SIGIR 2000

Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval

ACM SIGIR

July 24-28, 2000
Athens, Greece

Edited by Nicholas J. Belkin, Peter Ingwersen and Mun-Kew Leong

acm PRESS

Table of Contents

Salton Award Lecture

On theoretical argument in information retrieval 1
Stephen Robertson *(Microsoft Research, England)*

Session 1. Relevance

Relevance and contributing information types of searched documents in task performance ... 2
Pertti Vakkari. *(University of Tampere, Finland)*

Relevance feedback with a small number of relevance judgments: incremented relevance feedback versus document clustering 10
Makoto Iwayama. *(Hitachi, Ltd., Japan)*

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. *(Oregon Health Sciences University, USA)*

A novel method for the evaluation of Boolean query effectiveness across a wide operational range 25
Eero Sormunen. *(University of Tampere, Finland)*

Evaluating evaluation measure stability 33
Chris Buckley *(Sabir Research Inc, USA)* Ellen Voorhees. *(NIST, USA)*

IR evaluation methods for retrieving highly relevant documents 41
Kalervo Järvelin, Jaana Kekäläinen. *(University of Tampere, Finland)*

Session 3. Topic Detection and Tracking

Automatic generation of overview timelines 49
Russell Swan, James Allan. *(University of Massachusetts, Amherst, USA)*

Event tracking based on domain dependency 57
Fumiyo Fukumoto, Yoshimi Suzuki. *(Yamanashi University, Japan)*

Improving text categorization methods for event tracking 65
Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer. *(Carnegie Mellon University, USA)*

Session 4. Multimedia Information Retrieval

Evaluating a simple and effective music information retrieval method 73
J. Stephen Downie. *(University of Illinois at Urbana-Champaign, USA)* , Michael Nelson *(University of Western Ontario, Canada)*

Phonetic confusion matrix-based spoken document retrieval 81
Savitha Srinivasan, Dragutin Petkovic. *(IBM Almaden Research Center, USA)*

Multiple evidence combination in image retrieval: Diogenes searches for people on the web .. 88
Yusuk Alp Aslandogan, Clement T. Yu. *(University of Illinois at Chicago, USA)*

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 dictionaires 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. Chandrinos, 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. *(Just Research, Carnegie-Mellon University, USA)*

Building a question-answering test collection 200
Ellen M. Voorhees, Dawn Tice. *(NIST, USA)*

Session 9. Clustering

Document clustering using word clusters via the information bottleneck method 208
Noam Slonim, Naftali Tishby. *(The Hebrew University, Israel)*

Latent semantic space: iterative scaling improves precision of inter-document similarity measurement 216
Rie Kubota Ando. *(Cornell University and IBM T.J. Watson Laboratory, USA)*

An investigation of linguistic features and clustering algorithms for topical document clustering 224
Vasileios Hatzivassiloglou, Luis Gravano, Ankineedu Maganti. *(Columbia University, USA)*

The impact of database selection on distributed searching 232
Allison L. Powell, James C. French *(University of Virginia, USA)*, Jamie Callan *(Carnegie Mellon University, USA)*, Margaret Connell *(University of Massachusetts, Amherst, USA)*, Charles L. Viles *(University of North Carolina, Chapel Hill, USA)*

Session 10. Efficiency

Hill climbing algorithms for content-based retrieval of similar configurations 240
Dimitris Papadias. *(Hong Kong University of Science and Technology, China)*

Partial collection replication versus caching for information retrieval systems 248
Zhihong Lu, Kathryn S. McKinley *(University of Massachusetts, Amherst, USA)*

Session 11. Hypertext Classification

Hierarchical classification of web content 256
Susan Dumais *(Microsoft Research, USA)*, Hao Chen *(University of California, Berkeley, USA)*

A practical hypertext categorization method using links and incrementally available class information 264
Hyo-Jung Oh, Sung Hyon Myaeng, Mann-Ho Lee *(Chungnam National University, Korea)*

Session 12. World Wide Web Information Retrieval

Topical locality in the Web 272
Brian D. Davison *(Rutgers University, USA)*

Interactive internet search: keyword, directory and query reformulation mechanisms compared 280
Peter Bruza, Robert McArthur, Simon Dennis *(University of Queensland, Australia)*

Incorporating quality metrics in centralized/distributed information retrieval on the WWW ... 288
Xiaolan Zhu, Susan Gauch *(University of Kansas, USA)*

Does "authority" mean quality? Predicting expert quality ratings of web documents 296
Brian Amento *(Virginia Tech, USA)*, Loren Terveen, Will Hill *(AT&T Shannon Laboratories, USA)*

Posters.

Document classification on neural networks using only positive examples 304
Larry M. Manevitz and Malik Yousef *(University of Haifa, Israel)*

New paradigms in information visualization 307
Peter Au, Yike Guo, Stefa M. Rueger, Shalini Sewraz *(Imperial College, England)*

Latent semantic indexing model for Boolean query formulation 310
DaeHo Baek, HaeChang Rim *(Korea University, Korea)*, HeuiSeok Lim *(Chonan University, Korea)*

Generation of user profiles for information filtering research agenda 313
Tsvi Kuflik, Peretz Shoval *(University of the Negev, Israel)*

Variance based classifier comparison in text categorization 316
Atsuhiro Takasu, Kenro Aihara *(National Center for Science Information Systems, Japan)*

The use of phrases from query texts in information retrieval 318
Masumi Narita, Yasushi Ogawa *(Software Research Center, Ricoh Co., Ltd., Japan)*

Pseudo-frequency method: an efficient document ranking retrieval method for n-gram indexing 321
Yasushi Ogawa *(Software Research Center, Ricoh Co., Ltd., Japan)*

Lexical semantic relatedness and online news event detection 324
Nicola Stokes, Paula Hatch, Joe Carthy *(University College Dublin, Ireland)*

Modeling question response patterns by scaling of visualization 326
Mark E. Rorvig *(University of North Texas, USA)*

The effect of query type on subject searching behavior of image databases: an exploratory study 328
Efthimis N. Efthimiadis, Raya Fidel *(University of Washington, USA)*

The role of judge in a user based retrieval experiment 331
Mingfang Wu, Ross Wilkinson *(CSIRO – MIS, Australia)*, Michael Fuller *(Royal Melbourne Institute of Technology, Australia)*

Auto-construction of a live thesaurus from search term logs for interactive Web search 334
Shuie-Lung Juang, Wen-Hsiang Lu , Lee-Feng Chien *(Institute of Information Science, Academia Sinica, Taiwan)*, Hsiao-Tieh Pu *(Shih Hsin University, Taiwan)*

Cognitive approach for building user model in an information retrieval context 337
Amina Sayeb Belhassen, Nabil Ben Abdallah, Henda Hadjami Ben Ghezala *(PGL, ENSI Tunisia)*

Multimedia information retrieval from recorded presentations 339
Wolfgang Hurst, Rainer Muller, Christoph Mayer *(Computer Science Institute, University of Freiburg, Germany)*

Influence of speech recognition errors on topic detection 342
J. Scott McCarley, Martin Franz *(IBM TJ Watson Research Center, USA)*

Word document density & relevance scoring 345
Martin Franz, J. Scott McCarley *(IBM TJ Watson Research Center, USA)*

Ranking digital images using combination of evidence 348
Iadh Ounis *(University of Glasgow, Scotland)*

Collaborative filtering and the generalized vector space model 351
Ian Soboroff and Charles Nicholas *(University of Maryland, Baltimore County, USA)*

Theme-based retrieval of Web news 354
Nuno Maria, Mario J. Silva *(Universidadae de Lisboa, Portugal)*

Stemming and its effects on TFIDF ranking 357
Mark Kantrowitz *(Just Research, USA)*, Behrang Mohit, Vibhu Mittal *(Carnegie Mellon University, USA)*

Exploration of a heuristic approach to threshold learning in adaptive filtering 360
Chengxiang Zhai, Peter Jansen, David A. Evans *(CLARITECH Corp., USA)*

On the design and evaluation of a multi-dimensional approach to information retrieval 363
M. Catherine McCabe *(U.S. Government, USA)*, Jinho Lee, Abdur Chowdhury, David Grossman, Ophir Frieder *(Illinois Institute of Technology, USA)*

SWAMI: A framework for collaborative filtering algorithm development and evaluation 366
Danyel Fisher, Kris Hildrum, Jason Hong, Mark Newman, Megan Thomas, Richard Vuduc *(University of California, Berkeley, USA)*

Learning probabilistic models of the Web 369
Thomas Hofmann *(Brown University, USA)*

Effects of out of vocabulary words in spoken document retrieval 372
P. C. Woodland, S. E. Johnson, P. Jourlin, K. Spärck Jones *(Cambridge University, England)*

Towards an adaptive and task-specific ranking mechanism in web searching 375
Chen Ding, Chi-Hung Chi *(National University of Singapore, Singapore)*

Beyond the traditional query operators 377
Chen Ding, Chi-Hung Chi *(National University of Singapore, Singapore)*

Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems 379
Javed A. Aslam, Mark Montague *(Dartmouth College, USA)*

Information access for context-aware appliances 382
Gareth J. F. Jones, Peter J. Brown *(University of Exeter, England)*

Finding relevant passages using noun-noun compound: coherence vs. proximity 385
E. Hoenkamp, Rob de Groot *(Nijmegen Institute for Cognition and Information, The Netherlands)*

Demonstrations.

Semantic Explorer™ - Navigation in document collections;
Proxima Daily™ -Learning personal newspaper 388
Vadim Asadov, Serge Shumsky *(NeurOK, LLC, Russia)*

Integrated search tools for newspaper digital libraries 389
S.L. Mantzaris, B. Gatos, N. Gouraros, P. Tzavelis *(Lambrakis Press Archives, Greece)*

Managing photos with AT&T Shoebox 390
Timothy J. Mills, David Pye, David Sinclair, Kenneth R. Wood *(AT&T Laboratories, England)*

ClusterBook, a tool for dual information access 391
Gheorghe Muresan, David J. Harper, Ayse Göker, Peter Loweit *(Robert Gordon University, Scotland)*

Uexküll: an interactive visual user interface for document retrieval in vector space 392
Michael Preminger, Sandor Daranyi *(Oslo College, Norway)*

Time Mine: visualizing automatically constructed timelines 393
Russell Swan, James Allan *(University of Massachusetts, Amherst, USA)*

The Cambridge University multimedia document retrieval demo system 394
A. Tuerk, S.E. Johnson, P. Jourlin, K. Spärck Jones, P.C. Woodland *(Cambridge University, England)*

Author Index 395