ACL-IJCNLP 2009

NLPIR4DL 2009

2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries

Proceedings of the Workshop

7 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-58-9 / 1-932432-58-2

Introduction

In recent years, interest in scholarly publications in electronic forms has boomed, and several large-scale electronic digital libraries and citation indices are now used everyday by researchers.

The fact that formal citation metrics have become an increasingly large factor in decision-making by universities and funding bodies worldwide makes the need for research in such topics and for better methods for measuring the impact of work more pressing.

Current digital libraries collect and allow access to digital papers and their metadata (including citations), but largely do not attempt to analyze the items they collect.

The goal of this workshop is to investigate how developments in natural language processing and information retrieval techniques can advance the state-of-the-art in scholarly document understanding, analysis and retrieval.

We were amazed by the number of high-quality papers we received to this inaugural workshop, and by the innovativeness of the research that is done in this area. The contributions split into various areas, and we will here give a quick overview of what these are.

Full document text analysis can help design information access, namely automatic summarization and sentiment detection methods, automated recommendation and reviewing systems, and may provide data for visualizing scientific trends and bibliometrics. *Kaplan et al.*'s paper studies the interaction of citation contexts and co-reference for scientific summarization. Discourse analysis also is the focus of *Merity et al*'s paper which presents an ME-based approach to Argumentative Zoning, and of *Sándor and Vorndran*'s reviewing support system.

We are particularly proud to have two user studies on navigation and search at this workshop, because better systems for information access require such studies as a starting point. *Hearst and Stoica* present a user study and prototype system for faceted navigation in scholarly digital libraries, whereas the study by *Wan et al.* is collecting the browsing-specific information needs by medical searchers, in particular those that could be satisfied by citations and their contexts.

As far as improvement of academic search itself is concerned, the topic of *Shi et al.*'s paper is improved anchor text extraction.

Citation analysis takes this a step further, adding scientific social network analysis as another strand of evidence to enhance solutions to the above challenges. Web based digital libraries add download counts and Web 2.0 information such as tagging.

This workshop contains three papers on citation support in the strictest sense, namely *Hong et al.*, with a fast and lightweight reference string extractor, and *Romanello et al.*, with a recogniser for canonical references, and also a paper on the extraction of researcher affiliation, namely *Nagy et al.*

Aside from researchers, this workshop hopes to interest other stakeholders, namely implementers, publishers and policymakers. For instance, *Nanba and Takezawa*'s research into the Patent Classification Support goes in this direction. Even within computer science, many different scholarly sites exist – ACM Portal, IEEE Xplore, Google Scholar, PSU's CiteSeerX, MSRA's Libra, Tsinghua's

ArnetMiner, Trier's DBLP, UMass' Rexa, Hiroshima's PRESRI – and with this workshop we hope to bring a number of these contributers together. *Radev et al.*'s work on the ACL Anthology Network Corpus reports one such invaluable resource.

Today's publishers continue to seek new ways to be relevant to their consumers, in disseminating the right published works to their audience. Dr. Rick Lee, who is the Director for MIS and Electronic Publishing at the World Scientific Publishing Company, is our invited speaker and will talk about his company's strategy to serve content to the user in future-proof ways.

All that is left after this brief overview of the work in this workshop is to wish all participants a good and informative day.

The organisers of the first NLPIR4DL workshop,

Simone Teufel
Min-Yen Kan

Organizers:

Min-Yen Kan, National University of Singapore
Simone Teufel, University of Cambridge

Program Committee:

Colin Batchelor, Royal Society of Chemistry
Steven Bird, University of Melbourne and the Linguistic Data Consortium
Shannon Bradshaw, Drew University
Jason S Chang, National Tsing-hua University
Robert Dale, Macquarie University
Bonnie Dorr, University of Maryland
Curtis Dyreson, Utah State University
David Ellis, Facebook
C. Lee Giles, Pennsylvania State University
Dan Jurafsky, Stanford University
Noriko Kando, National Institute of Informatics, Japan
Dongwon Lee, Pennsylvania State University
Elizabeth Liddy, Syracuse University
Andrew McCallum, University of Massachusetts
Qiaozhu Mei, University of Illinois at Urbana-Champaign
Hidetsugu Nanba, Hiroshima University
Manabu Okumura, Tokyo Institute of Technology
Dragomir Radev, University of Michigan
Anna Ritchie, University of Cambridge
Mark Sanderson, University of Sheffield
John Swales, University of Michigan
Jie Tang, Tsinghua University
Michael Thelwall, University of Wolverhampton
Bonnie Webber, University of Edinburgh
Howard White, Drexel University

Additional Reviewers:

Johannes Schanda, Sheffield of University

Invited Speaker:

Rick Lee, World Scientific Publishing Company

Table of Contents

Conference Program

Friday, August 7, 2009

9:00	Opening Remarks
9:10–10:00	Invited Talk by Rick Lee of the World Scientific Publishing Company
10:00–10:30	Coffee Break
	Session 1: Metadata and Content
10:30–10:55	*Researcher affiliation extraction from homepages* István Nagy, Richárd Farkas and Márk Jelasity
10:55–11:20	*Anchor Text Extraction for Academic Search* Shuming Shi, Fei Xing, Mingjie Zhu, Zaiqing Nie and Ji-Rong Wen
11:20–11:45	*Accurate Argumentative Zoning with Maximum Entropy models* Stephen Merity, Tara Murphy and James R. Curran
11:45–12:10	*Classification of Research Papers into a Patent Classification System Using Two Translation Models* Hidetsugu Nanba and Toshiyuki Takezawa
12:10–13:50	Lunch Break
	Session 2: System Aspects
13:50–14:15	*Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences* Ágnes Sándor and Angela Vorndran
14:15–14:40	*Designing a Citation-Sensitive Research Tool: An Initial Study of Browsing-Specific Information Needs* Stephen Wan, Cécile Paris, Michael Muthukrishna and Robert Dale
14:40–15:05	*The ACL Anthology Network* Dragomir R. Radev, Pradeep Muthukrishnan and Vahed Qazvinian
15:05–15:30	*NLP Support for Faceted Navigation in Scholarly Collection* Marti A. Hearst and Emilia Stoica

Friday, August 7, 2009 (continued)

15:30–16:00 Coffee Break

Session 3: Citation Support

16:00–16:25 *FireCite: Lightweight real-time reference string extraction from webpages*
Ching Hoi Andy Hong, Jesse Prabawa Gozali and Min-Yen Kan

16:25–16:50 *Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields*
Matteo Romanello, Federico Boschetti and Gregory Crane

16:50–17:15 *Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach*
Dain Kaplan, Ryu Iida and Takenobu Tokunaga

17:15–18:00 Informal Demonstration Sessions - Wrap up