



ADCS 2007

Proceedings of the Twelfth Australasian  
Document Computing Symposium

10 December 2007

Edited by  
Amanda Spink, Andrew Turpin and Mingfang Wu



# **Proceedings of the Twelfth Australasian Document Computing Symposium**

The Melbourne Zoo,  
10 December 2007

Published by  
School of Computer Science and Information Technology,  
RMIT University,  
Melbourne VIC 3001, Australia.

## Editors

Amanda Spink  
Andrew Turpin  
Mingfang Wu

ISBN: 978 0 646 48437 2

<http://www.cs.rmit.edu.au/~aht/adcs2007>

**Proceedings of the Twelfth Australasian Document Computing Symposium**  
Melbourne  
10 December 2007

## **Chair's Preface**

These proceedings contain the papers of the Twelfth Australasian Document Computing Symposium hosted by the School of Computer Science and Information Technology at RMIT University, and held at the Melbourne Zoo.

The two keynote talks, ten papers and nine posters reflect the breadth of interest of the Australian research community in the area of document computing. It is also a highlight of ADCS this year that we are not only collocated with The Australasian Language Technology Workshop 2007, but are sharing a paper session, keynote talk, and social functions with the Australian natural language research community.

Of the 18 full papers submitted, 10 were accepted for presentation as full papers (55%), 5 were accepted as posters (28%), and 3 were rejected (17%). Of the 6 short papers submitted, 4 were accepted for presentation as posters (67%).

All full papers received at least three anonymous reviews by experts in the area, and every short paper received at least two anonymous reviews by experts in the area. Dual submissions were explicitly prohibited.

The members of the program committee and extra reviewers deserve special thanks for their professional reviews all received in the short time required for this ADCS conference. Reviewers not listed among the program committee include: Peter Bailey, Shlomo Geva, Alexander Krumholz, Jose Lay, Ben Martin, Johvan Pehcevski, and Tom Rowlands.

We would also like to thank RMIT School of Computer Science and IT, and NICTA (Victoria) for their generous sponsorship of the event.

The symposium includes many formal presentations, but perhaps its greatest benefit lies in the opportunity it provides for document computing practitioners and researchers to get together and informally share ideas and enthusiasm.

## **Symposium Co-Chairs**

Andrew Turpin RMIT University  
Mingfang Wu RMIT University

## **Program co-chairs**

Amanda Spink Queensland University of Technology  
Andrew Turpin RMIT University

## **Program Committee**

Peter Bruza Queensland University of Technology  
Bob Colomb University of Technology Malaysia  
Stijn Dekeyser University of Southern Queensland  
Peter Eklund University of Wollongong  
David Hawking CSIRO, Canberra  
Rob McArthur CSIRO, Canberra  
Alistair Moffat The University of Melbourne  
Gitesh K. Raikundalia Victoria University  
Falk Scholer RMIT University  
Saied Tahagohghi RMIT University  
Jamie Thom RMIT University  
Anne-Marie Vercoustre INRIA, France  
Anh Vo The University of Melbourne  
William Webber The University of Melbourne  
Ross Wilkinson CSIRO, Canberra  
Mingfang Wu RMIT University  
Justin Zobel NICTA

## **ADCS Advisory Committee**

Peter Bruza Queensland University of Technology  
Judy Kay The University of Sydney  
David Hawking CSIRO, Canberra  
Alistair Moffat The University of Melbourne  
Amanda Spink Queensland University of Technology  
Ross Wilkinson CSIRO, Canberra  
Justin Zobel NICTA

## Contents

### Keynote I

*How to Evaluate Information Retrieval: Why is it Receiving Attention Now?* vii  
Dr Tetsuya Sakai (Newswatch, Japan)

### Keynote II (Joint with ALTW2007)

*Measures of Measurements: Robust Evaluation of Search Systems* viii  
Professor Justin Zobel (NICTA, Victoria)

### Papers

*Score Standardization for Robust Comparison of Retrieval Systems* 1  
William Webber and Alistair Moffat (The University of Melbourne)  
Justin Zobel (NICTA, Victoria)

*IR Evaluation Using Multiple Assessors Per Topic* 9  
Andrew Trotman and Dylan Jenkinson (University of Otago)

*On the Distribution of User Persistence for Rank-Biased Precision* 17  
Laurence Park and Yuye Zhang (University of Melbourne and NICTA Victoria)

*Hybrid Bitvector Index Compression* 25  
Alistair Moffat and Shane Culpepper (The University of Melbourne)

*Source Code Authorship Attribution with n-Grams* 32  
Steven Burrows and Saied Tahaghoghi (RMIT University)

*Search and Navigation in Structured Document Retrieval: Comparison of User Behaviour in Search on Document Passages and XML Elements* 40  
Gabriella Kazai (Microsoft Research, Cambridge)

*Can Requests-For-Action and Commitments-To-Act Be Reliably Identified in Email Messages* 48  
Andrew Lampert and Cecile Paris (CSIRO, North Ryde)  
Robert Dale (Macquarie University)

*Use of Wikipedia Categories in Entity Ranking* 56  
James A. Thom (RMIT University) Jovan Pehcevski and  
Anne-Marie Vercoustre (INRIA, France)

*A Bottom-Up Term Extraction Approach for Web-Based Translation in Chinese-English IR Systems* 64  
Chengye Lu, Yue Xu and Shlomo Geva (Queensland University of Technology)

*Automatic Thread Classification for Linux User Forum Information Access* 72  
Timothy Baldwin, David Martinez and Richard B. Penman (University of Melbourne)

## Posters

<i>Multimedia Web Searching on a Meta-Search Engine</i> Dian Tjondronegoro (Queensland University of Technology) Amanda Spink (Queensland University of Technology) Bernard Jansen (The Pennsylvania State University, PA)	80
<i>Querying Image Ontology</i> Dayang Awang Iskandar (RMIT University) James Thom (RMIT University) Saied Tahaghoghi (RMIT University)	84
<i>Does Brandname Influence Perceived Search Result Quality? Yahoo! Google, and WebKumara</i> Peter Bailey (CSIRO, Canberra) Paul Thomas (The Australian National University) David Hawking (CSIRO, Canberra)	88
<i>Efficient Neighbourhood Estimation for Recommenders with Large Datasets</i> Li-Tung Weng (Queensland University of Technology) Yue Xu (Queensland University of Technology) Yuefeng Li (Queensland University of Technology) Richi Nayak (Queensland University of Technology)	92
<i>Document Composition and Content Selection Evaluation</i> Shijian Lu (CSIRO, North Ryde)	96
<i>A Comparison of Evaluation Measures Given How Users Perform on Search Tasks</i> James Thom (RMIT University) Falk Scholer (RMIT University)	100
<i>Integration of Information Filtering and Data Mining Process for Web Information Retrieval</i> Xujuan Zhou (Queensland University of Technology) Yuefeng Li (Queensland University of Technology) Peter Bruza (Queensland University of Technology) Yue Xu (Queensland University of Technology)	104
<i>A Framework for Measuring the Impact of Web Spam</i> Timothy Jones (The Australian National University) David Hawking (CSIRO, Canberra) Ramesh Sankaranarayana (The Australian National University)	108
<i>Predicting Query Performance for User-Based Search Tasks</i> Ying Zhao (RMIT University) Falk Scholer (RMIT University)	112

## **Keynote Talk I**

### *How to Evaluate Information Retrieval: Why is it Receiving Attention Now?*

My talk is comprised of three parts. PART ONE: I will introduce a couple of novel information access services that we provide at FreshEye, the Japanese Web portal run by NewsWatch, Inc. NewsWatch was founded by Toshiba in 1996, and was bought by Yahoo! JAPAN in 2006. PART TWO: I will then discuss why information retrieval evaluation is receiving a lot of attention in the research community now, and mention some challenges, including the relevance data incompleteness issue, and the possibility of evaluating online, nontraditional information access services like the ones I have mentioned in PART ONE. PART THREE: I will describe the ongoing activities at NTCIR, which is an international information retrieval evaluation effort for Asian languages. The latest tasks cover complex question answering, cross-language information retrieval, opinion extraction and patent mining/translation. I will conclude the talk by urging you to submit a paper to EVIA 2008, the Second International Workshop on Evaluating Information Access, which will be held together with NTCIR-7 in Tokyo.

### *About Tetsuya Sakai*

Tetsuya Sakai is the Director of the Natural Language Processing Laboratory at NewsWatch, Inc. which runs the Japanese Web portal FreshEye. In 1993, He received his Master's degree from Waseda University and joined Toshiba Corporate R&D Center. He received his Ph.D from Waseda University in 2000 for his work on information retrieval and filtering systems. From 2000 to 2001, he was a visiting researcher at the University of Cambridge Computer Laboratory. In February 2007, he left Toshiba to join NewsWatch, Inc. He has received several awards from the Information Processing Society of Japan (IPSJ) and the Institute of Electronics, Information and Communication Engineers (IEICE). He is currently the Asian Regional Representative of ACM SIGIR, and is on the steering committee of Asia Information Retrieval Symposium (AIRS) and the editorial board of the international journal "Information Retrieval".

## **Keynote Talk II**

### *Measures of Measurements: Robust Evaluation of Search Systems*

A good search system is one that helps a user to find useful documents. When building a new system, we hope, or hypothesise, that it will be more effective than existing alternatives. We apply a measure, which is often a drastic simplification, to establish whether the system is effective. Thus the ability of the system to help users and the measurement of this ability are only weakly connected, by assumptions that the researcher may not even be aware of. But how robust are these assumptions? If they are poor, is the research invalid? Such concerns apply not just to search, but to many other data-processing tasks. In this talk I introduce some of the recent developments in evaluation of search systems, and use these developments to examine some of the assumptions that underlie much of the research in this field.

### *About Justin Zobel*

Professor Justin Zobel is leading the Computing for Life Sciences initiative within National ICT Australia's Victorian Laboratory. He received his PhD from the University of Melbourne and for many years was based in the School of CS&IT at RMIT University, where he led the Search Engine group. He is an Editor-in-Chief of the International Journal of Information Retrieval, an associate editor of ACM Transactions on Information Systems and of Information Processing & Management, and was until recently Treasurer of ACM SIGIR. In the research community, he is best known for his role in the development of algorithms for efficient text retrieval. He is the author of "Writing for Computer Science" and his interests include search, bioinformatics, fundamental data structures, and research methods.