

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

New York University, NY, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Sung Hyon Myaeng Ming Zhou
Kam-Fai Wong Hong-Jiang Zhang (Eds.)

Information Retrieval Technology

Asia Information Retrieval Symposium, AIRS 2004
Beijing, China, October 18-20, 2004
Revised Selected Papers

Volume Editors

Sung Hyon Myaeng
Information and Communications University (ICU)
119 Munji-Ro, Yuseong-Gu, Daejeon, 305-714, South Korea
E-mail: myaeng@icu.ac.kr

Ming Zhou
Hong-Jiang Zhang
Microsoft Research Asia
5F, Beijing Sigma Center
No. 49 Zhichun Road Haidian District, Beijing 100080, China
E-mail: {mingzhou,hjzhang}@microsoft.com

Kam-Fai Wong
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
E-mail: kfwong@se.cuhk.edu.hk

Library of Congress Control Number: 2005921104

CR Subject Classification (1998): H.3, H.4, F.2.2, E.1, E.2

ISSN 0302-9743

ISBN 3-540-25065-4 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11398479 06/3142 5 4 3 2 1 0

Preface

The Asia Information Retrieval Symposium (AIRS) was established by the Asian information retrieval community after the successful series of Information Retrieval with Asian Languages (IRAL) workshops held in six different locations in Asia, starting from 1996. While the IRAL workshops had their focus on information retrieval problems involving Asian languages, AIRS covers a wider scope of applications, systems, technologies and theory aspects of information retrieval in text, audio, image, video and multimedia data. This extension of the scope reflects and fosters increasing research activities in information retrieval in this region and the growing need for collaborations across subdisciplines.

We are very pleased to report that we saw a sharp increase in the number of submissions and their quality, compared to the IRAL workshops. We received 106 papers from nine countries in Asia and North America, from which 28 papers (26%) were presented in oral sessions and 38 papers in poster sessions (36%). It was a great challenge for the Program Committee to select the best among the excellent papers. The low acceptance rates witness the success of this year's conference.

After a long discussion between the AIRS 2004 Steering Committee and Springer, the publisher agreed to publish our proceedings in the Lecture Notes in Computer Science (LNCS) series, which is SCI-indexed. We feel that this strongly attests to the excellent quality of the papers.

The attendees were cordially invited to participate in and take advantage of all the technical programs at this conference. A tutorial was given on the first day to introduce the state of the art in Web mining, an important application of Web document retrieval. Two keynote speeches covered two main areas of the conference: video retrieval and language issues. There were a total of eight oral sessions run, with two in parallel at a time, and two poster/demo sessions.

The technical and social programs, which we are proud of, were made possible by the hard-working people behind the scenes. In addition to the Program Committee members, we are thankful to the Organizing Committee (Shao-Ping Ma and Jianfeng Gao, Co-chairs), Interactive Posters/Demo Chair (Gary G. Lee), and the Special Session and Tutorials Chair (Wei-Ying Ma). We also thank the sponsoring organizations: Microsoft Research Asia, the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong, and LexisNexis for their financial support, the Department of Computer Science and Technology, Tsinghua University for local arrangements, the Chinese NewsML Community for website design and administration, Ling Huang for the logistics, Weiwei Sun for the conference webpage management, EONSO-LUTION for the conference management, and Springer for the postconference

LNCS publication. We believe that this conference set a very high standard for a regionally oriented conference, especially in Asia, and we hope that it continues as a tradition in the upcoming years.

Sung Hyon Myaeng and Ming Zhou (PC Co-chairs)
Kam-Fai Wong and Hong-Jiang Zhang (Conference Co-chairs)

Organization

General Conference Co-chairs

Kam-Fai Wong, Chinese University of Hong Kong, China
Hong-Jiang Zhang, Microsoft Research Asia, China

Program Co-chairs

Sung Hyon Myaeng, Information and Communications University (ICU),
South Korea
Ming Zhou, Microsoft Research Asia, China

Organization Committee Co-chairs

Jianfeng Gao, Microsoft Research Asia, China
Shao-Ping Ma, Tsinghua University, China

Special Session and Tutorials Chair

Wei-Ying Ma, Microsoft Research Asia, China

Interactive Posters/Demo Chair

Gary Geunbae Lee, POSTECH, South Korea

Steering Committee

Jun Adachi, National Institute of Informatics, Japan
Hsin-Hsi Chen, National Taiwan University, Taiwan
Lee-Feng Chien, Academia Sinica, Taiwan
Tetsuya Ishikawa, University of Tsukuba, Japan
Gary Geunbae Lee, POSTECH, South Korea
Mun-Kew Leong, Institute for Infocomm Research, Singapore
Helen Meng, Chinese University of Hong Kong, China
Sung-Hyon Myaeng, Information and Communications University, South Korea

Hwee Tou Ng, National University of Singapore, Singapore
Kam-Fai Wong, Chinese University of Hong Kong, China

Organizing Committee

Lee-Feng Chien, Academia Sinica, Taiwan (Publicity, Asia)
Susan Dumais, Microsoft, USA (Publicity, North America)
Jianfeng Gao, Microsoft Research Asia, China (Publication, Co-chair)
Mun-Kew Leong, Institute for Infocomm Research, Singapore (Finance)
Shao-Ping Ma, Tsinghua University, China (Local Organization, Co-chair)
Ricardo Baeza-Yates, University of Chile, Chile (Publicity, South America)
Shucaï Shi, BITI, China (Local Organization)
Dawei Song, DSTC, Australia (Publicity, Australia)
Ulich Thiel, IPSI, Germany (Publicity, Europe)
Chuanfa Yuan, Tsinghua University, China (Local Organization)

Program Committee

Peter Anick, Yahoo, USA
Hsin-Hsi Chen, National Taiwan University, Taiwan
Aitao Chen, University of California, Berkeley, USA
Lee-Feng Chien, Academia Sinica, Taiwan
Fabio Crestani, University of Strathclyde, UK
Edward A. Fox, Virginia Tech, USA
Jianfeng Gao, Microsoft Research Asia, China
Hani Abu-Salem, DePaul University, USA
Tetsuya Ishikawa, University of Tsukuba, Japan
Christopher Khoo, Nanyang Technological University, Singapore
Jung Hee Kim, North Carolina A&T University, USA
Minkoo Kim, Ajou University, South Korea
Munchurl Kim, Information and Communication University, South Korea
Kazuaki Kishida, Surugadai University, Japan
Kui-Lam Kwok, Queens College, City University of New York, USA
Wai Lam, Chinese University of Hong Kong, China
Gary Geunbae Lee, POSTECH, South Korea
Mun-Kew Leong, Institute for Infocomm Research, Singapore
Gena-Anne Levow, University of Chicago, USA
Hang Li, Microsoft Research Asia, China
Robert Luk, Hong Kong Polytechnic University, China
Gay Marchionini, University of North Carolina, Chapel Hill, USA
Helen Meng, Chinese University of Hong Kong, China
Hiroshi Nakagawa, University of Tokyo, Japan
Hwee Tou Ng, National University of Singapore, Singapore
Jian-Yun Nie, University of Montreal, Canada
Jon Patrick, University of Sydney, Australia
Ricardo Baeza-Yates, University of Chile, Chile
Hae-Chang Rim, Korea University, South Korea
Tetsuya Sakai, Toshiba Corporate R&D Center, Japan
Padmini Srinivasan, University of Iowa, USA
Tomek Strzalkowski, State University of New York, Albany, USA
Maosong Sun, Tsinghua University, China
Ulrich Thiel, Fraunhofer IPSI, Germany
Takenobu Tokunaga, Tokyo Institute of Technology, Japan
Hsin-Min Wang, Academia Sinica, Taiwan
Ross Wilkinson, CSIRO, Australia
Lide Wu, Fudan University, China
Jinxi Xu, BBN Technologies, USA
ChengXiang Zhai, University of Illinois, Urbana Champaign, USA
Min Zhang, Tsinghua University, China

Reviewers

Peter Anick
Yunbo Cao
Yee Seng Chan
Hsin-Hsi Chen
Zheng Chen
Aitao Chen
Tee Kiah Chia
Lee-Feng Chien
Fabio Crestani
Edward A. Fox
Jianfeng Gao
Hani Abu-Salem
Xuanjing Huang
Tetsuya Ishikawa
Christopher Khoo
Jung Hee Kim
Minkoo Kim
Munchurl Kim
Kazuaki Kishida

Kui-Lam Kwok
Wai Lam
Gary Geunbae Lee
Mun-Kew Leong
Gena-Anne Levow
Hang Li
Mu Li
Hongqiao Li
Chin-Yew Lin
Robert Luk
Wei-Ying Ma
Gay Marchionini
Helen Meng
Sung Hyon Myaeng
Hiroshi Nakagawa
Hwee Tou Ng
Jian-Yun Nie
Jon Patrick
Ricardo Baeza-Yates

Hae-Chang Rim
Tetsuya Sakai
Tomek Strzalkowski
Maosong Sun
Ulrich Thiel
Takenobu Tokunaga
Hsin-Min Wang
Haifeng Wang
Ross Wilkinson
Kam-Fai Wong
Lide Wu
Jinxi Xu
Peng Yu
Chunfa Yuan
ChengXiang Zhai
Hong-Jiang Zhang
Min Zhang
Ming Zhou
Jian-lai Zhou

Table of Contents

Information Organization

| | |
|---|----|
| Automatic Word Clustering for Text Categorization Using Global Information <i>Wenliang Chen, Xingzhi Chang, Huizhen Wang, Jingbo Zhu, Tianshun Yao</i> | 1 |
| Text Classification Using Web Corpora and EM Algorithms <i>Chen-Ming Hung, Lee-Feng Chien</i> | 12 |
| Applying CLIR Techniques to Event Tracking <i>Nianli Ma, Yiming Yang, Monica Rogati</i> | 24 |
| Document Clustering Using Linear Partitioning Hyperplanes and Reallocation <i>Canasai Kruengkrai, Virach Sornlertlamvanich, Hitoshi Isahara</i> | 36 |

Automatic Summarization

| | |
|--|----|
| Summary Generation Centered on Important Words <i>Dongli Han, Takashi Noguchi, Tomokazu Yago, Minoru Harada</i> | 48 |
| Sentence Compression Learned by News Headline for Displaying in Small Device <i>Kong Joo Lee, Jae-Hoon Kim</i> | 61 |
| Automatic Text Summarization Using Two-Step Sentence Extraction <i>Wooncheol Jung, Youngjoong Ko, Jungyun Seo</i> | 71 |
| Sentence Extraction Using Time Features in Multi-document Summarization <i>Jung-Min Lim, In-Su Kang, Jae-Hak J. Bae, Jong-Hyeok Lee</i> | 82 |

Alignment/Paraphrasing in IR

| | |
|--|----|
| Extracting Paraphrases of Japanese Action Word of Sentence Ending Part from Web and Mobile News Articles <i>Hiroshi Nakagawa, Hidetaka Masuda</i> | 94 |
|--|----|

| | |
|--|-----|
| Improving Transliteration with Precise Alignment of Phoneme Chunks and Using Contextual Features <i>Wei Gao, Kam-Fai Wong, Wai Lam</i> | 106 |
|--|-----|

| | |
|---|-----|
| Combining Sentence Length with Location Information to Align Monolingual Parallel Texts <i>Weigang Li, Ting Liu, Sheng Li</i> | 118 |
|---|-----|

Web Search

| | |
|--|-----|
| Effective Topic Distillation with Key Resource Pre-selection <i>Yiqun Liu, Min Zhang, Shaoping Ma</i> | 129 |
|--|-----|

| | |
|---|-----|
| Efficient PageRank with Same Out-Link Groups <i>Yizhou Lu, Xuezheng Liu, Hua Li, Benyu Zhang, Wensi Xi, Zheng Chen, Shuicheng Yan, Wei-Ying Ma</i> | 141 |
|---|-----|

| | |
|---|-----|
| Literal-Matching-Biased Link Analysis <i>Yinghui Xu, Kyoji Umemura</i> | 153 |
|---|-----|

Linguistic Issues in IR

| | |
|--|-----|
| Multilingual Relevant Sentence Detection Using Reference Corpus <i>Ming-Hung Hsu, Ming-Feng Tsai, Hsin-Hsi Chen</i> | 165 |
|--|-----|

| | |
|--|-----|
| A Bootstrapping Approach for Geographic Named Entity Annotation <i>Seungwoo Lee, Gary Geunbae Lee</i> | 178 |
|--|-----|

| | |
|---|-----|
| Using Verb Dependency Matching in a Reading Comprehension System <i>Kui Xu, Helen Meng</i> | 190 |
|---|-----|

Document/Query Models

| | |
|--|-----|
| Sense Matrix Model and Discrete Cosine Transform <i>Bing Swen</i> | 202 |
|--|-----|

| | |
|---|-----|
| Query Model Estimations for Relevance Feedback in Language Modeling Approach <i>Seung-Hoon Na, In-Su Kang, Kyonghi Moon, Jong-Hyeok Lee</i> | 215 |
|---|-----|

| | |
|--|-----|
| A Measure Based on Optimal Matching in Graph Theory for Document Similarity <i>Xiaojun Wan, Yuzin Peng</i> | 227 |
|--|-----|

| | |
|---|-----|
| Estimation of Query Model from Parsimonious Translation Model <i>Seung-Hoon Na, In-Su Kang, Sin-Jae Kang, Jong-Hyeok Lee</i> | 239 |
|---|-----|

Enabling Technology

| | |
|---|-----|
| Ranking the NTCIR Systems Based on Multigrade Relevance <i>Tetsuya Sakai</i> | 251 |
|---|-----|

| | |
|--|-----|
| X-IOTA: An Open XML Framework for IR Experimentation <i>Jean-Pierre Chevallet</i> | 263 |
|--|-----|

| | |
|--|-----|
| Recognition-Based Digitalization of Korean Historical Archives <i>Min Soo Kim, Sungho Ryu, Kyu Tae Cho, Taik Heon Rhee, Hyun Il Choi, Jin Hyung Kim</i> | 281 |
|--|-----|

| | |
|--|-----|
| On Bit-Parallel Processing of Multi-byte Text <i>Heikki Hyvrö, Jun Takaba, Ayumi Shinohara, Masayuki Takeda</i> | 289 |
|--|-----|

Mobile Applications

| | |
|--|-----|
| Retrieving Regional Information from Web by Contents Localness and User Location <i>Qiang Ma, Katsumi Tanaka</i> | 301 |
|--|-----|

| | |
|---|-----|
| Towards Understanding the Functions of Web Element <i>Xinyi Yin, Wee Sun Lee</i> | 313 |
|---|-----|

| | |
|--|-----|
| Clustering-Based Navigation of Image Search Results on Mobile Devices <i>Hao Liu, Xing Xie, Xiaou Tang, Wei-Ying Ma</i> | 325 |
|--|-----|

| | |
|-------------------------------|-----|
| Author Index | 337 |
|-------------------------------|-----|