Sharon McDonald
John Tait (Eds.)

LNCS 2997

Advances in Information Retrieval

26th European Conference on IR Research, ECIR 2004
Sunderland, UK, April 2004
Proceedings

Springer

Sharon McDonald John Tait (Eds.)

Advances in Information Retrieval

26th European Conference on IR Research, ECIR 2004
Sunderland, UK, April 5-7, 2004
Proceedings

Springer

Volume Editors

Sharon McDonald
John Tait
University of Sunderland, School of Computing and Technology
David Goldman Informatics Centre, St. Peter's Campus
Sunderland SR6 0DD, UK
E-mail: {sharon.mcdonald,john.tait}@sunderland.ac.uk

Library of Congress Control Number: 2004102629

CR Subject Classification (1998): H.3, H.2, I.2.3, I.2.6, H.4, H.5.4, I.7

ISSN 0302-9743
ISBN 3-540-21382-1 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 10996706 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 2997

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Preface

These proceedings contain the refereed full technical papers presented at the 26th Annual European Conference on Information Retrieval (ECIR 2004). ECIR is the annual conference of the British Computer Society's specialist group in Information Retrieval. This year the conference was held at the School of Computing and Technology at the University of Sunderland. ECIR began life as the Annual Colloquium on Information Retrieval Research. The colloquium was held in the UK each year until 1998 when the event was held in Grenoble, France. Since then the conference venue has alternated between the United Kingdom and Continental Europe, and the event was renamed the European Conference on Information Retrieval. In recent years, ECIR has continued to grow and has become the major European forum for the discussion of research in the field of Information Retrieval. To mark this metamorphosis from a small informal colloquium to a major event in the IR research calendar, the BCS-IRSG decided to rename the event to the European Conference on Information Retrieval.

ECIR 2004 received 88 full paper submissions, from across Europe and further afield including North America, China and Australia, a testament to the growing popularity and reputation of the conference. Out of the 88 submitted papers, 28 were accepted for presentation. All papers were reviewed by at least three reviewers. Among the accepted papers 11 have a student as the primary author, illustrating that the traditional student focus of the original colloquium is alive today.

The collection of papers presented in this book reflect a broad range of IR problems. Contributions from keynote speakers Gary Marchionini and Yorick Wilks kick start the proceedings with Marchionini's proposal for a new paradigm for IR, based on his emphasis on the interactive nature of IR tasks, and Wilks' thought provoking discussion of the role of NLP techniques in IR. The organization of the proceedings reflects the session structure of the conference, topics covered include user interaction, question answering, information models, classification, summarization, image retrieval, evaluation issues, cross language IR and categorization, summarization, information models, question answering, cross language IR, image retrieval and Web-based and XML retrieval.

I am indebted to many individuals for the quality of this year's conference proceedings. Specifically, I would like to acknowledge the significant efforts of the programme committee, my co-chair John Tait and posters chair Michael Oakes. Thank you for your hard work, and for meeting the tight deadlines imposed. It has been my pleasure to work with you to produce a high-quality conference programme. Thanks also to the conference gold sponsors, Microsoft Research, Canon UK, Leighton Internet, BCS-IRSG, and the University of Sunderland.

Finally, I would like to extend my thanks to Arthur Wyvill and John Cartledge for their work on the paper submission system, Zia Syed for his help in publicizing ECIR 2004 and Lesley Jenkins for her excellent administrative support. Most of all, I would like to thank my husband Alan Lumsden for his love and support as well as the invaluable contribution he made at various stages in the development of ECIR 2004.

January 2004 Sharon McDonald

Organization

ECIR 2004 was organized by the School of Computing and Technology, University of Sunderland, United Kingdom.

Programme Committee

Sharon McDonald, University of Sunderland, United Kingdom (Chair)
John Tait, University of Sunderland, United Kingdom (Chair)
Michael Oakes, University of Sunderland, United Kingdom (Posters Chair)

Andrew MacFarlane, City University, United Kingdom
Alan Smeaton, Dublin City University, Ireland
Alessandro Sperduti, University of Padova, Italy
Ali Asghar Shiri, University of Strathclyde, United Kingdom
Andreas Rauber, Vienna University of Technology, Austria.
Ari Pirkola, University of Tampere, Finland
Arjen de Vries, CWI, Netherlands
Avi Arampatzis, University of Utrecht, Netherlands
Ayse Göker, Robert Gordon University, United Kingdom
Barry Smyth, University College Dublin, Ireland
Chris Mellish, University of Aberdeen, United Kingdom
Claudio Carpineto, Fondazione Ugo Bordoni, Italy
David Harper, Robert Gordon University, United Kingdom
David Losada, University de Santiago de Compostela, Spain
Djoerd Hiemstra, University of Twente, Netherlands
Dunja Mladenić, Jožef Stefan Institute, Slovenia
Fabio Crestani, University of Strathclyde, United Kingdom
Fabrizio Sebastiani, National Council of Research, Italy
Gabriella Pasi, National Council of Research, Italy
Gareth Jones, Dublin City University, Ireland
Giambattista Amati, Fondazione Ugo Bordoni, Italy
Giuseppe Amato, National Council of Research, Italy
Gloria Bordogna, CNR, IDPA, Italy
Iadh Ounis, University of Glasgow, United Kingdom
Ian Ruthven, University of Strathclyde, United Kingdom
Ion Androutsopoulos, Athens University of Economics and Business, Greece
Jane Reid, Queen Mary, University of London, United Kingdom
Jesper Wiborg Schneider, Royal School of Library and Information Science, Denmark
Joemon Jose, University of Glasgow, United Kingdom
Johannes Füernkrantz, Austrian Research Institute for Artificial Intelligence, Austria

Josiane Mothe, University Paul Sabatier, France
Jussi Karlgren, Swedish Institute of Computer Science, Sweden
Kees Koster, University of Nijmegen, Netherlands
Keith van Rijsbergen, University of Glasgow, United Kingdom
Leif Azzopardi, University of Paisley, United Kingdom
Marcello Federico, ITC-irst, Italy
Margaret Graham, Northumbria University, United Kingdom
Mark Girolami, University of Glasgow, United Kingdom
Marko Grobelnik, Jožef Stefan Institute, Slovenia
Massimo Melucci, University of Padova, Italy
Micheline Beaulieu, University of Sheffield, United Kingdom
Mohand Boughanem, University Paul Sabatier, France
Monica Landoni, University of Strathclyde, United Kingdom
Mounia Lalmas, Queen Mary, University of London, United Kingdom
Nicholas Kushmerick, University College Dublin, Ireland
Norbert Fuhr, University of Duisburg-Essen, Germany
Pasquale Savino, National Council of Research, Italy
Patrick Gallinari, University of Paris, France
Peter Ingwersen, Royal School of Library and Information Science, Denmark
Pia Borlund, Royal School of Library and Information Science, Denmark
Ricardo Baeza-Yates, Univeristy of Chile, Chile
Robert Gaizauskas, Univeristy of Sheffield, United Kingdom
Sándor Dominich, University of Veszprém, Hungary
Tony Rose, Cancer Research, United Kingdom
Umberto Straccia, National Council of Research, Italy
Wessel Kraaij, TNO TPD, Netherlands
Yoelle Maarek, IBM Research, Isreal
Yves Chiaramella, Joseph Fourier University, France

Best Student Paper Award Committee

Sándor Dominich, University of Veszprém, Hungary (Chair)
Giambattista Amati, Fondazione Ugo Bordoni, Italy
Pia Borlund, Royal School of Library and Information Science, Denmark

Additional Reviewers

Anastasios Tombros, Queen Mary, University of London, United Kingdom
Christopher Stokoe, University of Sunderland, United Kingdom
Gilles Hubert, IRIT, France
Janez Brank, Jožef Stefan Institute, Slovenia
Theodora Tsikrika, Queen Mary, University of London, United Kingdom

Sponsoring Institutions

Microsoft Research

Leighton
Innovation as standard

Canon

Canon
Research Centre
Europe

University of
Sunderland

INFORMATION RETRIEVAL
SPECIALIST GROUP
BCS

Lecture Notes in Computer Science

For information about Vols. 1–2880

please contact your bookseller or Springer-Verlag

Vol. 3005: G.R. Raidl, S. Cagnoni, J. Branke, D.W. Corne, R. Drechsler, Y. Jin, C.G. Johnson, P. Machado, E. Marchiori, F. Rothlauf, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XVII, 562 pages. 2004.

Vol. 3004: J. Gottlieb, G.R. Raidl (Eds.), Evolutionary Computation in Combinatorial Optimization. X, 241 pages. 2004.

Vol. 2997: S. McDonald, J. Tait (Eds.), Advances in Information Retrieval. XIII, 421 pages. 2004.

Vol. 2996: V. Diekert, M. Habib (Eds.), STACS 2004. XVI, 658 pages. 2004.

Vol. 2995: C. Jensen, S. Poslad, T. Dimitrakos (Eds.), Trust Management. XIII, 377 pages. 2004.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004. (Subseries LNBI).

Vol. 2993: R. Alur, G.J. Pappas (Eds.), Hybrid Systems: Computation and Control. XII, 674 pages. 2004.

Vol. 2992: E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, E. Ferrari (Eds.), Advances in Database Technology - EDBT 2004. XVIII, 877 pages. 2004.

Vol. 2991: R. Alt, A. Frommer, R.B. Kearfott, W. Luther (Eds.), Numerical Software with Result Verification. X, 315 pages. 2004.

Vol. 2989: S. Graf, L. Mounier (Eds.), Model Checking Software. X, 309 pages. 2004.

Vol. 2988: K. Jensen, A. Podelski (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XIV, 608 pages. 2004.

Vol. 2987: I. Walukiewicz (Ed.), Foundations of Software Science and Computation Structures. XIII, 529 pages. 2004.

Vol. 2986: D. Schmidt (Ed.), Programming Languages and Systems. XII, 417 pages. 2004.

Vol. 2985: E. Duesterwald (Ed.), Compiler Construction. X, 313 pages. 2004.

Vol. 2984: M. Wermelinger, T. Margaria-Steffen (Eds.), Fundamental Approaches to Software Engineering. XII, 389 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004. (Subseries LNBI).

Vol. 2982: N. Wakamiya, M. Solarski, J. Sterbenz (Eds.), Active Networks. XI, 308 pages. 2004.

Vol. 2981: C. Müller-Schloer, T. Ungerer, B. Bauer (Eds.), Organic and Pervasive Computing – ARCS 2004. XI, 339 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004. (Subseries LNAI).

Vol. 2978: R. Groz, R.M. Hierons (Eds.), Testing of Communicating Systems. XII, 225 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004. (Subseries LNAI).

Vol. 2976: M. Farach-Colton (Ed.), LATIN 2004: Theoretical Informatics. XV, 626 pages. 2004.

Vol. 2973: Y. Lee, J. Li, K.-Y. Whang, D. Lee (Eds.), Database Systems for Advanced Applications. XXIV, 925 pages. 2004.

Vol. 2970: F. Fernández Rivera, M. Bubak, A. Gómez Tato, R. Doallo (Eds.), Grid Computing. XI, 328 pages. 2004.

Vol. 2964: T. Okamoto (Ed.), Topics in Cryptology – CT-RSA 2004. XI, 387 pages. 2004.

Vol. 2963: R. Sharp, Higher Level Hardware Synthesis. XVI, 195 pages. 2004.

Vol. 2962: S. Bistarelli, Semirings for Soft Constraint Solving and Programming. XII, 279 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004. (Subseries LNAI).

Vol. 2960: P.D. Mosses (Ed.), CASL Reference Manual. XVII, 528 pages. 2004.

Vol. 2958: L. Rauchwerger (Ed.), Languages and Compilers for Parallel Computing. XI, 556 pages. 2004.

Vol. 2957: P. Langendoerfer, M. Liu, I. Matta, V. Tsaoussidis (Eds.), Wired/Wireless Internet Communications. XI, 307 pages. 2004.

Vol. 2954: F. Crestani, M. Dunlop, S. Mizzaro (Eds.), Mobile and Ubiquitous Information Access. X, 299 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004. (Subseries LNAI).

Vol. 2952: N. Guelfi, E. Astesiano, G. Reggio (Eds.), Scientific Engineering of Distributed Java Applications. X, 157 pages. 2004.

Vol. 2951: M. Naor (Ed.), Theory of Cryptography. XI, 523 pages. 2004.

Vol. 2949: R. De Nicola, G. Ferrari, G. Meredith (Eds.), Coordination Models and Languages. X, 323 pages. 2004.

Vol. 2948: G.L. Mullen, A. Poli, H. Stichtenoth (Eds.), Finite Fields and Applications. VIII, 263 pages. 2004.

Vol. 2947: F. Bao, R. Deng, J. Zhou (Eds.), Public Key Cryptography – PKC 2004. XI, 455 pages. 2004.

Vol. 2946: R. Focardi, R. Gorrieri (Eds.), Foundations of Security Analysis and Design II. VII, 267 pages. 2004.

Vol. 2943: J. Chen, J. Reif (Eds.), DNA Computing. X, 225 pages. 2004.

Vol. 2941: M. Wirsing, A. Knapp, S. Balsamo (Eds.), Radical Innovations of Software and Systems Engineering in the Future. X, 359 pages. 2004.

Vol. 2940: C. Lucena, A. Garcia, A. Romanovsky, J. Castro, P.S. Alencar (Eds.), Software Engineering for Multi-Agent Systems II. XII, 279 pages. 2004.

Vol. 2939: T. Kalker, I.J. Cox, Y.M. Ro (Eds.), Digital Watermarking. XII, 602 pages. 2004.

Vol. 2937: B. Steffen, G. Levi (Eds.), Verification, Model Checking, and Abstract Interpretation. XI, 325 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004. (Subseries LNAI).

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004. (Subseries LNAI).

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004. (Subseries LNAI).

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004. (Subseries LNAI).

Vol. 2919: E. Giunchiglia, A. Tacchella (Eds.), Theory and Applications of Satisfiability Testing. XI, 530 pages. 2004.

Vol. 2917: E. Quintarelli, Model-Checking Based Data Retrieval. XVI, 134 pages. 2004.

Vol. 2916: C. Palamidessi (Ed.), Logic Programming. XII, 520 pages. 2003.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004. (Subseries LNAI).

Vol. 2914: P.K. Pandya, J. Radhakrishnan (Eds.), FST TCS 2003: Foundations of Software Technology and Theoretical Computer Science. XIII, 446 pages. 2003.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003. (Subseries LNAI).

Vol. 2911: T.M.T. Sembok, H.B. Zaman, H. Chen, S.R. Urs, S.H. Myaeng (Eds.), Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. XX, 703 pages. 2003.

Vol. 2910: M.E. Orlowska, S. Weerawarana, M.M.P. Papazoglou, J. Yang (Eds.), Service-Oriented Computing - ICSOC 2003. XIV, 576 pages. 2003.

Vol. 2909: R. Solis-Oba, K. Jansen (Eds.), Approximation and Online Algorithms. VIII, 269 pages. 2004.

Vol. 2908: K. Chae, M. Yung (Eds.), Information Security Applications. XII, 506 pages. 2004.

Vol. 2907: I. Lirkov, S. Margenov, J. Wasniewski, P. Yalamov (Eds.), Large-Scale Scientific Computing. XI, 490 pages. 2004.

Vol. 2906: T. Ibaraki, N. Katoh, H. Ono (Eds.), Algorithms and Computation. XVII, 748 pages. 2003.

Vol. 2905: A. Sanfeliu, J. Ruiz-Shulcloper (Eds.), Progress in Pattern Recognition, Speech and Image Analysis. XVII, 693 pages. 2003.

Vol. 2904: T. Johansson, S. Maitra (Eds.), Progress in Cryptology - INDOCRYPT 2003. XI, 431 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003. (Subseries LNAI).

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003. (Subseries LNAI).

Vol. 2901: F. Bry, N. Henze, J. Ma luszyński (Eds.), Principles and Practice of Semantic Web Reasoning. X, 209 pages. 2003.

Vol. 2900: M. Bidoit, P.D. Mosses (Eds.), Casl User Manual. XIII, 240 pages. 2004.

Vol. 2899: G. Ventre, R. Canonico (Eds.), Interactive Multimedia on Next Generation Networks. XIV, 420 pages. 2003.

Vol. 2898: K.G. Paterson (Ed.), Cryptography and Coding. IX, 385 pages. 2003.

Vol. 2897: O. Balet, G. Subsol, P. Torguet (Eds.), Virtual Storytelling. XI, 240 pages. 2003.

Vol. 2896: V.A. Saraswat (Ed.), Advances in Computing Science – ASIAN 2003. VIII, 305 pages. 2003.

Vol. 2895: A. Ohori (Ed.), Programming Languages and Systems. XIII, 427 pages. 2003.

Vol. 2894: C.S. Laih (Ed.), Advances in Cryptology - ASIACRYPT 2003. XIII, 543 pages. 2003.

Vol. 2893: J.-B. Stefani, I. Demeure, D. Hagimont (Eds.), Distributed Applications and Interoperable Systems. XIII, 311 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003. (Subseries LNAI).

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003. (Subseries LNAI).

Vol. 2890: M. Broy, A.V. Zamulin (Eds.), Perspectives of System Informatics. XV, 572 pages. 2003.

Vol. 2889: R. Meersman, Z. Tari (Eds.), On The Move to Meaningful Internet Systems 2003: OTM 2003 Workshops. XIX, 1071 pages. 2003.

Vol. 2888: R. Meersman, Z. Tari, D.C. Schmidt (Eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. XXI, 1546 pages. 2003.

Vol. 2887: T. Johansson (Ed.), Fast Software Encryption. IX, 397 pages. 2003.

Vol. 2886: I. Nyström, G. Sanniti di Baja, S. Svensson (Eds.), Discrete Geometry for Computer Imagery. XII, 556 pages. 2003.

Vol. 2885: J.S. Dong, J. Woodcock (Eds.), Formal Methods and Software Engineering. XI, 683 pages. 2003.

Vol. 2884: E. Najm, U. Nestmann, P. Stevens (Eds.), Formal Methods for Open Object-Based Distributed Systems. X, 293 pages. 2003.

Vol. 2883: J. Schaeffer, M. Müller, Y. Björnsson (Eds.), Computers and Games. XI, 431 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003. (Subseries LNAI).

Vol. 2881: E. Horlait, T. Magedanz, R.H. Glitho (Eds.), Mobile Agents for Telecommunication Applications. IX, 297 pages. 2003.

Table of Contents

Keynote Papers

From Information Retrieval to Information Interaction 1
Gary Marchionini

IR and AI: Traditions of Representation and Anti-representation in Information Processing 12
Yorick Wilks

User Studies

A User-Centered Approach to Evaluating Topic Models 27
Diane Kelly, Fernando Diaz, Nicholas J. Belkin, James Allan

A Study of User Interaction with a Concept-Based Interactive Query Expansion Support Tool 42
Hideo Joho, Mark Sanderson, Micheline Beaulieu

Searcher's Assessments of Task Complexity for Web Searching 57
David J. Bell, Ian Ruthven

Question Answering

Evaluating Passage Retrieval Approaches for Question Answering 72
Ian Roberts, Robert Gaizauskas

Identification of Relevant and Novel Sentences Using Reference Corpus 85
Hsin-Hsi Chen, Ming-Feng Tsai, Ming-Hung Hsu

Answer Selection in a Multi-stream Open Domain Question Answering System 99
Valentin Jijkoun, Maarten de Rijke

Information Models

A Bidimensional View of Documents for Text Categorisation 112
Giorgio Maria Di Nunzio

Query Difficulty, Robustness, and Selective Application of Query Expansion 127
Giambattista Amati, Claudio Carpineto, Giovanni Romano

Combining CORI and the Decision-Theoretic Approach for Advanced Resource Selection 138
Henrik Nottelmann, Norbert Fuhr

Predictive Top-Down Knowledge Improves Neural Exploratory Bottom-Up Clustering 154
Chihli Hung, Stefan Wermter, Peter Smith

Classification

Contextual Document Clustering 167
Vladimir Dobrynin, David Patterson, Niall Rooney

Complex Linguistic Features for Text Classification: A Comprehensive Study 181
Alessandro Moschitti, Roberto Basili

Eliminating High-Degree Biased Character Bigrams for Dimensionality Reduction in Chinese Text Categorization 197
Dejun Xue, Maosong Sun

Summarization

Broadcast News Gisting Using Lexical Cohesion Analysis 209
Nicola Stokes, Eamonn Newman, Joe Carthy, Alan F. Smeaton

From Text Summarisation to Style-Specific Summarisation for Broadcast News 223
Heidi Christensen, BalaKrishna Kolluru, Yoshihiko Gotoh, Steve Renals

Image Retrieval

Relevance Feedback for Cross Language Image Retrieval 238
Paul Clough, Mark Sanderson

NN^k Networks for Content-Based Image Retrieval 253
Daniel Heesch, Stefan Rüger

Integrating Perceptual Signal Features within a Multi-facetted Conceptual Model for Automatic Image Retrieval 267
Mohammed Belkhatir, Philippe Mulhem, Yves Chiaramella

Evaluation Issues

Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary 283
Jaap Kamps

A Study of the Assessment of Relevance for the INEX'02 Test Collection ... 296
Gabriella Kazai, Sherezad Masood, Mounia Lalmas

A Simulated Study of Implicit Feedback Models ... 311
Ryen W. White, Joemon M. Jose, C.J. van Rijsbergen, Ian Ruthven

Cross Language IR

Cross-Language Information Retrieval Using EuroWordNet and Word Sense Disambiguation ... 327
Paul Clough, Mark Stevenson

Fault-Tolerant Fulltext Information Retrieval in Digital Multilingual Encyclopedias with Weighted Pattern Morphing ... 338
Wolfram M. Esser

Measuring a Cross Language Image Retrieval System ... 353
Mark Sanderson, Paul Clough, Catherine Paterson, Wai Tung Lo

Web-Based and XML IR

An Optimistic Model for Searching Web Directories ... 364
Fidel Cacheda, Ricardo Baeza-Yates

Content-Aware DataGuides: Interleaving IR and DB Indexing Techniques for Efficient Retrieval of Textual XML Data ... 378
Felix Weigel, Holger Meuss, François Bry, Klaus U. Schulz

Performance Analysis of Distributed Architectures to Index One Terabyte of Text ... 394
Fidel Cacheda, Vassilis Plachouras, Iadh Ounis

Applying the Divergence from Randomness Approach for Content-Only Search in XML Documents ... 409
Mohammad Abolhassani, Norbert Fuhr

Author Index ... 421

From Information Retrieval to Information Interaction

Gary Marchionini

University of North Carolina at Chapel Hill, School of Information and Library Science
100 Manning Hall
Chapel Hill, NC 27599, USA
march@ils.unc.edu

Abstract. This paper argues that a new paradigm for information retrieval has evolved that incorporates human attention and mental effort and takes advantage of new types of information objects and relationships that have emerged in the WWW environment. One aspect of this new model is attention to highly interactive user interfaces that engage people directly and actively in information seeking. Two examples of these kinds of interfaces are described.

1 Introduction

Information retrieval (IR) is hot. After 40 years of systematic research and development, often ignored by the public, technology and a global information economy have conspired to make IR a crucial element of the emerging cyberinfrastrucure and a field of interest for the best and brightest students. The new exciting employers are Google, Amazon, and eBay and the extant giants like IBM and Microsoft have active IR research and development groups. In many ways, research in IR had plateaued until the WWW breathed new life into it by supporting a global marketplace of electronic information exchange. In fact, I argue that the IR problem itself has fundamentally changed and a new paradigm of information interaction has emerged. This argument is made in two parts: first, the evolution of IR will be considered by a broad look at today's information environment and trends in IR research and development and second, examples of attempts to address IR as an interactive process that engages human attention and mental effort will be given.

2 Information Objects and People

As a scientific area, IR uses analysis to break down the whole problem into components and first focus on the components that promise to yield to our techniques. IR has always been fundamentally concerned with information objects and with the people who create, find, and use those objects; however, because people are less predictable and more difficult and expensive to manipulate experimentally, IR research logically focused on the information objects first. Traditionally, information objects have been taken to be documents and queries and research has centered on two basic issues: representation of those objects and definition of the relationships

S. McDonald and J. Tait (Eds.): ECIR 2004, LNCS 2997, pp. 1–11, 2004.

among them. Representation is a classical issue in philosophy, information science (e.g., Heilprin argued that compression was the central representation problem [9]), and artificial intelligence. The IR community has demonstrated a variety of effective representations for documents and queries, including linguistic (e.g., controlled vocabulary) assignments and a large variety of mathematical assignments (e.g., vectors) based on term-occurrence, relevance probability estimates, and more recently hyperlink graphs. IR research has mainly focused on equality (e.g., of index terms) and similarity relationships—similarity between/among objects—and developed a large variety of matching algorithms that are exploited in today's retrieval systems. A schematic for the traditional IR problem is depicted in Figure 1.

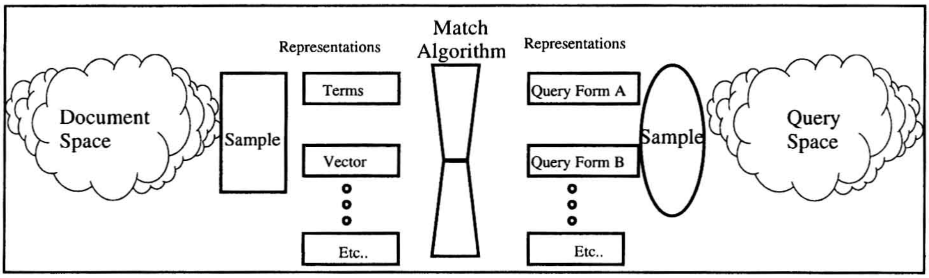

Fig. 1. Content-Centered Retrieval as Matching Document Representations to Query Representations

The figure shows that samples of document and query objects from the respective universe of all objects are each represented in some fashion, most often using the same representation form. For example, a simple approach used in early commercial retrieval systems was to represent documents and queries with terms assigned from a controlled vocabulary and simply match overlaps. A more contemporary example returns ranked sets of similarities by representing documents and queries as vectors of inverse document frequency values for a specific set of terms in the sample ordered by cosine similarity. In cases where the document and query representations are in different forms (e.g., different metadata schemes or human languages), crosswalks, translations, or interlingua must also be added to the process. This content-centered paradigm has driven creative work and led to mainly effective retrieval systems (e.g., SMART, Okapi, Iquery), however, progress toward improving both recall and precision seems to have reached a diminishing return state.

Two important changes have been taking place in the electronic information environment that expand this schema and stimulate new kinds of IR research and development. These changes are due to new types and properties of information objects and to increasing attention to human participation in the IR process. The IR community has begun to recognize these changes as illustrated by the two grand research and development challenges identified for IR research at a recent strategic workshop [1]: global information access ("Satisfy human information needs through natural, efficient interaction with an automated system that leverages world-wide structured and unstructured data in any language."), and contextual retrieval ("Combine search technologies and knowledge about query and user context into a single framework in order to provide the most 'appropriate' answer for a user's information needs." P.330).