

Proceedings of the 21st Annual International ACM SIGIR Conference
on Research and Development in Information Retrieval

Melbourne, Australia

ninety-eight

SIGIR

August 24–28, 1998

Editors

*W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen,
Ross Wilkinson and Justin Zobel*



Contents

Keynote Address

The Future of Internet Search 1
Steve Kirsch (*Infoseek*)

Session 1 — Opening Session

Advantages of Query Biased Summaries in Information Retrieval 2
Anastasios Tombros (*University of Glasgow*), Mark Sanderson (*University of Massachusetts*)

A Theory of Term Weighting Based on Exploratory Data Analysis 11
Warren R. Greiff (*University of Massachusetts*)

New Techniques for Open-Vocabulary Spoken Document Retrieval 20
Martin Wechsler, Eugen Munteanu, Peter Schäuble (*ETH Zürich*)

Session 2 — Event Detection and Clustering

A Study on Retrospective and On-Line Event Detection 28
Yiming Yang, Tom Pierce, Jaime Carbonell (*Carnegie Mellon University*)

On-Line New Event Detection and Tracking 37
James Allan, Ron Papka, Victor Lavrenko (*University of Massachusetts*)

Web Document Clustering: A Feasibility Demonstration 46
Oren Zamir, Oren Etzioni (*University of Washington*)

Session 3A — Cross-Language Retrieval

The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval 55
Ari Pirkola (*University of Tampere*)

Resolving Ambiguity for Cross-Language Retrieval 64
Lisa Ballesteros, W. Bruce Croft (*University of Massachusetts*)

Cross-Language Information Retrieval with the UMLS Metathesaurus 72
David Eichmann, Miguel E. Ruiz, Padmini Srinivasan (*University of Iowa*)

Session 3B — Categorization

Using a Generalized Instance Set for Automatic Text Categorization 81
Wai Lam, Chao Yang Ho (*Chinese University of Hong Kong*)

Automatic Essay Grading using Text Categorization Techniques 90
Leah S. Larkey (*University of Massachusetts*)

Distributional Clustering of Words for Text Classification 96
L. Douglas Baker (*Carnegie Mellon University*), Andrew Kachites McCallum (*Just Research*)

Session 4A — Distributed Retrieval

Improved Algorithms for Topic Distillation in a Hyperlinked Environment 104
Krishna Bharat, Monika R. Henzinger (*Digital Equipment Corporation*)

Effective Retrieval with Distributed Collections 112
Jinxi Xu, Jamie Callan (*University of Massachusetts*)

Evaluating Database Selection Techniques: A Testbed and Experiment.....	121
James C. French, Allison L. Powell (<i>University of Virginia</i>), Charles L. Viles (<i>University of North Carolina</i>), Travis Emmitt, Kevin J. Prey (<i>University of Virginia</i>)	
Session 4B — Using Structure	
The Impact of Query Structure and Query Expansion on Retrieval Performance.....	130
Jaana Kekäläinen, Kalervo Järvelin (<i>University of Tampere</i>)	
A Flexible Model for Retrieval of SGML Documents.....	138
Sung Hyon Myaeng, Dong-Hyun Jang (<i>Chungnam National University</i>), Mun-Seok Kim, Zong-Cheol Zoo (<i>Systems Engineering Research Institute</i>)	
Discovering Typical Structures of Documents: A Road Map Approach.....	146
Ke Wang, Huiqing Liu (<i>National University of Singapore</i>)	
Session 5A — Interactive Retrieval	
A Cognitive Model for Searching for Ill-Defined Targets on the Web: The Relationship between Search Strategies and User Satisfaction.....	155
Mari Saito, Kazunori Ohmura (<i>Sony</i>)	
Comparing Interactive Information Retrieval Systems Across Sites: The TREC-6 Interactive Track Matrix Experiment.....	164
Eric Lagergren, Paul Over (<i>National Institute of Standards and Technology</i>)	
Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems	173
Russell C. Swan, James Allan (<i>University of Massachusetts</i>)	
Session 5B — Combination of Evidence	
Modeling and Combining Evidence Provided by Document Relationships Using Probabilistic Argumentation Systems.....	182
Justin Picard (<i>Université de Neuchâtel</i>)	
Predicting the Performance of Linearly Combined IR Systems.....	190
Christopher C. Vogt, Garrison W. Cottrell (<i>University of California, San Diego</i>)	
Experiments in Japanese Text Retrieval and Routing using the NEAT System.....	197
Gareth J.F. Jones, Tetsuya Sakai, Masahiro Kajiura, Kazuo Sumita (<i>Toshiba</i>)	
Session 6A — Query and Profile Modification	
Improving Automatic Query Expansion.....	206
Mandar Mitra (<i>Cornell University</i>), Amit Singhal (<i>AT&T Labs</i>), Chris Buckley (<i>Sabir Research</i>)	
Boosting and Rocchio Applied to Text Filtering.....	215
Robert E. Schapire, Yoram Singer, Amit Singhal (<i>AT&T Labs</i>)	
Learning While Filtering Documents.....	224
Jamie Callan (<i>University of Massachusetts</i>)	
Session 6B — Information Retrieval Experiments	
Spatial Querying for Image Retrieval: A User-Oriented Evaluation.....	232
Joemon M. Jose, Jonathan Furner, David J. Harper (<i>Robert Gordon University</i>)	

Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach	241
Shian-Hua Lin (<i>National Cheng Kung University</i>), Chi-Sheng Shih, Meng Chang Chen, Jan-Ming Ho, Ming-Tat Ko (<i>Academia Sinica</i>), Yueh-Ming Huang (<i>National Cheng Kung University</i>)	
Improving Two-Stage Ad-Hoc Retrieval for Short Queries	250
K.L. Kwok, M. Chan (<i>City University of New York</i>)	
Session 7A — Retrieval Models	
DOLORES: A System for Logic-Based Retrieval of Multimedia Objects	257
Norbert Fuhr, Norbert Gövert, Thomas Rölleke (<i>University of Dortmund</i>)	
RELIEF: Combining Expressiveness and Rapidity into a Single System	266
Iadh Ounis, Marius Paşca (<i>Université de Grenoble</i>)	
A Language Modeling Approach to Information Retrieval	275
Jay M. Ponte, W. Bruce Croft (<i>University of Massachusetts</i>)	
Session 7B — Efficiency	
Efficient Construction of Large Test Collections	282
Gordon V. Cormack, Christopher R. Palmer (<i>University of Waterloo</i>), Charles L.A. Clarke (<i>University of Toronto</i>)	
Compressed Inverted Files with Reduced Decoding Overheads	290
Vo Ngoc Anh, Alistair Moffat (<i>University of Melbourne</i>)	
Fast Searching on Compressed Text Allowing Errors	298
Edleno Silva de Moura (<i>Universidade Federal de Minas Gerais</i>), Gonzalo Navarro (<i>Universidad de Chile</i>), Nivio Ziviani (<i>Universidade Federal de Minas Gerais</i>), Ricardo Baeza-Yates (<i>Universidad de Chile</i>)	
Session 8 — Information Retrieval Experiments	
How Reliable are the Results of Large-Scale Information Retrieval Experiments?	307
Justin Zobel (<i>RMIT</i>)	
Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness	315
Ellen M. Voorhees (<i>National Institute of Standards and Technology</i>)	
Measures of Relative Relevance and Ranked Half-Life: Performance Indicators for Interactive IR	324
Pia Borlund, Peter Ingwersen (<i>Royal School of Library and Information Science</i>)	
Panel Session	
Tools for Searching the Web	332
Donna Harman, Paul Over (<i>National Institute of Standards and Technology</i>)	
Poster Abstracts	
Modern Classical Document Indexing: A Linguistic Contribution to Knowledge-Based IR	333
Bas van Bakel (<i>University of Twente</i>)	
The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries	335
Jaime Carbonell, Jade Goldstein (<i>Carnegie Mellon University</i>)	

A Method for Scoring Correlated Features in Query Expansion	337
Martin Franz, Salim Roukos (<i>IBM T.J. Watson Research Center</i>)	
Using Maps as a User Interface to a Digital Library.....	339
Mountaz Hascoët (<i>Université Paris-sud</i>), Xavier Soinard (<i>Electricité de France</i>)	
Comparison between Proximity Operation and Dependency Operation in Japanese Full-Text Retrieval.....	341
Yasuaki Hyoudo, Kazuhiko Niimi, Takashi Ikeda (<i>Gifu University</i>)	
Term-Ordered Query Evaluation versus Document-Ordered Query Evaluation for Large Document Databases.....	343
Marcin Kaszkiel, Justin Zobel (<i>RMIT</i>)	
Lessons From BMIR-J2: A Test Collection for Japanese IR Systems	345
Tsuyoshi Kitani (<i>NTT</i>), Yasushi Ogawa (<i>Ricoh</i>), Tetsuya Ishikawa (<i>ULIS</i>), Haruo Kimoto (<i>NTT</i>), Ikuo Keshi (<i>Sharp</i>), Jun Toyoura (<i>Mitsubishi Electric</i>), Toshikazu Fukushima (<i>NEC</i>), Kunio Matsui (<i>Fujitsu Laboratories</i>), Yoshihiro Ueda (<i>Fuji Xerox</i>), Tetsuya Sakai (<i>Toshiba</i>), Takenobu Tokunaga (<i>Tokyo Institute of Technology</i>), Hiroshi Tsuruoka (<i>University of Tokyo</i>), Hidekazu Nakawatase (<i>NTT</i>), Teru Agata (<i>Keio University</i>)	
Automatically Locating, Extracting and Analyzing Tabular Data	347
William Kornfeld, John Wattecamps	
Using Global Colour Features for General Photographic Image Indexing and Retrieval.....	349
Ting-Sheng Lai, John Tait (<i>University of Sunderland</i>)	
Automatic Acquisition of Phrasal Knowledge for English-Chinese Bilingual Information Retrieval	351
Ming-Jer Lee, Lee-Feng Chien (<i>Academia Sinica</i>)	
Visual Interactions with a Multidimensional Ranked List	353
Anton Leouski, James Allan (<i>University of Massachusetts</i>)	
Predicting Query Times.....	355
Rodger McNab, Yong Wang, Ian H. Witten, Carl Gutwin (<i>University of Waikato</i>)	
The WebCluster Project: Using Clustering for Mediating Access to the World Wide Web.....	357
Mourad Mechkour, David J. Harper, Gheorghe Muresan (<i>Robert Gordon University</i>)	
Automatic Abstracting of Magazine Articles: The Creation of 'Highlight' Abstracts	359
Marie-Francine Moens, Jos Dumortier (<i>Katholieke Universiteit Leuven</i>)	
Optimizing Recall/Precision Scores in IR over the WWW.....	361
Matthew Montebello (<i>Cardiff University</i>)	
Interactive Multidimensional Document Visualization.....	363
Josiane Mothe, Taoufiq Dkaki (<i>Institut de Recherche en Informatique de Toulouse</i>)	
Speech Retrieval using Phonemes with Error Correction	365
Corinna Ng, Justin Zobel (<i>RMIT</i>)	
Optimizing Query Evaluation in <i>n</i> -Gram Indexing.....	367
Yasushi Ogawa, Toru Matsuda (<i>Ricoh</i>)	
Four Text Classification Algorithms Compared on a Dutch Corpus.....	369
Hein Ragas (<i>Cap Gemini Netherlands</i>), Cornelis H.A. Koster (<i>University of Nijmegen</i>)	

Automatic Acquisition of Terminological Relations from a Corpus for Query Expansion	371
Jean-David Sta (<i>Electricité de France</i>)	
Keyword Extraction of Radio News using Term Weighting with an Encyclopedia and Newspaper Articles	373
Yoshimi Suzuki, Fumiyo Fukumoto, Yoshihiro Sekiguchi (<i>Yamanashi University</i>)	
Efficient Search Server Assignment in a Disproportionate System Environment	375
Toru Takaki, Tsuyoshi Kitani (<i>NTT</i>)	
Multilingual Keyword Extraction for Term Suggestion	377
Yuen-Hsien Tseng (<i>Fu Jen Catholic University</i>)	
Experiments of Collecting WWW Information using Distributed WWW Robots	379
Hayato Yamana (<i>Electrotechnical Laboratory</i>), Kent Tamura (<i>IBM Tokyo Research Laboratory</i>), Hiroyuki Kawano, Satoshi Kamei (<i>Kyoto University</i>), Masanori Harada (<i>University of Tokyo</i>), Hideki Nishimura (<i>Sharp</i>), Isao Asai (<i>Osaka Prefecture University</i>), Hiroyuki Kusumoto (<i>Keio University</i>), Yoichi Shinoda (<i>JAIST</i>), Yoichi Muraoka (<i>Waseda University</i>)	
Demonstration Abstracts	
Presenting Web Site Search Results in Context: A Demonstration	381
Michael Chen, Marti A. Hearst (<i>University of California, Berkeley</i>)	
Personal Browser	382
Yi-Shiou Chen, Schy Chiou, Yuan-Kai Wang, Wen-Lian Hsu (<i>Academia Sinica</i>)	
Towards a Fast Precision-Oriented Image Retrieval System	383
Yves Chiaramella, Philippe Mulhem, Mourad Mechkour, Iadh Ounis, Marius Paşca (<i>Université de Grenoble</i>)	
Teraphim: An Engine for Distributed Information Retrieval	384
Owen de Kretser, Alistair Moffat (<i>University of Melbourne</i>), Justin Zobel (<i>RMIT</i>)	
Cheshire II: Combining Probabilistic and Boolean Retrieval	385
Ray R. Larson (<i>University of California, Berkeley</i>)	
A Research Prototype Image Retrieval System	386
S. Nepal, M.V. Ramakrishna, J.A. Thom (<i>RMIT</i>)	
The Structured Information Manager (SIM)	387
Ron Sacks-Davis, Alan Kent (<i>RMIT</i>)	
PWA: An Extended Probabilistic Web Algebra	388
Dan Smith, Rattasit Sukhahuta (<i>University of East Anglia</i>)	
Cafe: An Indexed Approach to Searching Genomic Databases	389
Hugh E. Williams (<i>RMIT</i>)	
Fast Speculative Search Engine on the Highly Parallel Computer EM-X	390
Hayato Yamana, Hanpei Koike, Yuetsu Kodama, Hirofumi Sakane, Yoshinori Yamaguchi (<i>Electrotechnical Laboratory</i>)	
Author Index	392