CONFERENCE REPORT

The Seventeenth Australasian Document Computing Symposium

Andrew Trotman
University of Otago,
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Sally Jo Cunningham
University of Waikato
Hamilton, New Zealand
sallyjo@waikato.ac.nz

Laurianne Sitbon
Queensland University of Technology
Brisbane, Australia
laurianne.sitbon@qut.edu.au

Abstract

The Seventeenth Australian Document Computing Symposium was held in Dunedin, New Zealand on the 5th and 6th of December 2012. In total twenty four papers were submitted. From those eleven were accepted for full presentation and 8 for short presentation. A poster session was held jointly with the Australasian Language Technology Workshop.

1 Introduction

The Australasian Document Computing Symposium was first run in 1996 and has been the annual regional Information Retrieval and Document Computing conference in Australasia ever since. In 2012 it was run in Dunedin (New Zealand), the first time the event has been run outside of Australia.

In total 24 papers were submitted, and similar numbers have been seen for several years. Each paper was, in full, single-blind reviewed by three independent qualified experts in the field. Reviewing was at the normal high level expected at SIGIR supported academic events. From the 24 papers, 11 were accepted for full presentation and 8 were selected for short presentation and poster (poster boaster and poster). The PC chairs are grateful to the reviewers for the quality of their reviews.

Two invited talks were given, one by Charles L. A. Clarke (University of Waterloo) and the other by Nigel Stanger (University of Otago).

Sponsorship was provided by: ACM SIGIR, Google, NICTA, Bing, the University of Otago, and Funnelback. A small proportion of this was used for logistics (rent of poster boards, coffee, and lunches), but the vast majority was used for student travel with every applicant receiving high levels of financial support. Funnelback directly sponsored the best paper award, which was won by Petri & Culpepper for their paper "Efficient indexing algorithms for approximate pattern matching in text" [11]. The ADCS steering committee would like to thank these organizations for their contributions.

The social program consisted of a pay-your-own-way "banquet" at a near-by pizza house, and an optional pay-for-yourself boat trip down the spectacularly scenic Otago Harbor (in the rain). Most of the 34 registrants chose to attend both.

The remainder of this report gives an overview of the proceedings of ADCS 2012. The interested reader is encouraged to consult the full papers which are held in the ACM digital library; and to participate in future years.

ADCS 2012 was an ACM SIGIR in-cooperation event.

2 Keynote 1

ADCS 2012 started with an invited talk by Charles L. A. Clarke (University of Waterloo). Charlie presented his recent work on Time-Biased Gain. In this work he is trying to progress Information Retrieval metrics to include some aspect of user behavior. In particular he includes a user model in the metric. This model can include such characteristics as the user reading behavior – a form of length normalization, and so on. By modeling the gain in information and the user behavior necessary to get it he is better able to score search engine results than is possible with models with implicit (or no) user model such as precision or Mean Average Precision.

He has published aspects of this work in prestigious international conferences including SIGIR [16], CIKM [15], and HCIR [14]. In his ADCS presentation he brought this work together and discussed with participants how they might further explore this topic.

3 Session 1

The first paper session saw three full papers on search engines.

Crane & Trotman [2] discussed their experiments in indexing Clueweb09 category B and category A using the ATIRE search engine. In particular they concentrated on indexing efficiency and search precision when using the Waterloo spam list, stop words, and static pruning algorithms. They found that Category A can be indexed on a single server in less than a day, and when searching with BM25 such efficiency decisions do not statistically significantly affect precision.

Petri & Culpepper [11] won the Funnelback best paper award for their work on building indexes capable of supporting approximate search. They use a variant on the Context Bound Burrows-Wheeler Transform that they call the Variable Depth Burrows-Wheeler Transform. The efficiency of their algorithm was compared to that of q-gram approaches and they show improvements.

Hawking & Jones [4] re-examine the question of document reordering for search engines. Their particular interest lies in search engine efficiency rather than the more traditional compression reasons for reordering. They first index the collection then in a post process re-order documents to build a new index out of the old index that is ordered optimally for early document-at-a-time stopping. Their re-orderer is fast and their resultant indexes can be searched with little-to-no loss in precision in less time than the original un-ordered index.

4 Session 2

The second paper session saw three full papers on users.

Kim, Thomas, Sankaranarayana, & Gedeon [6] are interested in mobile searching and conducted a study to examine user behavior on a desktop computer and to compare it with that on a mobile device. In their experiments they looked at fixation time and scanning strategy. Their results suggest that users look-ahead less on a mobile device than on a large screen device, but they still take longer to search. They suggest further HCI work is needed for IR on small screens.

Thomas [20] conducted a study into how users navigate web sites and shows that user difficulty can be modeled from web logs. He then takes this model and applies it to live web sites to identify hot-spots of user difficulty. A web tool he built can be used by webmasters to examine how changes they make affect levels of difficulty.

Kinley, Tjondronegoro, Partridge, & Edwards [7] conducted a user study with 50 users to examine the effects of query type on query reformation behavior. Specifically, they looked at addition, removal, and replacement of search terms in a query, along with repeated queries and query

replacement. They examined exploratory, factual, and abstract search. They found that query type does affect the way queries are reformulated

5 Session 3

The third session saw two full papers, one on metrics and the other on sentence retrieval.

Moffat, Scholer & Thomas [10] examined information retrieval metrics and plausible user models for those metrics. They argue that for some of the IR metrics there is no plausible model. They suggest that users are more likely to end their session as they find more relevant material, but less slowly if it took longer to find that information in the first place. They propose an experiment to test this and are expecting results in the near future.

Bando, Scholer & Turpin [1] show that sentence length bias exists in the TREC novelty track assessments. This bias suggests that prior results of the Novelty track may not be conclusive. Of particular interest is that the inclusion of a sentence length component in the ranking algorithm, and using feedback, appears to result in substantial improvements in precision. They go on to propose an alternative evaluation strategy that appears to be unbiased.

6 Minutes of the ADCS 2012 Business Meeting

A joint discussion with ALTA resulted in the decision to hold ADCS 2013 and ALTA 2013 together (again) in Brisbane in December 2013. The general chair will be Laurianne Sitbon and one of the PC chairs will be Guido Zuccon. The conference will again be two days. The parallel session approach with ADCS and ALTA sharing tea breaks and lunch breaks was considered successful and will be continued.

The ADCS / SIGIR relationship was considered beneficial and continuing to run in-cooperation with ACM SIGIR was strongly favored. There was agreement that SIGIR student travel allowances were of substantial benefit (thanks go to SIGIR for their support).

A discussion on the benefits of the ACM Digital Library showed strong support from young academics and students. More established academics raised the issue of copyright transfer to the ACM (which is new for ADCS) and the formal lodging of papers in the digital library making ADCS a conference rather than a symposium and that this might subtly change the nature of the event. General discussion suggested this might be for the better. Agreement was raised to discuss copyright with the ACM.

7 Keynote 2

In the second keynote Nigel Stanger (University of Otago) discussed his unpublished work in building small digital museums for local communities. His work with the Cardrona Valley raised awareness of the gulf between the understanding of a digital museum by an academic and the general public in a remote community. Community members were, for example, concerned that giving a photo to an online museum would result in them not retaining the photograph. They were also, rightly, concerned with copyright and privacy. Stanger's recent work has focused on digital preservation and magnetic media and he showed a device for reading the magnetic flux from the surface of a disk so that it could be decoded on a desktop computer. Such devices are being used to rescue disks from early Apple and Commodore computers.

8 ADCS / ALTA Joint Session

The joint ADCS / ALTA session saw two papers from ADCS and two from ALTA; these papers were selected by the chairs on the basis of being of interest to both communities. This report

discusses only the ADCS contributions to that session. The interested reader is referred to the proceedings of the Australasian Language Technology Associate Workshop for the ALTA contributions [12, 13].

Ghahremanloo, Thom, & Magee [3] used the METHONTOLOGY approach to design an ontology to represent the knowledge about sustainability indicator sets. They presented two design models: generic and specific. The ontology models were evaluated by the ROMEO approach using seen and unseen indicator sets. They concluded that where an ontology needs to be designed for both seen and unseen indicator systems, a generic and reusable design is preferable.

Koopman, Zuccon, Bruza, Sitbon, & Lawley [8] present a retrieval model for medical IR that utilizes a graph-based document representation as an alternative to bag-of-words; the graph is made up of concepts from the SNOMED CT medical ontology. Their evaluation using the TREC Medical Records track shows that utilizing this graph-based concept representation is more effective than using terms and / or a bag-of-words document representation.

9 Posters

The ADCS poster boaster session saw the authors of 8 short contributions presenting their work in a 6 minute each time slot. This work was on a broad range of document computing and information retrieval topics including: language identification [9]; analysis of parliamentary question time [21]; the release of a new corpus for cross-lingual information retrieval [19]; concept based information retrieval using subsumption relations [23]; the identification of entity-related attribute-value pairs in documents [5]; term dependencies in query expansion [18]; similarity of document signatures [22]; and public health information dissemination [17].

This poster boaster session was followed by a joint poster session with ALTA in which 16 posters were seen.

10 References

[1] Bando, L.L., F. Scholer, A. Turpin, *Sentence length bias in TREC novelty track judgements*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 55-61.

[2] Crane, M., A. Trotman, *Effects of spam removal on search engine efficiency and effectiveness*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 1-8.

[3] Ghahremanloo, L., J.A. Thom, L. Magee, *An ontology derived from heterogeneous sustainability indicator set documents*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 72-79.

[4] Hawking, D., T. Jones, *Reordering an index to speed query processing without loss of effectiveness*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 17-24.

[5] Hou, J., R. Nayak, J. Zhang, *Finding additional semantic entity information for search engines*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 115-122.

[6] Kim, J., P. Thomas, R. Sankaranarayana, T. Gedeon, *Comparing scanning behaviour in web search on small and large screens*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 25-30.

[7] Kinley, K., D. Tjondronegoro, H. Partridge, S. Edwards, *Relationship between the nature of the search task types and query reformulation behaviour*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 39-46.

[8] Koopman, B., G. Zuccon, P. Bruza, L. Sitbon, M. Lawley, *Graph-based concept weighting for*

medical information retrieval, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 80-87.

[9] Milne, R.M., R.A. O'Keefe, A. Trotman, *A study in language identification*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 88-95.

[10] Moffat, A., F. Scholer, P. Thomas, *Models and metrics: IR evaluation as a user process*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 47-54.

[11] Petri, M., J.S. Culpepper, *Efficient indexing algorithms for approximate pattern matching in text*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 9-16.

[12] Sarker, A., D. Molla, C. Paris, *Towards Two-step Multi-document Summarisation for Evidence Based Medicine: A Quantitative Analysis* in *Proceedings of the Australasian Language Technology Association Workshop 2012*. 2012. p. 79-87.

[13] Smith, A.G., C.X.S. Zee, A.L. Uitdenbogerd, *In Your Eyes: Identifying Clichés in Song Lyrics*, in *Proceedings of the Australasian Language Technology Association Workshop 2012*. 2012. p. 88-96.

[14] Smucker, M.D., C.L.A. Clarke, *Modeling user variance in time-biased gain*, in *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*. 2012.

[15] Smucker, M.D., C.L.A. Clarke, *Stochastic simulation of time-biased gain*, in *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012. p. 2040-2044.

[16] Smucker, M.D., C.L.A. Clarke, *Time-based calibration of effectiveness measures*, in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 2012. p. 95-104.

[17] Steele, R., D. Dumbrell, *Putting the public into public health information dissemination: social media and health-related web pages*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 135-138.

[18] Symonds, M., P. Bruza, G. Zuccon, L. Sitbon, I. Turner, *Is the unigram relevance model term independent?: classifying term dependencies in query expansion*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 123-127.

[19] Tang, L.-X., S. Geva, A. Trotman, *An English-translated parallel corpus for the CJK Wikipedia collections*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 104-110.

[20] Thomas, P., *Explaining difficulty navigating a website using page view data*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 31-28.

[21] Turpin, A., *An attempt to measure the quality of questions in question time of the Australian Federal Parliament*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 96-103.

[22] Vries, C.M.D., S. Geva, *Pairwise similarity of TopSig document signatures*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 128-134.

[23] Zuccon, G., B. Koopman, A. Nguyen, D. Vickers, L. Butt, *Exploiting medical hierarchies for concept-based information retrieval*, in *Proceedings of the Seventeenth Australasian Document Computing Symposium*. 2012. p. 111-114.