

Table of Contents

Salton Award Lecture

On theoretical argument in information retrieval	1
Stephen Robertson (<i>Microsoft Research, England</i>)	

Session 1. Relevance

Relevance and contributing information types of searched documents in task performance ...	2
Pertti Vakkari. (<i>University of Tampere, Finland</i>)	
Relevance feedback with a small number of relevance judgments: incremented relevance feedback versus document clustering	10
Makoto Iwayama. (<i>Hitachi, Ltd., Japan</i>)	

Session 2. Evaluation

Do batch and user evaluations give the same results?	17
William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, Daniel Olson. (<i>Oregon Health Sciences University, USA</i>)	
A novel method for the evaluation of Boolean query effectiveness across a wide operational range	25
Eero Sormunen. (<i>University of Tampere, Finland</i>)	
Evaluating evaluation measure stability	33
Chris Buckley (<i>Sabir Research Inc, USA</i>) Ellen Voorhees. (<i>NIST, USA</i>)	
IR evaluation methods for retrieving highly relevant documents	41
Kalervo Järvelin, Jaana Kekäläinen. (<i>University of Tampere, Finland</i>)	

Session 3. Topic Detection and Tracking

Automatic generation of overview timelines	49
Russell Swan, James Allan. (<i>University of Massachusetts, Amherst, USA</i>)	
Event tracking based on domain dependency	57
Fumiyo Fukumoto, Yoshimi Suzuki. (<i>Yamanashi University, Japan</i>)	
Improving text categorization methods for event tracking	65
Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer. (<i>Carnegie Mellon University, USA</i>)	

Session 4. Multimedia Information Retrieval

Evaluating a simple and effective music information retrieval method	73
J. Stephen Downie. (<i>University of Illinois at Urbana-Champaign, USA</i>) , Michael Nelson (<i>University of Western Ontario, Canada</i>)	
Phonetic confusion matrix-based spoken document retrieval	81
Savitha Srinivasan, Dragutin Petkovic. (<i>IBM Almaden Research Center, USA</i>)	
Multiple evidence combination in image retrieval: Diogenes searches for people on the web ..	88
Yusuk Alp Aslandogan, Clement T. Yu. (<i>University of Illinois at Chicago, USA</i>)	

Session 5. Theory and Practice in Information Retrieval

Link-based and content-based evidential information in a belief network model 96
Ilmério R. Silva (*Universidade Federal de Uberlândia, Brazil*) Berthier Ribeiro-Neto, Pável
Calado, Nívio Ziviani (*Universidade Federal de Minas Gerais, Brazil*) Edleno S. Moura
(*Universidade do Amazonas, Brazil*)

The feature quantity: an information-theoretic perspective of tfidf-like measures 104
Akiko Aizawa. (*National Center for Science Information Systems, Japan*)

INSYDER - An information assistant for business intelligence 112
Harald Reiterer, Gabriela Mussler, Thomas H. Mann, Siegfried Handschuh. (*University of
Konstanz, Germany*)

Structured translation for cross-language IR 120
Ruth Sperer, Douglas W. Oard. (*University of Maryland, USA*)

Session 6. Natural Language Processing and Summarization for Information Retrieval

Automatic adaptation of proper noun dictionaries through co-operation of machine
learning and probabilistic methods 128
Georgios Petasis, Giorgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos.
(*National Centre for Scientific Research "Demokritos", Greece*), Alessandro Cucchiarelli
(*Università di Ancona, Italy*) Paola Velardi (*Università di Roma "La Sapienza", Italy*)

Document centered approach for text normalization 136
Andrei Mikheev. (*University of Edinburgh, Scotland*)

OCELOT: a system for summarizing web pages 144
Adam Berger (*Carnegie Mellon University, USA*), Vibhu O. Mittal (*Just Research, USA*)

Extracting sentence segments for text summarization: a machine learning approach 152
Wesley T. Chuang (*UCLA, USA*), Jihoon Yang (*HRL Laboratories, LLC, USA*)

Session 7. Information Filtering

An experimental comparison of naive Bayesian and keyword-based anti-spam
filtering with personal email messages 160
Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinou, Constantine D.
Spyropoulos. (*National Center for Scientific Research "Demokritos", Greece*)

Text filtering by boosting naive Bayes classifiers 168
Yu-Hwan Kim, Shang-Yoon Hahn, Byoung-Tak Zhang (*Seoul National University, Korea*)

Document filtering method using non-relevant information profile 176
Keiichiro Hoashi, Kazunori Matsumoto, Naomi Inoue, Kazuo Hashimoto (*KDD R&D
Laboratories, Inc., Japan*)

Session 8. Question Answering

Question-answering by predictive annotation 184
John Prager, Eric Brown, Anni Coden (*IBM T.J. Watson Laboratories, USA*),
Dragomir Radev (*University of Michigan, USA*)

Bridging the lexical chasm: statistical approaches to answer-finding	192
Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, Vibhu Mittal. (<i>Just Research, Carnegie-Mellon University, USA</i>)	
Building a question-answering test collection	200
Ellen M. Voorhees, Dawn Tice. (<i>NIST, USA</i>)	
Session 9. Clustering	
Document clustering using word clusters via the information bottleneck method	208
Noam Slonim, Naftali Tishby. (<i>The Hebrew University, Israel</i>)	
Latent semantic space: iterative scaling improves precision of inter-document similarity measurement	216
Rie Kubota Ando. (<i>Cornell University and IBM T.J. Watson Laboratory, USA</i>)	
An investigation of linguistic features and clustering algorithms for topical document clustering	224
Vasileios Hatzivassiloglou, Luis Gravano, Ankinedu Maganti. (<i>Columbia University, USA</i>)	
The impact of database selection on distributed searching	232
Allison L. Powell, James C. French (<i>University of Virginia, USA</i>), Jamie Callan (<i>Carnegie Mellon University, USA</i>), Margaret Connell (<i>University of Massachusetts, Amherst, USA</i>), Charles L. Viles (<i>University of North Carolina, Chapel Hill, USA</i>)	
Session 10. Efficiency	
Hill climbing algorithms for content-based retrieval of similar configurations	240
Dimitris Papadias. (<i>Hong Kong University of Science and Technology, China</i>)	
Partial collection replication versus caching for information retrieval systems	248
Zhihong Lu, Kathryn S. McKinley (<i>University of Massachusetts, Amherst, USA</i>)	
Session 11. Hypertext Classification	
Hierarchical classification of web content	256
Susan Dumais (<i>Microsoft Research, USA</i>), Hao Chen (<i>University of California, Berkeley, USA</i>)	
A practical hypertext categorization method using links and incrementally available class information	264
Hyo-Jung Oh, Sung Hyon Myaeng, Mann-Ho Lee (<i>Chungnam National University, Korea</i>)	
Session 12. World Wide Web Information Retrieval	
Topical locality in the Web	272
Brian D. Davison (<i>Rutgers University, USA</i>)	
Interactive internet search: keyword, directory and query reformulation mechanisms compared	280
Peter Bruza, Robert McArthur, Simon Dennis (<i>University of Queensland, Australia</i>)	
Incorporating quality metrics in centralized/distributed information retrieval on the WWW ...	288
Xiaolan Zhu, Susan Gauch (<i>University of Kansas, USA</i>)	

Does "authority" mean quality? Predicting expert quality ratings of web documents	296
Brian Amento (<i>Virginia Tech, USA</i>), Loren Terveen, Will Hill (<i>AT&T Shannon Laboratories, USA</i>)	
Posters.	
Document classification on neural networks using only positive examples	304
Larry M. Manevitz and Malik Yousef (<i>University of Haifa, Israel</i>)	
New paradigms in information visualization	307
Peter Au, Yike Guo, Stefa M. Rueger, Shalini Sewraz (<i>Imperial College, England</i>)	
Latent semantic indexing model for Boolean query formulation	310
DaeHo Baek, HaeChang Rim (<i>Korea University, Korea</i>), HeuiSeok Lim (<i>Chonan University, Korea</i>)	
Generation of user profiles for information filtering research agenda	313
Tsvi Kuflik, Peretz Shoval (<i>University of the Negev, Israel</i>)	
Variance based classifier comparison in text categorization	316
Atsuhiko Takasu, Kenro Aihara (<i>National Center for Science Information Systems, Japan</i>)	
The use of phrases from query texts in information retrieval	318
Masumi Narita, Yasushi Ogawa (<i>Software Research Center, Ricoh Co., Ltd., Japan</i>)	
Pseudo-frequency method: an efficient document ranking retrieval method for n-gram indexing	321
Yasushi Ogawa (<i>Software Research Center, Ricoh Co., Ltd., Japan</i>)	
Lexical semantic relatedness and online news event detection	324
Nicola Stokes, Paula Hatch, Joe Carthy (<i>University College Dublin, Ireland</i>)	
Modeling question response patterns by scaling of visualization	326
Mark E. Rorvig (<i>University of North Texas, USA</i>)	
The effect of query type on subject searching behavior of image databases: an exploratory study	328
Efthimis N. Efthimiadis, Raya Fidel (<i>University of Washington, USA</i>)	
The role of judge in a user based retrieval experiment	331
Mingfang Wu, Ross Wilkinson (<i>CSIRO – MIS, Australia</i>), Michael Fuller (<i>Royal Melbourne Institute of Technology, Australia</i>)	
Auto-construction of a live thesaurus from search term logs for interactive Web search	334
Shuie-Lung Juang, Wen-Hsiang Lu, Lee-Feng Chien (<i>Institute of Information Science, Academia Sinica, Taiwan</i>), Hsiao-Tieh Pu (<i>Shih Hsin University, Taiwan</i>)	
Cognitive approach for building user model in an information retrieval context	337
Amina Sayeb Belhassen, Nabil Ben Abdallah, Henda Hadjami Ben Ghezala (<i>PGL, ENSI Tunisia</i>)	
Multimedia information retrieval from recorded presentations	339
Wolfgang Hurst, Rainer Muller, Christoph Mayer (<i>Computer Science Institute, University of Freiburg, Germany</i>)	

Influence of speech recognition errors on topic detection	342
J. Scott McCarley, Martin Franz (<i>IBM TJ Watson Research Center, USA</i>)	
Word document density & relevance scoring	345
Martin Franz, J. Scott McCarley (<i>IBM TJ Watson Research Center, USA</i>)	
Ranking digital images using combination of evidence	348
Iadh Ounis (<i>University of Glasgow, Scotland</i>)	
Collaborative filtering and the generalized vector space model	351
Ian Soboroff and Charles Nicholas (<i>University of Maryland, Baltimore County, USA</i>)	
Theme-based retrieval of Web news	354
Nuno Maria, Mario J. Silva (<i>Universidade de Lisboa, Portugal</i>)	
Stemming and its effects on TFIDF ranking	357
Mark Kantrowitz (<i>Just Research, USA</i>), Behrang Mohit, Vibhu Mittal (<i>Carnegie Mellon University, USA</i>)	
Exploration of a heuristic approach to threshold learning in adaptive filtering	360
Chengxiang Zhai, Peter Jansen, David A. Evans (<i>CLARITECH Corp., USA</i>)	
On the design and evaluation of a multi-dimensional approach to information retrieval	363
M. Catherine McCabe (<i>U.S. Government, USA</i>), Jinho Lee, Abdur Chowdhury, David Grossman, Ophir Frieder (<i>Illinois Institute of Technology, USA</i>)	
SWAMI: A framework for collaborative filtering algorithm development and evaluation	366
Danyel Fisher, Kris Hildrum, Jason Hong, Mark Newman, Megan Thomas, Richard Vuduc (<i>University of California, Berkeley, USA</i>)	
Learning probabilistic models of the Web	369
Thomas Hofmann (<i>Brown University, USA</i>)	
Effects of out of vocabulary words in spoken document retrieval	372
P. C. Woodland, S. E. Johnson, P. Jourlin, K. Spärck Jones (<i>Cambridge University, England</i>)	
Towards an adaptive and task-specific ranking mechanism in web searching	375
Chen Ding, Chi-Hung Chi (<i>National University of Singapore, Singapore</i>)	
Beyond the traditional query operators	377
Chen Ding, Chi-Hung Chi (<i>National University of Singapore, Singapore</i>)	
Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems	379
Javed A. Aslam, Mark Montague (<i>Dartmouth College, USA</i>)	
Information access for context-aware appliances	382
Gareth J. F. Jones, Peter J. Brown (<i>University of Exeter, England</i>)	
Finding relevant passages using noun-noun compound: coherence vs. proximity	385
E. Hoenkamp, Rob de Groot (<i>Nijmegen Institute for Cognition and Information, The Netherlands</i>)	

Demonstrations.

Semantic Explorer™ - Navigation in document collections; Proxima Daily™ -Learning personal newspaper	388
Vadim Asadov, Serge Shumsky (<i>NeurOK, LLC, Russia</i>)	
Integrated search tools for newspaper digital libraries	389
S.L. Mantzaris, B. Gatos, N. Gouraros, P. Tzavelis (<i>Lambrakis Press Archives, Greece</i>)	
Managing photos with AT&T Shoebox	390
Timothy J. Mills, David Pye, David Sinclair, Kenneth R. Wood (<i>AT&T Laboratories, England</i>)	
ClusterBook, a tool for dual information access	391
Gheorghe Muresan, David J. Harper, Ayse Göker, Peter Loweit (<i>Robert Gordon University, Scotland</i>)	
Uexküll: an interactive visual user interface for document retrieval in vector space	392
Michael Preminger, Sandor Daranyi (<i>Oslo College, Norway</i>)	
Time Mine: visualizing automatically constructed timelines	393
Russell Swan, James Allan (<i>University of Massachusetts, Amherst, USA</i>)	
The Cambridge University multimedia document retrieval demo system	394
A. Tuerk, S.E. Johnson, P. Jourlin, K. Spärck Jones, P.C. Woodland (<i>Cambridge University, England</i>)	
Author Index	395