Choudhury

Ontology based perspective determination and its implications for searching

The Motion Picture Industry generates and uses a vast quantity of data for a variety of purposes. However, limitations of current information search methods have led to inefficiencies and introduced difficulties in finding the right information. We propose a domain ontology based search methodology to add a layer of semantics to search engine. Using the ontology the perspective of the user in relation to the concepts present in a given query can be determined. This may in turn lead to a better search experience for the user, as well as more relevant results being returned from a search. In order to achieve this we are developing the Loculus: a domain ontology for the Motion Picture Industry. The Loculus ontology is to be used with the Loculus system to provide a tool which the industry practitioners can use to better utilize the vast store of information that is constantly expanding.

Luiz Augusto Pizzato

Indexing on Semantic Roles for Question Answering

Semantic Role Labeling (SRL) has been used successfully in several stages of automated Question Answering (QA) systems but its inherent slow procedures make it difficult to use at the indexing stage of the document retrieval component. In my presentation, we will demonstrate how SRL can be used to index documents and how it can improve QA performance. We will also discuss whether simpler, but faster models such as the Question Prediction Language Model (QPLM) are more beneficial to IR and QA than traditional SRL.

Timothy Jones

Investigating the effect of spam results on user experience

Spam web pages are pages that are designed to manipulate the relevance ranking provided by web search engines. The presence of spam pages in search engine results can damage result quality. A result from an earlier user experiment surprisingly suggests that while the presence of spam pages in search results does damage result quality, this damage does not increase as the number of spam pages present in the top ten search results increases. In this talk, we outline a further series of user experiments to investigate the effect of web spam on search result quality from a user perspective. These experiments compare the effect of spam documents in search result pages to the effect of irrelevant documents in search result pages, and the effect on result quality as

the amount of spam present in result pages increases. The experiments we outline should generate a more complete understanding of user reaction to web spam.

Andrew Lampert

Improving Email Search and Task Management

Search functionality in many existing email clients is notoriously slow and unhelpful. In particular, representing the content of email messages as unstructured bags of words is limiting, particularly for the many people who employ their email client as a task-tracking tool. Several studies have identified problems with "keeping track of lots of concurrent actions: One's own to-dos and to-dos one expects from others" using existing email clients [1]. We are working to augment existing email software with the ability to automatically identify incoming and outgoing requests and commitments. Uncovering this latent task information is an important step towards incorporating aspects of workflow and priority in email clients to provide better support for task management.

More generally, we are interested in how email search can be improved, beyond our work on better task management. In particular, we are interested in addressing email-specific search challenges such as how to make use of click data and how to compensate for the lack of citation-style external evidence used in web search by algorithms such as PageRank. We're also interested in discussing ideas that exploit the latent social graph, conversation structure and other characteristics of email to improve the utility and effectiveness of email search and task management.

[1] V. Bellotti, N. Ducheneaut, M. Howard and I. Smith. "Taking email to task: The design and evaluation of a task management centred email tool", in Proceedings of CHI 2003, Ft Lauderdale, Florida, pp. 345-352.

Long Presentation 2

David Hawking

Promoting Diversity in Search

Across a searcher population, diversity in search results is essential to accommodate divergent user needs and interpretations. An individual searcher may also appreciate diversity of viewpoints, diversity of sources, and lack of repetition in results. There are a range of methods for promoting diversity. Diversity can be encouraged in ranking, but there are alternative ways of presenting diversity in results, and ways in which diversity within a ranked list can be exposed and accessed. There is a need for evaluation methods to model needs for diversity.

Short Presentations - Session 2

Alexander Krumpholz

XML tag type classification for structured document retrieval

Structured document (e.g. XML) retrieval systems aim at retrieving the most specific elements of a document matching given query terms. XML elements can be used to mark-up the structure of a document by using different tags to declare text blocks like chapters, paragraphs or headings. However, XML elements are also used to specify regions of different style like font size. Both types of elements are usually treated equally in the retrieval process. Ignoring the style tags and only considering structure tags as potential retrieval units should improve the retrieval quality. We split the documents into sentences using standard NLP techniques and identify tags that cut through the sentence structure. Such elements are likely to be of style tag nature. Clustering can be used to identify structural elements which are potential retrieval units, and stylistic elements which could be used in term up-weighting or down-weighting.

Cecile Paris

Summarisation for Search on Heterogeneous Sources

Often, finding information relevant to a user's query may involve searching across multiple heterogeneous data sources, including both online public and offline personal data. Meta-search engines typically send user queries to each of these sources of data, providing a common interface with which to view the results. However, while such search engines may save the user time by automatically farming queries out to the many possible data sources, the question of how best to present results remains. For example, search results could be presented for each data source in isolation or interleaved and a single ranked list. Either method requires the user to manually integrate the search results, identifying how the different documents relate to one another. In our work, we are exploring the use of summarisation techniques to help provide overviews of the information space. In particular, we look at using update summarisation methods to generate "Supplement Summaries". That is, a generated

summary may high-light how data from one source complements and differs from data from another source. In this way, the user may be assisted in identifying how different data sources relate to each other.

David Martinez

ILIAD (Improved Linux Information Access by Data Mining)

The project ILIAD (Improved Linux Information Access by Data Mining) is concerned with information access in the domain of Linux troubleshooting, based on Linux web user forum data. In this 3-year ARC project (2006-2008), we are crawling the information contained in different user forums and mailing lists (e.g. linuxquestions), and studying better ways to deliver this information by relying on Language Technologies (LT) and domain-specific tools.

The information extraction tools that we have developed include tailored classifiers to determine forum-thread characteristics (e.g. solved vs unsolved threads) and domain-specific LT tools (i.e. Named Entity Recognition and deep parsing). We have also studied the role of features such as the author profiling or the use of semantic vector indexing for the document collection.

Finally, we apply these tools to an Information Retrieval task that simulates the information needs of real forum participants, and boost the retrieval of relevant pieces of information by relying on our domain-specific tools and features.

Jai Gupta

In search of better retrieval system: learning from our brain

Human brain operates with approximately 20,000 – 30,000 lexicon (as against 91,600 unique lexicon claimed by WordNet) but it is still an efficient retrieval system as compared to the super fast search engines and still go wrong. This efficiency come from it's ability to store over billions of pages of information within a frame of semantic network. The feeders or input channels to this system are primarily sensory in nature from developmental perspective. The input modules work within a parallel and/or distributed processing models. For some retrieval task parallel process is more dominant and for some distributed process and or some parallel or distributed alone is sufficient. If a general purpose semantic model integrating is developed incorporating inputs from various sciences, we should be able to develop a superior retrieval system that is more semantic based and that does not heavily depend on lexical properties at word, syntax and text level.

Short presentations - Session 3

Jim Brander

Internet Searching Using Active Structure

Active Structure, when used for searching in free text, involves the processes of

- Tagging
- Parsing
- Structure building
- Stitching together of the discourse
- Building a search structure

When attempting to create an accurate and complete representation of the knowledge in a document, it runs at about the speed of an attentive human reading for full understanding in one pass (very slowly, in other words). Internet search, with its billions of pages across all subjects, is a different problem. We are currently scoping the problem by estimating the time on one aspect of an enterprise website, then looking to broaden the base without losing too much meaning in the process. Millions of machines is not an obstacle - Google has hundreds of thousands for key word searching - but routing the knowledge would be.

Tuomo Kakkonen

POSELSA - Extending Latent Semantic Analysis with Part-of-speech Information

Latent Semantic Analysis (LSA) is a widely used Information Retrieval method based on the "bag-of-words" assumption. According to general conception, however, syntax plays a key role in representing meaning of a text, which means that models such as LSA that take this simplifying assumption are inadequate. Several approaches have been applied in order to enhance LSA with syntactic and morphological knowledge, but most of them have failed to improve on the performance of the basic bag-of-words LSA, and somewhat surprisingly, have resulted in worse performance.

This talk introduces several LSA models in which part-of-speech (POS) information is used for adding local information about the internal relations between the words in sentences. Our experiments on an agglutinative language that has a relatively free word order, Finnish, show that the addition of morphological and syntactic information in the form of POS tag sequences can significantly improve the performance of LSA.

Peter Eklund

Local Analysis of Web Search using Formal Concept Analysis

SearchSleuth is a program developed to experiment with the automated local analysis of Web search using formal concept analysis. The program extends a standard search interface to include a conceptual neighborhood centered on a formal concept derived from the initial query. This neighborhood of the concept derived from the search terms is decorated with its upper and lower neighbors representing more general and more specialized concepts respectively.

Student presentations - Session 2

Jack Tseng

Improving Visual Search Process using User Centered Query Concept Mapping Model

Research in visual retrieval has been focusing on the various approaches to bridge the semantic gap between user and system for the past few years. Recently, as led by NIST TRECVID research, the contributions have been confined in query concept mapping (QUCOM) techniques which try to detect concept terms from user queries and match with pre-defined concept terms in the system. Some achievements have been shown in both lexical and statistical approaches. Concept-based systems which utilize semantic objects or annotations for retrieval have also been proposed (eg., the ALIPR system). However, none of them has addressed the issues of the ever changing nature in user's query intent, despite the prominence of such phenomenon has been revealed in many search log analysis studies, especially for Web searches.

In this research, we present a new approach to improve query concept extraction and mapping from consecutive modifications in user's queries. From our preliminary query modification analysis, we have discovered the dominance of constant replacing modification sequence (i.e. substitute with some new terms while leaving other terms in subsequent queries) in user's query modification behavior. Query concept mapping based on these replacing modification sequences should improve modeling the dynamics of query concept movement. The retrieval process can also be enhanced by suggesting related concept terms according to the salient concepts identified from query modifications. We aim to tackle visual retrieval problems from user's perspective, as well as to expand current state-of-the-art query processing and concept mapping techniques to achieve user-centered visual search.

Kenneth Treharne

Human Factors in Visualisation for Improvement of Information Search

Information retrieval (IR) systems are as only as good as the underlying implementation allows them. For any information search, IR systems cannot know what the human user is thinking and often the human user does not exactly know what they are looking for! Therefore, we should now focus more on both the human factors and the interface (i.e. search tools) to learn how to best present IR results and what sub tasks (new or old) a user engages to effectively and efficiently search. The Flinders AILab is interested in evaluating tools and techniques that explore alternatives to the result presentation interfaces of current IR systems. Of particular interest are graphically augmented representations in opposition to current interfaces that present predominantly textual summaries of potentially useful document, video and audio media. AILab research is exploring techniques for the graphic representation of IR result sets in an interactive information map/space using latent semantic analysis, and evaluating techniques for Metadata visualisation in both 2D and 3D; all the while, analyses of the user logs are revealing an interaction-dialogue taking place between the user and the IR system. With a better understanding of the human factor, we will implement search engines in such a way as to elicit more of the unspecified and background information associated with information queries that search engines need but currently only 'guestimate'. An online delivery methodology for each evaluation apparatus ensures a wider and more valid experimental population whilst still collecting a rich dataset of precise measurements for information retrieval task performance. Our evaluations use a variety of technologies, including Java Applets and the Mozilla Firefox Extension facility for our Flingle IR result visualisation extension.

Tina Du

New Cognitive Model of Web IR Interaction: Integrating Multitasking, Cognitive Coordination and Cognitive Shifts

Web searching is an important element of information behaviour and human computer interaction which includes multitasking processes and the allocation of cognitive resources among several tasks, and shifts in cognitive, problem and knowledge states. In addition to multitasking, cognitive coordination and cognitive shifts are also important but under-explored aspects of Web searching. This project aims to model the relationship between multitasking, cognitive coordination and cognitive shifts during Web information retrieval. Key findings based on the pilot study include: (1) the major factor influencing multiple information problems search ordering was personal interest; (2) participants experienced complex cognitive coordination process embedded within Web search interaction; (3) study participants experienced both positive and negative cognitive shifts at all levels; (4) Web search interaction is shown to be a multitasking process during which information problems ordering, task switching, task and mental coordinating occur, and at deeper level, cognitive shifts take place. A model based on the key findings is provided to illustrate this relationship. Also the implications for more effective Web based IR systems development are discussed.

Workshop Contact List

Surname	First Name	Affiliation	Email address
Bennett	Graham	NAB	Graham.J.Bennett@nab.com.au
Brander	Jim	Interactive Engineering	jim.brander@pacific.net.au
Choudhury	Sharmin	Queensland University of Technology	t.choudhury@qut.edu.au
Dale	Robert	Macquarie University	rdale@ics.mq.edu.au
Du	Jia Tina	Queensland University of Technology	jia.du@student.qut.edu.au
Eklund	Peter	University of Wollongong	peklund@uow.edu.au
Gupta	Jai	Ballarat University	guptajs1@gmail.com
Hawking	David	Funnelback / ANU	david.hawking@acm.org
Jones	Timothy	ANU	tim.jones@anu.edu.au
Kakkonen	Tuomo	University of Joensuu, Finland	tkakkone@ics.joensuu.fi
Krumpholz	Alexander	ANU/CSIRO	alexander.krumpholz@csiro.au
Lampert	Andrew	CSIRO ICT Centre / Macquarie University	Andrew.Lampert@csiro.au
Martinez	David	University of Melbourne	davidm@csse.unimelb.edu.au
Molla-Aliod	Diego	Macquarie University	dmollaaliod@gmail.com
Paris	Cecile	CSIRO	Cecile.Paris@csiro.au
Pizzato	Luiz Augusto	Macquarie University	pizzato@ics.mq.edu.au
Pohl	Stefan	NICTA, The University of Melbourne	spohl@csse.unimelb.edu.au
Poole	Brett	Yahoo!	brettp@yahoo-inc.com
Rowlands	Tom	CSIRO / ANU	tom.rowlands@csiro.au
Sarvnaz	Karimi	NICTA	skarimi@unimelb.edu.au
Scholer	Falk	RMIT University	falk.scholer@rmit.edu.au
Treharne	Kenneth	Flinders University	kenneth.treharne@flinders.edu.au
Tseng	Jack	Queensland University of Technology	j.tseng@qut.edu.au
Uitdenbogerd	Alexandra	RMIT University	sandrau@rmit.edu.au
Watters	Paul	University of Ballarat	p.watters@ballarat.edu.au
Wu	Mingfang	RMIT University	mingfang.wu@rmit.edu.au