

SIGIR '95



Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval

Edited by

Edward A. Fox, Peter Ingwersen, and Raya Fidel



Sponsored by ACM SIGIR

held in cooperation with

DD (Denmark) CEPIS-EIRSG (Europe) GI (Germany)
AICA-GLIR (Italy) IPSJ (Japan) BCS-IRSG (UK)

with support from **Microsoft** *and*

University of Washington



Computing & Communications

Seattle, Washington, USA July 9 - July 13, 1995

Special issue of the SIGIR Forum

Table of Contents



KEYNOTE ADDRESS

- Digital Libraries: Bridging the Two Cultures**1
Terry Winograd, *Stanford University*

DISTRIBUTED IR AND THE INTERNET

Chair: Paul Lindner

- Netserf: Using Semantic Knowledge to Find Internet Information Archives**4
Anil S. Chakravarthy, Kenneth B. Haase, *MIT Media Laboratory*

- Dissemination of Collection Wide Information in a Distributed
Information Retrieval System**12
Charles L. Viles, James C. French, *University of Virginia*

- Searching Distributed Collections with Inference Networks**21
James P. Callan, Zhihong Lu, W. Bruce Croft, *University of Massachusetts*

EFFICIENCY TECHNIQUES

Chair: Peter Willett

- Fast Evaluation of Structured Queries for Information Retrieval**30
Eric W. Brown, *University of Massachusetts*

- Efficient Recompression Techniques for Dynamic Full-Text Retrieval
Systems**39
Shmuel T. Klein, *Bar-Ilan University*

ADVANCED SYSTEMS

Chair: IJsbrand Jan Aalbersberg

- Design of a Reusable IR Framework**49
Gabriele Sonnenberger, Hans-Peter Frei, *Union Bank of Switzerland*

- Parallel Text Retrieval on a High Performance Super Computer Using
the Vector Space Model**58
Pavlos Efraimidis, Christos Glymidakis, Basilis Mamalis,
Paul Spirakis, Basil Tampakas, *University of Patras*

TEXT SUMMARIZATION

Chair: Karen Sparck-Jones

- A Trainable Document Summarizer**68
Julian Kupiec, Jan Pedersen, Francine Chen, *Xerox Palo Alto Research Center*

- Generating Summaries of Multiple News Articles**74
Kathleen McKeown, Dragomir R. Radev, *Columbia University*

INTEGRATING STRUCTURED AND UNSTRUCTURED INFORMATION

Chair: E. J. Yannakoudakis

- Integrating IR and RDBMS Using Cooperative Indexing**84
Samuel DeFazio, Amjad Daoud, Lisa Ann Smith, Jagannathan Srinivasan,
Oracle Corp., and W. Bruce Croft, Jamie Callan, *Univ. of Massachusetts*
- A Language for Queries on Structure and Contents of Textual Databases**93
Gonzalo Navarro, Ricardo Baeza-Yates, *University of Chile*
- An NF2 Relational Interface For Document Retrieval, Restructuring and Aggregation**102
Kalervo Jarvelin, Timo Niemi, *University of Tampere*

NATURAL LANGUAGE PROCESSING

Chair: Haruo Kimoto

- Fast and Quasi-Natural Language Search for Gigabits of Chinese Texts**112
Lee-Feng Chien, *Academia Sinica*
- A New Character-based Indexing Organization using Frequency Data for Japanese Documents**121
Yasushi Ogawa, *Ricoh Co., Ltd.*
- Little Words Can Make a Big Difference for Text Classification**130
Ellen Riloff, *University of Utah*

KEYNOTE ADDRESS

- Evaluation of Evaluation in Information Retrieval**137
Tefko Saracevic, *Rutgers University*

USER STUDIES

Chair: Phil Smith

- Searchers and Searchers: Differences between the Most and Least Consistent Searchers**149
Mirja Iivonen, *University of Tampere*
- Information Processing in the Context of Medical Care**158
Valerie Florance, *University of Rochester Medical Center*, and
Gary Marchionini, *University of Maryland*
- Towards New Measures of Information Retrieval Evaluation**164
William R. Hersh, Diane L. Elliot, David H. Hickam,
Stephanie L. Wolf, *Oregon Health Sciences University*, and
Anna Molnar, Christine Leichtenstien, *University of Ulm*

FUSION STRATEGIES

Chair: Richard Tong

- Learning Collection Fusion Strategies**172
Ellen M. Voorhees, Narendra K.Gupta, Ben Johnson-Laird,
Siemens Corporate Research, Inc.

Combining Multiple Evidence from Different Properties of Weighting Schemes	180
Joon Ho Lee, <i>Cornell University</i>	

Efficient Processing of Vague Queries using a Data Stream Approach	189
Ulrich Pfeifer, Robert Fuhr, <i>University of Dortmund</i>	

SEARCH INTERFACES

Chair: Maristella Agosti

Document Analysis for Visualization	199
David Dubin, <i>University of Pittsburgh</i>	

Users' Model of the Information Space: the case for two search models	205
Sylvia Willie, Peter Bruza, <i>Queensland University of Tech.</i>	

COGNITION AND ASSOCIATION

Chair: Rik Belew

The Newspaper Image Database: Empirical Supported Analysis of Users' Typology and Word Association Clusters	212
Susanne Ornager, <i>Royal School of Librarianship</i>	

Human Memory Models and Term Association	219
Gerda Ruge, <i>Sietec, Germany</i>	

AUTOMATIC CLASSIFICATION

Chair: Ray Larson

A Comparison of Classifiers and Document Representations for the Routing Problem	229
Hinrich Schutze and Jan Pederson, <i>Xerox Palo Alto Research Center</i> , David A. Hull, <i>Rank Xerox Research Center</i>	

A Case-Based Approach To Intelligent Information Retrieval	238
Jody J. Daniels, Edwina L. Rissland, <i>University of Massachussets</i>	

Evaluating and Optimizing Autonomous Text Classification Systems	246
David D. Lewis, <i>AT&T Bell Laboratories</i>	

TEXT CATEGORIZATION

Chair: Elizabeth D. Liddy

Noise Reduction in a Statistical Approach to Text Categorization	256
Yiming Yang, <i>Mayo Clinic</i>	

Partial Orders for Document Representation: A New Methodology for Combining Document Features	264
Steven Finch, <i>University of Edinburgh</i>	

Cluster-Based Text Categorization: A Comparison of Category Search Strategies	273
Makato Iwayama, <i>Hitachi Ltd.</i> , Takenobu Tokunaga, <i>Tokyo Inst. of Tech.</i> ,	

RETRIEVAL LOGIC

Chair: Fabrizio Sebastiani

Probabilistic Datalog - A Logic For Powerful Retrieval Methods282
Norbert Fuhr, *University of Dortmund*

Probability Kinematics in Information Retrieval291
F. Crestani, *Univ. of Padova*, and C.J. Van Rijsbergen, *Univ. of Glasgow*

An Image Retrieval Model Based on Classical Logic300
Carlo Meghini, *Istituto di Elaborazione Dell'Informazione*

TERM STATISTICS

Chair: Donna Harman

One Term Or Two?310
Kenneth W. Church, *AT&T Bell Laboratories*

Detecting Content-Bearing Words by Serial Clustering319
A. Bookstein, *University of Chicago*, USA, S.T. Klein, *Bar-Ilan Univ.*,
T. Raita, *University of Turku*

**Applying Probabilistic Term Weighting to OCR Text in the Case of a
Large Alphabetic Library Catalogue**328
Elke Mittendorf, Peter Schauble, Paraic Sheridan, *Swiss Federal Inst. of Tech.*

FEEDBACK METHODS

Chair: Howard Turtle

Relevance Feedback With Too Much Data337
James Allan, *University of Massachussetts*

On the Reuse of Past Optimal Queries344
Vijay V. Raghavan, Hayri Sever, *University of Southwestern Louisiana*

Optimization of Relevance Feedback Weights351
Chris Buckley, Gerard Salton, *Cornell University*

PANEL ON FUNDING FOR IR RESEARCH358

Chair: Efthimis Efthimiadis

Panelists:

Maria Zemankova, *National Science Foundation*
Milton Corn, *National Library of Medicine*

PANEL ON EDUCATION FOR IR358

Chair: Kazem Taghva

Panelists:

Edward Fox, *Virginia Tech*
Stephen Robertson, *City University, London*
Nicholas Belkin, *Rutgers University*
David Lewis, *AT&T Bell Laboratories*
Donna Harman, *NIST*

SYSTEMS DEMONSTRATIONS: Abstracts359

Chair: Efthimis Efthimiadis, *UCLA*

POSTERS: Abstracts369

Chair: Elizabeth D. Liddy, *Syracuse University*

POST-CONFERENCE RESEARCH WORKSHOPS377

VIRI: Visual Information Retrieval Interfaces377

Program Committee: Robert H. Korfage, Xia Lin, and David S. Dubin

Z39.50 and the IR Research Community377

Program Committee: Clifford Lynch and Ray Larson

Information Retrieval and Databases378

Program Committee: David Harper and Peter Schauble

Curriculum Development in Computer Information Science:

A Framework for Developing a New Curriculum in IR378

Program Committee: Edward A. Fox, Doris K. Lidtke, Michael C.
Mulder, Edie M. Rasmussen, and Kazem Taghva

IR and Automatic Construction of Hypermedia379

Program Committee: Maristella Agosti and James Allan