

Modern Information Retrieval

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Modern Information Retrieval

ACM PRESS BOOKS

This book is published as part of ACM Press Books – a collaboration between the Association for Computing (ACM) and Addison Wesley Longman Limited. ACM is the oldest and largest educational and scientific society in the information technology field. Through its high-quality publications and services, ACM is a major force in advancing the skills and knowledge of IT professionals throughout the world. For further information about ACM, contact:

ACM Member Services

1515 Broadway, 17th Floor

New York, NY 10036-5701

Phone: 1-212-626-0500

Fax: 1-212-944-1318

E-mail: acmhelp@acm.org

URL: <http://www.acm.org/>

ACM European Service Center

108 Cowley Road

Oxford OX4 1JF

United Kingdom

Phone: +44-1865-382388

Fax: +44-1865-381388

E-mail: acm_europe@acm.org

URL: <http://www.acm.org/>

Modern Information Retrieval

Ricardo Baeza-Yates
Berthier Ribeiro-Neto



ACM Press
New York



Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • Madrid • Mexico City • Amsterdam • Munich • Paris • Milan

Pearson Education Limited

Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsoned.co.uk

Copyright © 1999 by the ACM press, A Division of the Association for Computing Machinery, Inc. (ACM).

The rights of the authors of this Work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP.

While the publisher has made every attempt to trace all copyright owners and obtain permission to reproduce material, in a few cases this has proved impossible. Copyright holders of material which has not been acknowledged are encouraged to contact the publisher.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Addison Wesley Longman Limited has made every attempt to supply trade mark information about manufacturers and their products mentioned in this book. A list of the trademark designations and their owners appears on page viii.

Typeset in Computer Modern by 56
Printed and bound in the United States of America

First printed 1999

ISBN 0-201-39829-X

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Baeza-Yates, R. (Ricardo)

Modern information retrieval / Ricardo Baeza-Yates, Berthier Ribeiro-Neto.
p. cm.

Includes bibliographical reference and index.

ISBN 0-201-39829-X

I. Information storage and retrieval systems. I. Ribeiro, Berthier de Araújo Neto, 1960-. II. Title.

Z667.B34 1999

025.04-dc21

99-10033

CIP

10 9 8 7

07 06 05 04 03

Preface

Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out of date and this has led to the introduction of new IR books. Nevertheless, we believe that there is still great need for a book that approaches the field in a rigorous and complete way from a computer-science perspective (as opposed to a user-centered perspective). This book is an effort to partially fulfill this gap and should be useful for a first course on information retrieval as well as for a graduate course on the topic.

The book comprises two portions which complement and balance each other. The core portion includes nine chapters authored or coauthored by the designers of the book. The second portion, which is fully integrated with the first, is formed by six state-of-the-art chapters written by leading researchers in their fields. The same notation and glossary are used in all the chapters. Thus, despite the fact that several people have contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Furthermore, unlike a collection of chapters, we have carefully designed the contents and organization of the book to present a cohesive view of all the important aspects of modern information retrieval.

From IR models to indexing text, from IR visual tools and interfaces to the Web, from IR multimedia to digital libraries, the book provides both breadth of coverage and richness of detail. It is our hope that, given the now clear relevance and significance of information retrieval to modern society, the book will contribute to further disseminate the study of the discipline at information science, computer science, and library science departments throughout the world.

Ricardo Baeza-Yates, Santiago, Chile
Berthier Ribeiro-Neto, Belo Horizonte, Brazil
January, 1999

To Helena, Rosa, and our children

Amo los libros
exploradores,
libros con bosque o nieve,
profundidad o cielo

de Oda al Libro (I),

Pablo Neruda

I love books
that explore,
books with a forest or snow,
depth or sky

from Ode to the Book (I),

Pablo Neruda

território de homens livres
que será nosso país
e será pátria de todos.
Irmãos, cantai ese mundo
que não verei, mas virá
um dia, dentro de mil anos,
talvez mais. . . não tenho pressa.

de Cidade Prevista no livro
A Rosa do Povo, 1945

Carlos Drummond de Andrade

territory of free men
that will be our country
and will be the nation of all
Brothers, sing this world
which I'll not see, but which will come
one day, in a thousand years,
maybe more. . . no hurry.

from Prevised City in the book
The Rose of the People, 1945

Carlos Drummond de Andrade

Acknowledgements

We would like to deeply thank the various people who, during the several months in which this endeavor lasted, provided us with useful and helpful assistance. Without their care and consideration, this book would likely not have matured.

First, we would like to thank all the chapter contributors, for their dedication and interest. To Elisa Bertino, Eric Brown, Barbara Catania, Christos Faloutsos, Elena Ferrari, Ed Fox, Marti Hearst, Gonzalo Navarro, Edie Rasmussen, Ohm Sornil, and Nivio Ziviani, who contributed with writings that reflect expertise we certainly do not fully profess ourselves. And for all their patience throughout an editing and cross-reviewing process which constitutes a rather difficult balancing act.

Second, we would like to thank all the people who demonstrated interest in publishing this book, particularly Scott Delman and Doug Sery.

Third, we would like to commend the interest, encouragement, and great job done by Addison Wesley Longman throughout the overall process, represented by Keith Mansfield, Karen Sutherland, Bridget Allen, David Harison, Sheila Chatten, Helen Hodge and Lisa Talbot. The reviewers they contacted read an early (and rather preliminary) proposal of this book and provided us with good feedback and invaluable insights. The chapter on Parallel and Distributed IR was moved from the part on Applications of IR (where it did not fit well) to the part on Text IR due to the objective argument of an unknown referee. A separate chapter on Retrieval Evaluation was only included after another zealous referee strongly made the case for the importance of this subject.

Fourth, we would like to thank all the people who discussed this project with us. Doug Oard provided us with an early critique of the proposal. Gary Marchionini was an earlier supporter and provided us with useful contacts during the process. Bruce Croft encouraged our efforts from the beginning. Alberto Mendelzon provided us with an initial proposal and a compilation of references for the chapter on searching the Web. Ed Fox found time in a rather busy schedule to provide us with an insightful review of the introduction (which resulted in a great improvement) and a thorough review of the chapter on Modeling. Marti Hearst demonstrated interest in our proposal early on, provided assistance throughout the editing process, and has been an enthusiastic supporter and partner.

Fifth, we thank the support of our institutions, the Departments of Computer Science of the University of Chile and of the Federal University of Minas Gerais, as well as the funding provided by national research agencies (CNPq in Brazil and CONICYT in Chile) and international collaboration projects, in particular CYTED project VII.13 AMYRI (Environment for Information Managing and Retrieval in the World Wide Web) and Finep project SIAM (Information Systems for Mobile Computers) under the Pronex program.

Most important, to Helena, Rosa, and our children, who put up with a string of trips abroad, lost weekends, and odd working hours.

List of Trademarks

Alta Vista is a trademark of Compaq Computer Corporation

FrameMaker is a trademark of Adobe Systems Incorporated

IBM SP2 is a trademark of International Business Machines Corporation

Netscape Communicator is a trademark of Netscape Communications Corporation

Solaris, Sun 3/50 and Sun UltraSparc-1 are trademarks of Sun Microsystems, Inc.

Thinking Machines CM-2 is a trademark of Thinking Machines Corporation

Unix is licensed through X/Open Company Ltd

Word is a trademark of Microsoft Corporation

WordPerfect is a trademark of of Corel Corporation

Contents

Preface	v
Acknowledgements	vii
Biographies	xvii
1 Introduction	1
1.1 Motivation	1
1.1.1 Information versus Data Retrieval	1
1.1.2 Information Retrieval at the Center of the Stage	2
1.1.3 Focus of the Book	3
1.2 Basic Concepts	3
1.2.1 The User Task	4
1.2.2 Logical View of the Documents	5
1.3 Past, Present, and Future	6
1.3.1 Early Developments	6
1.3.2 Information Retrieval in the Library	7
1.3.3 The Web and Digital Libraries	7
1.3.4 Practical Issues	8
1.4 The Retrieval Process	9
1.5 Organization of the Book	10
1.5.1 Book Topics	11
1.5.2 Book Chapters	12
1.6 How to Use this Book	15
1.6.1 Teaching Suggestions	15
1.6.2 The Book's Web Page	16
1.7 Bibliographic Discussion	17
2 Modeling	19
2.1 Introduction	19
2.2 A Taxonomy of Information Retrieval Models	20
2.3 Retrieval: Ad hoc and Filtering	21

2.4	A Formal Characterization of IR Models	23
2.5	Classic Information Retrieval	24
2.5.1	Basic Concepts	24
2.5.2	Boolean Model	25
2.5.3	Vector Model	27
2.5.4	Probabilistic Model	30
2.5.5	Brief Comparison of Classic Models	34
2.6	Alternative Set Theoretic Models	34
2.6.1	Fuzzy Set Model	34
2.6.2	Extended Boolean Model	38
2.7	Alternative Algebraic Models	41
2.7.1	Generalized Vector Space Model	41
2.7.2	Latent Semantic Indexing Model	44
2.7.3	Neural Network Model	46
2.8	Alternative Probabilistic Models	48
2.8.1	Bayesian Networks	48
2.8.2	Inference Network Model	49
2.8.3	Belief Network Model	56
2.8.4	Comparison of Bayesian Network Models	59
2.8.5	Computational Costs of Bayesian Networks	60
2.8.6	The Impact of Bayesian Network Models	61
2.9	Structured Text Retrieval Models	61
2.9.1	Model Based on Non-Overlapping Lists	62
2.9.2	Model Based on Proximal Nodes	63
2.10	Models for Browsing	65
2.10.1	Flat Browsing	65
2.10.2	Structure Guided Browsing	66
2.10.3	The Hypertext Model	66
2.11	Trends and Research Issues	69
2.12	Bibliographic Discussion	69
3	Retrieval Evaluation	73
3.1	Introduction	73
3.2	Retrieval Performance Evaluation	74
3.2.1	Recall and Precision	75
3.2.2	Alternative Measures	82
3.3	Reference Collections	84
3.3.1	The TREC Collection	84
3.3.2	The CACM and ISI Collections	91
3.3.3	The Cystic Fibrosis Collection	94
3.4	Trends and Research Issues	96
3.5	Bibliographic Discussion	96
4	Query Languages	99
4.1	Introduction	99
4.2	Keyword-Based Querying	100

4.2.1	Single-Word Queries	100
4.2.2	Context Queries	101
4.2.3	Boolean Queries	102
4.2.4	Natural Language	103
4.3	Pattern Matching	104
4.4	Structural Queries	106
4.4.1	Fixed Structure	108
4.4.2	Hypertext	108
4.4.3	Hierarchical Structure	109
4.5	Query Protocols	113
4.6	Trends and Research Issues	114
4.7	Bibliographic Discussion	116
5	Query Operations	117
5.1	Introduction	117
5.2	User Relevance Feedback	118
5.2.1	Query Expansion and Term Reweighting for the Vector Model	118
5.2.2	Term Reweighting for the Probabilistic Model	120
5.2.3	A Variant of Probabilistic Term Reweighting	121
5.2.4	Evaluation of Relevance Feedback Strategies	122
5.3	Automatic Local Analysis	123
5.3.1	Query Expansion Through Local Clustering	124
5.3.2	Query Expansion Through Local Context Analysis	129
5.4	Automatic Global Analysis	131
5.4.1	Query Expansion based on a Similarity Thesaurus	131
5.4.2	Query Expansion based on a Statistical Thesaurus	134
5.5	Trends and Research Issues	137
5.6	Bibliographic Discussion	138
6	Text and Multimedia Languages and Properties	141
6.1	Introduction	141
6.2	Metadata	142
6.3	Text	144
6.3.1	Formats	144
6.3.2	Information Theory	145
6.3.3	Modeling Natural Language	145
6.3.4	Similarity Models	148
6.4	Markup Languages	149
6.4.1	SGML	149
6.4.2	HTML	152
6.4.3	XML	154
6.5	Multimedia	156
6.5.1	Formats	157
6.5.2	Textual Images	158
6.5.3	Graphics and Virtual Reality	159

6.5.4	HyTime	159
6.6	Trends and Research Issues	160
6.7	Bibliographic Discussion	162
7	Text Operations	163
7.1	Introduction	163
7.2	Document Preprocessing	165
7.2.1	Lexical Analysis of the Text	165
7.2.2	Elimination of Stopwords	167
7.2.3	Stemming	168
7.2.4	Index Terms Selection	169
7.2.5	Thesauri	170
7.3	Document Clustering	173
7.4	Text Compression	173
7.4.1	Motivation	173
7.4.2	Basic Concepts	175
7.4.3	Statistical Methods	176
7.4.4	Dictionary Methods	183
7.4.5	Inverted File Compression	184
7.5	Comparing Text Compression Techniques	186
7.6	Trends and Research Issues	188
7.7	Bibliographic Discussion	189
8	Indexing and Searching	191
8.1	Introduction	191
8.2	Inverted Files	192
8.2.1	Searching	195
8.2.2	Construction	196
8.3	Other Indices for Text	199
8.3.1	Suffix Trees and Suffix Arrays	199
8.3.2	Signature Files	205
8.4	Boolean Queries	207
8.5	Sequential Searching	209
8.5.1	Brute Force	209
8.5.2	Knuth-Morris-Pratt	210
8.5.3	Boyer-Moore Family	211
8.5.4	Shift-Or	212
8.5.5	Suffix Automaton	213
8.5.6	Practical Comparison	214
8.5.7	Phrases and Proximity	215
8.6	Pattern Matching	215
8.6.1	String Matching Allowing Errors	216
8.6.2	Regular Expressions and Extended Patterns	219
8.6.3	Pattern Matching Using Indices	220
8.7	Structural Queries	222
8.8	Compression	222

8.8.1	Sequential Searching	223
8.8.2	Compressed Indices	224
8.9	Trends and Research Issues	226
8.10	Bibliographic Discussion	227
9	Parallel and Distributed IR	229
9.1	Introduction	229
9.1.1	Parallel Computing	230
9.1.2	Performance Measures	231
9.2	Parallel IR	232
9.2.1	Introduction	232
9.2.2	MIMD Architectures	233
9.2.3	SIMD Architectures	240
9.3	Distributed IR	249
9.3.1	Introduction	249
9.3.2	Collection Partitioning	251
9.3.3	Source Selection	252
9.3.4	Query Processing	253
9.3.5	Web Issues	254
9.4	Trends and Research Issues	255
9.5	Bibliographic Discussion	256
10	User Interfaces and Visualization	257
10.1	Introduction	257
10.2	Human-Computer Interaction	258
10.2.1	Design Principles	258
10.2.2	The Role of Visualization	259
10.2.3	Evaluating Interactive Systems	261
10.3	The Information Access Process	262
10.3.1	Models of Interaction	262
10.3.2	Non-Search Parts of the Information Access Process	265
10.3.3	Earlier Interface Studies	266
10.4	Starting Points	267
10.4.1	Lists of Collections	267
10.4.2	Overviews	268
10.4.3	Examples, Dialogs, and Wizards	276
10.4.4	Automated Source Selection	278
10.5	Query Specification	278
10.5.1	Boolean Queries	279
10.5.2	From Command Lines to Forms and Menus	280
10.5.3	Faceted Queries	281
10.5.4	Graphical Approaches to Query Specification	282
10.5.5	Phrases and Proximity	286
10.5.6	Natural Language and Free Text Queries	287
10.6	Context	289
10.6.1	Document Surrogates	289

10.6.2	Query Term Hits Within Document Content	289
10.6.3	Query Term Hits Between Documents	293
10.6.4	SuperBook: Context via Table of Contents	296
10.6.5	Categories for Results Set Context	297
10.6.6	Using Hyperlinks to Organize Retrieval Results	299
10.6.7	Tables	301
10.7	Using Relevance Judgements	303
10.7.1	Interfaces for Standard Relevance Feedback	304
10.7.2	Studies of User Interaction with Relevance Feedback Systems	305
10.7.3	Fetching Relevant Information in the Background	307
10.7.4	Group Relevance Judgements	308
10.7.5	Pseudo-Relevance Feedback	308
10.8	Interface Support for the Search Process	309
10.8.1	Interfaces for String Matching	309
10.8.2	Window Management	311
10.8.3	Example Systems	312
10.8.4	Examples of Poor Use of Overlapping Windows	317
10.8.5	Retaining Search History	317
10.8.6	Integrating Scanning, Selection, and Querying	318
10.9	Trends and Research Issues	321
10.10	Bibliographic Discussion	322
11	Multimedia IR: Models and Languages	325
11.1	Introduction	325
11.2	Data Modeling	328
11.2.1	Multimedia Data Support in Commercial DBMSs	329
11.2.2	The MULTOS Data Model	331
11.3	Query Languages	334
11.3.1	Request Specification	335
11.3.2	Conditions on Multimedia Data	335
11.3.3	Uncertainty, Proximity, and Weights in Query Expressions	337
11.3.4	Some Proposals	338
11.4	Trends and Research Issues	341
11.5	Bibliographic Discussion	342
12	Multimedia IR: Indexing and Searching	345
12.1	Introduction	345
12.2	Background — Spatial Access Methods	347
12.3	A Generic Multimedia Indexing Approach	348
12.4	One-dimensional Time Series	353
12.4.1	Distance Function	353
12.4.2	Feature Extraction and Lower-bounding	353
12.4.3	Experiments	355
12.5	Two-dimensional Color Images	357

12.5.1	Image Features and Distance Functions	357
12.5.2	Lower-bounding	358
12.5.3	Experiments	360
12.6	Automatic Feature Extraction	360
12.7	Trends and Research Issues	361
12.8	Bibliographic Discussion	363
13	Searching the Web	367
13.1	Introduction	367
13.2	Challenges	368
13.3	Characterizing the Web	369
13.3.1	Measuring the Web	369
13.3.2	Modeling the Web	371
13.4	Search Engines	373
13.4.1	Centralized Architecture	373
13.4.2	Distributed Architecture	375
13.4.3	User Interfaces	377
13.4.4	Ranking	380
13.4.5	Crawling the Web	382
13.4.6	Indices	383
13.5	Browsing	384
13.5.1	Web Directories	384
13.5.2	Combining Searching with Browsing	386
13.5.3	Helpful Tools	387
13.6	Metasearchers	387
13.7	Finding the Needle in the Haystack	389
13.7.1	User Problems	389
13.7.2	Some Examples	390
13.7.3	Teaching the User	391
13.8	Searching using Hyperlinks	392
13.8.1	Web Query Languages	392
13.8.2	Dynamic Search and Software Agents	393
13.9	Trends and Research Issues	393
13.10	Bibliographic Discussion	395
14	Libraries and Bibliographical Systems	397
14.1	Introduction	397
14.2	Online IR Systems and Document Databases	398
14.2.1	Databases	399
14.2.2	Online Retrieval Systems	403
14.2.3	IR in Online Retrieval Systems	404
14.2.4	'Natural Language' Searching	406
14.3	Online Public Access Catalogs (OPACs)	407
14.3.1	OPACs and Their Content	408
14.3.2	OPACs and End Users	410
14.3.3	OPACs: Vendors and Products	410

14.3.4	Alternatives to Vendor OPACs	410
14.4	Libraries and Digital Library Projects	412
14.5	Trends and Research Issues	412
14.6	Bibliographic Discussion	413
15	Digital Libraries	415
15.1	Introduction	415
15.2	Definitions	417
15.3	Architectural Issues	418
15.4	Document Models, Representations, and Access	420
15.4.1	Multilingual Documents	420
15.4.2	Multimedia Documents	421
15.4.3	Structured Documents	421
15.4.4	Distributed Collections	422
15.4.5	Federated Search	424
15.4.6	Access	424
15.5	Prototypes, Projects, and Interfaces	425
15.5.1	International Range of Efforts	427
15.5.2	Usability	428
15.6	Standards	429
15.6.1	Protocols and Federation	429
15.6.2	Metadata	430
15.7	Trends and Research Issues	431
15.8	Bibliographical Discussion	432
	Appendix: Porter's Algorithm	433
	Glossary	437
	References	455
	Index	501

Biographies

Biographies of Main Authors

Ricardo Baeza-Yates received a bachelor degree in Computer Science in 1983 from the University of Chile. Later, he received an MSc in Computer Science (1985), a professional title in electrical engineering (1985), and an MEng in EE (1986) from the same university. He received his PhD in Computer Science from the University of Waterloo, Canada, in 1989. He has been the president of the Chilean Computer Science Society (SCCC) from 1992 to 1995 and from 1997 to 1998. During 1993, he received the Organization of the American States award for young researchers in exact sciences. Currently, he is a full professor at the Computer Science Department of the University of Chile, where he was the chairperson in the period 1993 to 1995. He is coauthor of the second edition of the *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1991; and coeditor of *Information Retrieval: Algorithms and Data Structures*, Prentice Hall, 1992. He has also contributed several papers to journals published by professional organizations such as ACM, IEEE, and SIAM.

His research interests include algorithms and data structures, text retrieval, graphical interfaces, and visualization applied to databases. He currently coordinates an IberoAmerican project on models and techniques for searching the Web financed by the Spanish agency Cyted. He has been a visiting professor or an invited speaker at several conferences and universities around the world, as well as referee for several journals, conferences, NSF, etc. He is a member of the ACM, AMS, EATCS, IEEE, SCCC, and SIAM.

Berthier Ribeiro-Neto received a bachelor degree in Math, a BS degree in Electrical Engineering, and an MS degree in Computer Science, all from the Federal University of Minas Gerais, Brazil. In 1995, he was awarded a Ph.D. in Computer Science from the University of California at Los Angeles. Since then, he has been with the Computer Science Department of the Federal University of Minas Gerais where he is an Associate Professor.

His main interests are information retrieval systems, digital libraries, interfaces for the Web, and video on demand. He has been involved in a number

of research projects financed through Brazilian national agencies such as the Ministry of Science and Technology (MCT) and the National Research Council (CNPq). From the projects currently underway, the two main ones deal with wireless information systems (project SIAM financed within program PRONEX) and video on demand (project ALMADEM financed within program PROTEM III). Dr Ribeiro-Neto is also involved with an IberoAmerican project on information systems for the Web coordinated by Professor Ricardo Baeza-Yates. He was the chair of SPIRE'98 (String Processing and Information Retrieval South American Symposium), is the chair of SBBD'99 (Brazilian Symposium on Databases), and has been on the committee of several conferences in Brazil, in South America and in the USA. He is a member of ACM, ASIS, and IEEE.

Biographies of Contributors

Elisa Bertino is Professor of Computer Science in the Department of Computer Science of the University of Milano where she heads the Database Systems Group. She has been a visiting researcher at the IBM Research Laboratory (now Almaden) in San Jose, at the Microelectronics and Computer Technology Corporation in Austin, Texas, and at Rutgers University in Newark, New Jersey. Her main research interests include object-oriented databases, distributed databases, deductive databases, multimedia databases, interoperability of heterogeneous systems, integration of artificial intelligence and database techniques, and database security. In those areas, Professor Bertino has published several papers in refereed journals, and in proceedings of international conferences and symposia. She is a coauthor of the books *Object-Oriented Database Systems — Concepts and Architectures*, Addison-Wesley 1993; *Indexing Techniques for Advanced Database Systems*, Kluwer 1997; and *Intelligent Database Systems*, Addison-Wesley forthcoming. She is or has been on the editorial boards of the following scientific journals: the *IEEE Transactions on Knowledge and Data Engineering*, the *International Journal of Theory and Practice of Object Systems*, the *Very Large Database Systems (VLDB) Journal*, the *Parallel and Distributed Database Journal*, the *Journal of Computer Security, Data & Knowledge Engineering*, and the *International Journal of Information Technology*.

Eric Brown has been a Research Staff Member at the IBM T.J. Watson Research Center in Yorktown Heights, NY, since 1995. Prior to that he was a Research Assistant at the Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst. He holds a BSc from the University of Vermont and an MS and PhD from the University of Massachusetts, Amherst. Dr. Brown conducts research in large scale information retrieval systems, automatic text categorization, and hypermedia systems for digital libraries and knowledge management. He has published a number of papers in the field of information retrieval.

Barbara Catania is a researcher at the University of Milano, Italy. She received an MS degree in Information Sciences in 1993 from the University of

Genova and a PhD in Computer Science in 1998 from the University of Milano. She has also been a visiting researcher at the European Computer-Industry Research Center, Munich, Germany. Her main research interests include multimedia databases, constraint databases, deductive databases, and indexing techniques in object-oriented and constraint databases. In those areas, Dr Catania has published several papers in refereed journals, and in proceedings of international conferences and symposia. She is also a coauthor of the book *Indexing Techniques for Advanced Database Systems*, Kluwer 1997.

Christos Faloutsos received a BSc in Electrical Engineering (1981) from the National Technical University of Athens, Greece and an MSc and PhD in Computer Science from the University of Toronto, Canada. Professor Faloutsos is currently a faculty member at Carnegie Mellon University. Prior to joining CMU he was on the faculty of the Department of Computer Science at the University of Maryland, College Park. He has spent sabbaticals at IBM-Almaden and AT&T Bell Labs. He received the Presidential Young Investigator Award from the National Science Foundation in 1989, two 'best paper' awards (SIGMOD 94, VLDB 97), and three teaching awards. He has published over 70 refereed articles and one monograph, and has filed for three patents. His research interests include physical database design, searching methods for text, geographic information systems, indexing methods for multimedia databases, and data mining.

Elena Ferrari is an Assistant Professor at the Computer Science Department of the University of Milano, Italy. She received an MS in Information Sciences in 1992 and a PhD in Computer Science in 1998 from the University of Milano. Her main research interests include multimedia databases, temporal object-oriented data models, and database security. In those areas, Dr Ferrari has published several papers in refereed journals, and in proceedings of international conferences and symposia. She has been a visiting researcher at George Mason University in Fairfax, Virginia, and at Rutgers University in Newark, New Jersey.

Dr Edward A. Fox holds a PhD and MS in Computer Science from Cornell University, and a BS from MIT. Since 1983 he has been at Virginia Polytechnic Institute and State University (Virginia Tech), where he serves as Associate Director for Research at the Computing Center, Professor of Computer Science, Director of the Digital Library Research Laboratory, and Director of the Internet Technology Innovation Center. He served as vice chair and chair of ACM SIGIR from 1987 to 1995, helped found the ACM conferences on multimedia and digital libraries, and serves on a number of editorial boards. His research is focused on digital libraries, multimedia, information retrieval, WWW/Internet, educational technologies, and related areas.

Marti Hearst is an Assistant Professor at the University of California Berkeley in the School of Information Management and Systems. From 1994 to 1997 she was a Member of the Research Staff at Xerox PARC. She received her BA, MS, and PhD degrees in Computer Science from the University of California at Berkeley. Professor Hearst's research focuses on user interfaces and robust language analysis for information access systems, and on furthering the understanding of how people use and understand such systems.

Gonzalo Navarro received his first degrees in Computer Science from ESLAI (Latin American Superior School of Informatics) in 1992 and from the University of La Plata (Argentina) in 1993. In 1995 he received his MSc in Computer Science from the University of Chile, obtaining a PhD in 1998. Between 1990 and 1993 he worked at IBM Argentina, on the development of interactive applications and on research on multimedia and hypermedia. Since 1994 he has worked in the Department of Computer Science of the University of Chile, doing research on design and analysis of algorithms, textual databases, and approximate search. He has published a number of papers and also served as referee on different journals (*Algorithmica*, *TOCS*, *TOIS*, etc.) and at conferences (SIGIR, CPM, ESA, etc.).

Edie Rasmussen is an Associate Professor in the School of Information Sciences, University of Pittsburgh. She has also held faculty appointments at institutions in Malaysia, Canada, and Singapore. Dr Rasmussen holds a BSc from the University of British Columbia and an MSc degree from McMaster University, both in Chemistry, an MLS degree from the University of Western Ontario, and a PhD in Information Studies from the University of Sheffield. Her current research interests include indexing and information retrieval in text and multimedia databases.

Ohm Sornil is currently a PhD candidate in the Department of Computer Science at Virginia Polytechnic and State University and a scholar of the Royal Thai Government. He received a BEng in Electrical Engineering from Kasetsart University, Thailand, in 1993 and an MS in Computer Science from Syracuse University in 1997. His research interests include information retrieval, digital libraries, communication networks, and hypermedia.

Nivio Ziviani is a Professor of Computer Science at the Federal University of Minas Gerais in Brazil, where he heads the laboratory for Treating Information. He received a BS in Mechanical Engineering from the Federal University of Minas Gerais in 1971, an MSc in Informatics from the Catholic University of Rio in 1976, and a PhD in Computer Science from the University of Waterloo, Canada, in 1982. He has obtained several research funds from the Brazilian Research Council (CNPq), Brazilian Agencies CAPES and FINEP, Spanish Agency CYTED (project AMYRI), and private institutions. He currently coordinates a four year project on Web and wireless information systems (called SIAM) financed by the Brazilian Ministry of Science and Technology. He is co-founder of the Miner Technology Group, owner of the Miner Family of agents to search the Web. He is the author of several papers in journals and conference proceedings covering topics in the areas of algorithms and data structures, information retrieval, text indexing, text searching, text compression, and related areas. Since January of 1998, he is the editor of the 'News from Latin America' section in the Bulletin of the European Association for Theoretical Computer Science. He has been chair and member of the program committee of several conferences and is a member of ACM, EATICS and SBC.

Chapter 1

Introduction

1.1 Motivation

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested. Unfortunately, characterization of the *user information need* is not a simple problem. Consider, for instance, the following hypothetical user information need in the context of the World Wide Web (or just the Web):

Find all the pages (documents) containing information on college tennis teams which: (1) are maintained by an university in the USA and (2) participate in the NCAA tennis tournament. To be relevant, the page must include information on the national ranking of the team in the last three years and the email or phone number of the team coach.

Clearly, this full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a *query* which can be processed by the search engine (or IR system).

In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of *information* as opposed to the retrieval of *data*.

1.1.1 Information versus Data Retrieval

Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. In fact, the user of an IR system is concerned more with retrieving *information* about a

subject than with retrieving data which satisfies a given query. A data retrieval language aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression. Thus, for a data retrieval system, a single erroneous object among a thousand retrieved objects means total failure. For an information retrieval system, however, the retrieved objects might be inaccurate and small errors are likely to go unnoticed. The main reason for this difference is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. On the other hand, a data retrieval system (such as a relational database) deals with data that has a well defined structure and semantics.

Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic. To be effective in its attempt to satisfy the user information need, the IR system must somehow ‘interpret’ the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This ‘interpretation’ of a document content involves extracting syntactic and semantic information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. Thus, the notion of *relevance* is at the center of information retrieval. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.

1.1.2 Information Retrieval at the Center of the Stage

In the past 20 years, the area of information retrieval has grown well beyond its primary goals of indexing text and searching for useful documents in a collection. Nowadays, research in IR includes modeling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc. Despite its maturity, until recently, IR was seen as a narrow area of interest mainly to librarians and information experts. Such a tendentious vision prevailed for many years, despite the rapid dissemination, among users of modern personal computers, of IR tools for multimedia and hypertext applications. In the beginning of the 1990s, a single fact changed once and for all these perceptions — the introduction of the World Wide Web.

The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. Its success is based on the conception of a standard user interface which is always the same no matter what computational environment is used to run the interface. As a result, the user is shielded from details of communication protocols, machine location, and operating systems. Further, any user can create his own Web documents and make them point to any other Web documents without restrictions. This is a key aspect because it turns the Web into a new publishing medium accessible to everybody. As an immediate

consequence, any Web user can push his personal agenda with little effort and almost at no cost. This universe without frontiers has attracted tremendous attention from millions of people everywhere since the very beginning. Furthermore, it is causing a revolution in the way people use computers and perform their daily tasks. For instance, home shopping and home banking are becoming very popular and have generated several hundred million dollars in revenues.

Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy his information need, the user might navigate the space of Web links (i.e., the *hyperspace*) searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient. For naive users, the problem becomes harder, which might entirely frustrate all their efforts. The main obstacle is the absence of a well defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality. These difficulties have attracted renewed interest in IR and its techniques as promising solutions. As a result, almost overnight, IR has gained a place with other technologies at the center of the stage.

1.1.3 Focus of the Book

Despite the great increase in interest in information retrieval, modern textbooks on IR with a broad (and extensive) coverage of the various topics in the field are still difficult to find. In an attempt to partially fulfill this gap, this book presents an overall view of research in IR from a computer scientist's perspective. This means that the focus of the book is on computer algorithms and techniques used in information retrieval systems. A rather distinct viewpoint is taken by librarians and information science researchers, who adopt a human-centered interpretation of the IR problem. In this interpretation, the focus is on trying to understand how people interpret and use information as opposed to how to structure, store, and retrieve information automatically. While most of this book is dedicated to the computer scientist's viewpoint of the IR problem, the human-centered viewpoint is discussed to some extent in the last two chapters.

We put great emphasis on the integration of the different areas which are closely related to the information retrieval problem and thus, should be treated together. For that reason, besides covering text retrieval, library systems, user interfaces, and the Web, this book also discusses visualization, multimedia retrieval, and digital libraries.

1.2 Basic Concepts

The effective retrieval of relevant information is directly affected both by the *user task* and by the *logical view of the documents* adopted by the retrieval system, as we now discuss.

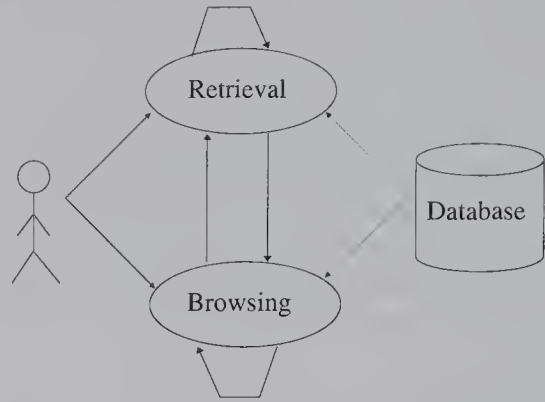


Figure 1.1 Interaction of the user with the retrieval system through distinct tasks.

1.2.1 The User Task

The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need. With a data retrieval system, a query expression (such as, for instance, a regular expression) is used to convey the constraints that must be satisfied by objects in the answer set. In both cases, we say that the user searches for useful information executing a *retrieval* task.

Consider now a user who has an interest which is either poorly defined or which is inherently broad. For instance, the user might be interested in documents about car racing in general. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to car racing. For instance, he might find interesting documents about Formula 1 racing, about car manufacturers, or about the ‘24 Hours of Le Mans.’ Furthermore, while reading about the ‘24 Hours of Le Mans’, he might turn his attention to a document which provides directions to Le Mans and, from there, to documents which cover tourism in France. In this situation, we say that the user is *browsing* the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.

In this book, we make a clear distinction between the different tasks the user of the retrieval system might be engaged in. His task might be of two distinct types: information or data *retrieval* and *browsing*. Classic information retrieval systems normally allow information or data retrieval. Hypertext systems are usually tuned for providing quick browsing. Modern digital library and Web interfaces might attempt to combine these tasks to provide improved retrieval capabilities. However, combination of retrieval and browsing is not yet a well

established approach and is not the dominant paradigm.

Figure 1.1 illustrates the interaction of the user through the different tasks we identify. Information and data retrieval are usually provided by most modern information retrieval systems (such as Web interfaces). Further, such systems might also provide some (still limited) form of browsing. While combining information and data retrieval with browsing is not yet a common practice, it might become so in the future.

Both retrieval and browsing are, in the language of the World Wide Web, ‘pulling’ actions. That is, the user requests the information in an interactive manner. An alternative is to do retrieval in an automatic and permanent fashion using software agents which *push* the information towards the user. For instance, information useful to a user could be extracted periodically from a news service. In this case, we say that the IR system is executing a particular retrieval task which consists of *filtering* relevant information for later inspection by the user. We briefly discuss filtering in Chapter 2.

1.2.2 Logical View of the Documents

Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a *logical view of the document*. For a precise definition of the concept of a document and its characteristics, see Chapter 6.

Modern computers are making it possible to represent a document by its full set of words. In this case, we say that the retrieval system adopts a *full text* logical view (or representation) of the documents. With very large collections, however, even modern computers might have to reduce the set of representative keywords. This can be accomplished through the elimination of *stopwords* (such as articles and connectives), the use of *stemming* (which reduces distinct words to their common grammatical root), and the identification of noun groups (which eliminates adjectives, adverbs, and verbs). Further, compression might be employed. These operations are called *text operations* (or transformations) and are covered in detail in Chapter 7. Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of *index terms*.

The full text is clearly the most complete logical view of a document but its usage usually implies higher computational costs. A small set of categories (generated by a human specialist) provides the most concise logical view of a document but its usage might lead to retrieval of poor quality. Several intermediate logical views (of a document) might be adopted by an information retrieval system as illustrated in Figure 1.2. Besides adopting any of the intermediate representations, the retrieval system might also recognize the internal structure normally present in a document (e.g., chapters, sections, subsections, etc.). This

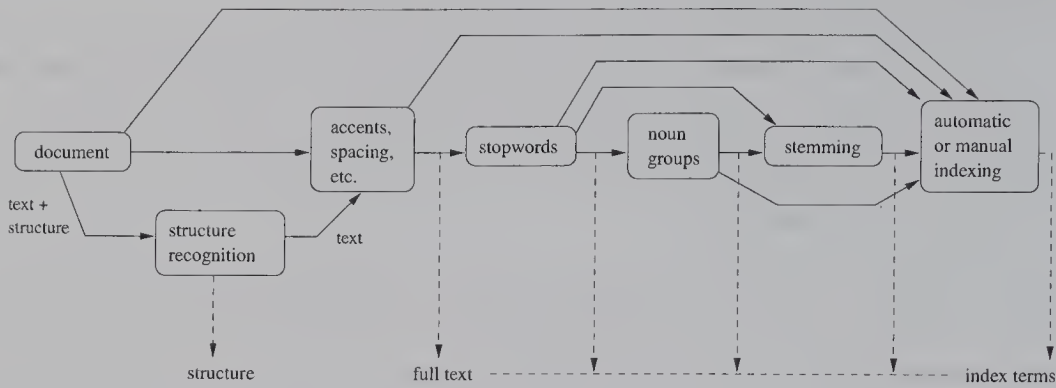


Figure 1.2 Logical view of a document: from full text to a set of index terms.

information on the structure of the document might be quite useful and is required by structured text retrieval models such as those discussed in Chapter 2.

As illustrated in Figure 1.2, we view the issue of logically representing a document as a continuum in which the logical view of a document might shift (smoothly) from a full text representation to a higher level representation specified by a human subject.

1.3 Past, Present, and Future

1.3.1 Early Developments

For approximately 4000 years, man has organized information for later retrieval and usage. A typical example is the table of contents of a book. Since the volume of information eventually grew beyond a few books, it became necessary to build specialized data structures to ensure faster access to the stored information. An old and popular data structure for faster information retrieval is a collection of selected words or concepts with which are associated pointers to the related information (or documents) — the *index*. In one form or another, indexes are at the core of every modern information retrieval system. They provide faster access to the data and allow the query processing task to be speeded up. A detailed coverage of indexes and their usage for searching can be found in Chapter 8.

For centuries, indexes were created manually as categorization hierarchies. In fact, most libraries still use some form of categorical hierarchy to classify their volumes (or documents), as discussed in Chapter 14. Such hierarchies have usually been conceived by human subjects from the library sciences field. More recently, the advent of modern computers has made possible the construction of large indexes automatically. Automatic indexes provide a view of the retrieval problem which is much more related to the system itself than to the user need.

In this respect, it is important to distinguish between two different views of the IR problem: a computer-centered one and a human-centered one.

In the computer-centered view, the IR problem consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms which improve the ‘quality’ of the answer set. In the human-centered view, the IR problem consists mainly of studying the behavior of the user, of understanding his main needs, and of determining how such understanding affects the organization and operation of the retrieval system. According to this view, keyword based query processing might be seen as a strategy which is unlikely to yield a good solution to the information retrieval problem in the long run.

In this book, we focus mainly on the computer-centered view of the IR problem because it continues to be dominant in the market place.

1.3.2 Information Retrieval in the Library

Libraries were among the first institutions to adopt IR systems for retrieving information. Usually, systems to be used in libraries were initially developed by academic institutions and later by commercial vendors. In the first generation, such systems consisted basically of an automation of previous technologies (such as card catalogs) and basically allowed searches based on author name and title. In the second generation, increased search functionality was added which allowed searching by subject headings, by keywords, and some more complex query facilities. In the third generation, which is currently being deployed, the focus is on improved graphical interfaces, electronic forms, hypertext features, and open system architectures.

Traditional library management system vendors include Endeavor Information Systems Inc., Innovative Interfaces Inc., and EOS International. Among systems developed with a research focus and used in academic libraries, we distinguish Okapi (at City University, London), MELVYL (at University of California), and Cheshire II (at UC Berkeley). Further details on these library systems can be found in Chapter 14.

1.3.3 The Web and Digital Libraries

If we consider the search engines on the Web today, we conclude that they continue to use indexes which are very similar to those used by librarians a century ago. What has changed then?

Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and the boom of the Web. First, it became a lot cheaper to have access to various sources of information. This allows reaching a wider audience than ever possible before. Second, the advances in all kinds of digital communication provided greater access to networks. This implies that the information source is available even if distantly located and that

the access can be done quickly (frequently, in a few seconds). Third, the freedom to post whatever information someone judges useful has greatly contributed to the popularity of the Web. For the first time in history, many people have free access to a large publishing medium.

Fundamentally, low cost, greater access, and publishing freedom have allowed people to use the Web (and modern digital libraries) as a highly *interactive* medium. Such interactivity allows people to exchange messages, photos, documents, software, videos, and to ‘chat’ in a convenient and low cost fashion. Further, people can do it at the time of their preference (for instance, you can buy a book late at night) which further improves the convenience of the service. Thus, high interactivity is the fundamental and current shift in the communication paradigm. Searching the Web is covered in Chapter 13, while digital libraries are covered in Chapter 15.

In the future, three main questions need to be addressed. First, despite the high interactivity, people still find it difficult (if not impossible) to retrieve information relevant to their information needs. Thus, in the dynamic world of the Web and of large digital libraries, which techniques will allow retrieval of higher quality? Second, with the ever increasing demand for access, quick response is becoming more and more a pressing factor. Thus, which techniques will yield faster indexes and smaller query response times? Third, the quality of the retrieval task is greatly affected by the user interaction with the system. Thus, how will a better understanding of the user behavior affect the design and deployment of new information retrieval strategies?

1.3.4 Practical Issues

Electronic commerce is a major trend on the Web nowadays and one which has benefited millions of people. In an electronic transaction, the buyer usually has to submit to the vendor some form of credit information which can be used for charging for the product or service. In its most common form, such information consists of a credit card number. However, since transmitting credit card numbers over the Internet is not a safe procedure, such data is usually transmitted over a fax line. This implies that, at least in the beginning, the transaction between a new user and a vendor requires executing an off-line procedure of several steps before the actual transaction can take place. This situation can be improved if the data is encrypted for security. In fact, some institutions and companies already provide some form of encryption or automatic authentication for security reasons.

However, security is not the only concern. Another issue of major interest is privacy. Frequently, people are willing to exchange information as long as it does not become public. The reasons are many but the most common one is to protect oneself against misuse of private information by third parties. Thus, privacy is another issue which affects the deployment of the Web and which has not been properly addressed yet.

Two other very important issues are copyright and patent rights. It is far

from clear how the wide spread of data on the Web affects copyright and patent laws in the various countries. This is important because it affects the business of building up and deploying large digital libraries. For instance, is a site which supervises all the information it posts acting as a publisher? And if so, is it responsible for a misuse of the information it posts (even if it is not the source)?

Additionally, other practical issues of interest include scanning, optical character recognition (OCR), and cross-language retrieval (in which the query is in one language but the documents retrieved are in another language). In this book, however, we do not cover practical issues in detail because it is not our main focus. The reader interested in details of practical issues is referred to the interesting book by Lesk [501].

1.4 The Retrieval Process

At this point, we are ready to detail our view of the retrieval process. Such a process is interpreted in terms of component subprocesses whose study yields many of the chapters in this book.

To describe the retrieval process, we use a simple and generic software architecture as shown in Figure 1.3. First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following: (a) the documents to be used, (b) the operations to be performed on the text, and (c) the text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them.

Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the *inverted file* as indicated in Figure 1.3. The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times.

Given that the document database is indexed, the retrieval process can be initiated. The user first specifies a *user need* which is then parsed and transformed by the same text operations applied to the text. Then, *query operations* might be applied before the actual *query*, which provides a system representation for the user need, is generated. The query is then processed to obtain the *retrieved documents*. Fast query processing is made possible by the index structure previously built.

Before been sent to the user, the retrieved documents are ranked according to a *likelihood* of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a *user feedback* cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation

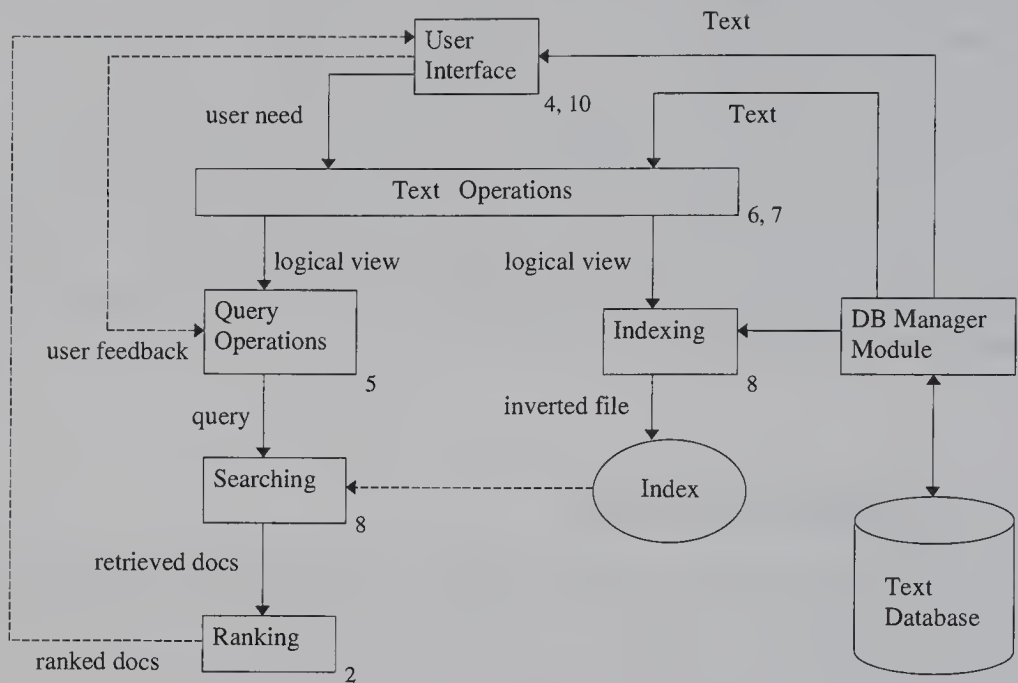


Figure 1.3 The process of retrieving information (the numbers beside each box indicate the chapters that cover the corresponding topic).

of the real user need.

The small numbers outside the lower right corner of various boxes in Figure 1.3 indicate the chapters in this book which discuss the respective subprocesses in detail. A brief introduction to each of these chapters can be found in section 1.5.

Consider now the user interfaces available with current information retrieval systems (including Web search engines and Web browsers). We first notice that the user almost never declares his information need. Instead, he is required to provide a direct representation for the query that the system will execute. Since most users have no knowledge of text and query operations, the query they provide is frequently inadequate. Therefore, it is not surprising to observe that poorly formulated queries lead to poor retrieval (as happens so often on the Web).

1.5 Organization of the Book

For ease of comprehension, this book has a straightforward structure in which four main parts are distinguished: text IR, human-computer interaction (HCI)

for IR, multimedia IR, and applications of IR. *Text IR* discusses the classic problem of searching a collection of documents for useful information. *HCI for IR* discusses current trends in IR towards improved user interfaces and better data visualization tools. *Multimedia IR* discusses how to index document images and other binary data by extracting features from their content and how to search them efficiently. On the other hand, document images that are predominantly text (rather than pictures) are called *textual images* and are amenable to automatic extraction of keywords through metadescriptors, and can be retrieved using text IR techniques. *Applications of IR* covers modern applications of IR such as the Web, bibliographic systems, and digital libraries. Each part is divided into topics which we now discuss.

1.5.1 Book Topics

The four parts which compose this book are subdivided into eight topics as illustrated in Figure 1.4. These eight topics are as follows.

The topic *Retrieval Models & Evaluation* discusses the traditional models of searching text for useful information and the procedures for evaluating an information retrieval system. The topic *Improvements on Retrieval* discusses techniques for transforming the query and the text of the documents with the aim of improving retrieval. The topic *Efficient Processing* discusses indexing and searching approaches for speeding up the retrieval. These three topics compose the first part on Text IR.

The topic *Interfaces & Visualization* covers the interaction of the user with the information retrieval system. The focus is on interfaces which facilitate the process of specifying a query and provide a good visualization of the results.

The topic *Multimedia Modeling & Searching* discusses the utilization of multimedia data with information retrieval systems. The focus is on modeling, indexing, and searching multimedia data such as voice, images, and other binary data.

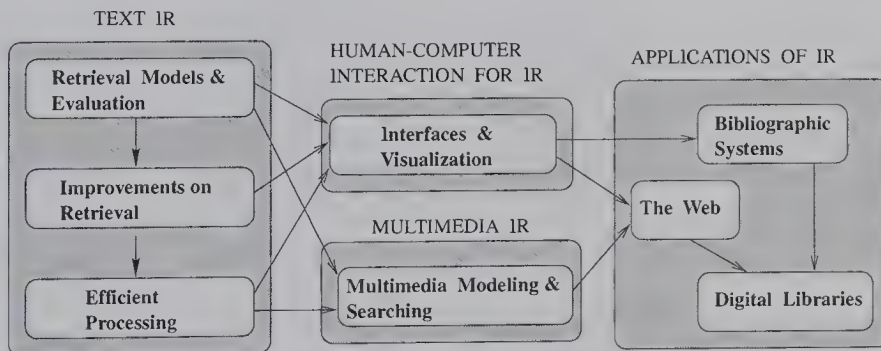


Figure 1.4 Topics which compose the book and their relationships.

The part on applications of IR is composed of three interrelated topics: *The Web*, *Bibliographic Systems*, and *Digital Libraries*. Techniques developed for the first two applications support the deployment of the latter.

The eight topics distinguished above generate the 14 chapters, besides this introduction, which compose this book and which we now briefly introduce.

1.5.2 Book Chapters

Figure 1.5 illustrates the overall structure of this book. The reasoning which yielded the chapters from 2 to 15 is as follows.

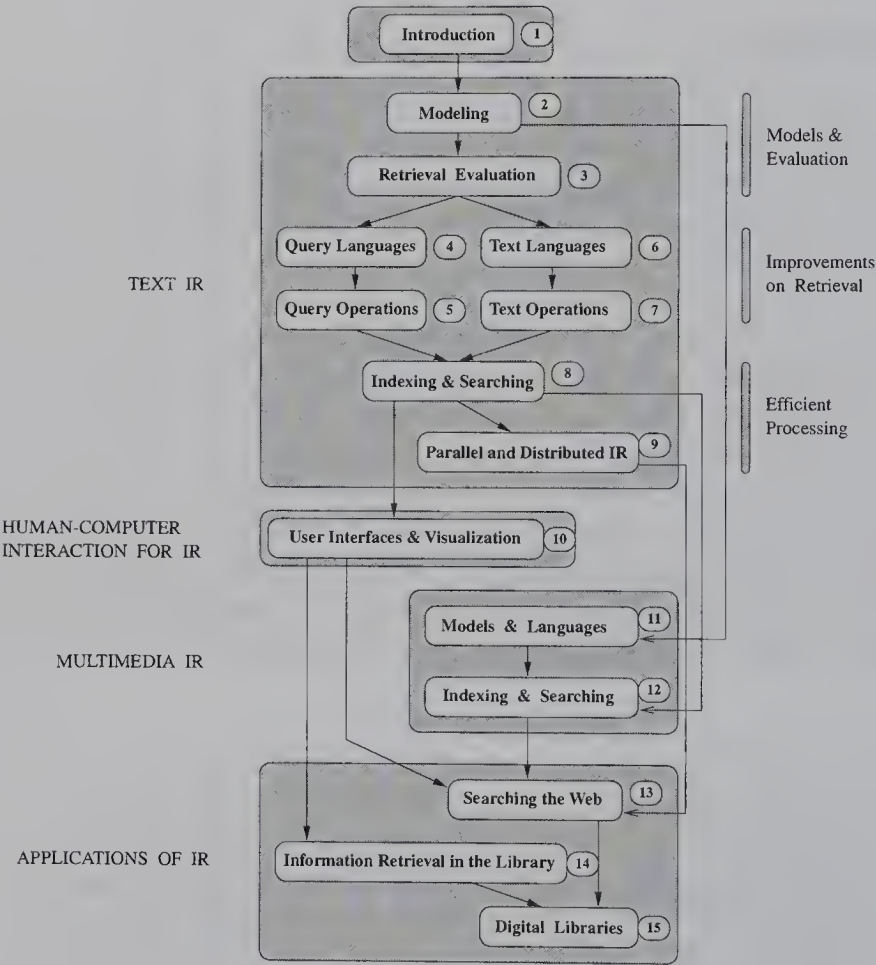


Figure 1.5 Structure of the book.

In the traditional keyword-based approach, the user specifies his information need by providing sets of keywords and the information system retrieves the documents which best approximate the user query. Also, the information system

might attempt to rank the retrieved documents using some measure of relevance. This ranking task is critical in the process of attempting to satisfy the user information need and is the main goal of *modeling* in IR. Thus, information retrieval models are discussed early in Chapter 2. The discussion introduces many of the fundamental concepts in information retrieval and lays down much of the foundation for the subsequent chapters. Our coverage is detailed and broad. Classic models (Boolean, vector, and probabilistic), modern probabilistic variants (belief network models), alternative paradigms (extended Boolean, generalized vector, latent semantic indexing, neural networks, and fuzzy retrieval), structured text retrieval, and models for browsing (hypertext) are all carefully introduced and explained.

Once a new retrieval algorithm (maybe based on a new retrieval model) is conceived, it is necessary to evaluate its performance. Traditional evaluation strategies usually attempt to estimate the costs of the new algorithm in terms of time and space. With an information retrieval system, however, there is the additional issue of evaluating the relevance of the documents retrieved. For this purpose, text reference collections and evaluation procedures based on variables other than time and space are used. Chapter 3 is dedicated to the discussion of *retrieval evaluation*.

In traditional IR, queries are normally expressed as a set of keywords which is quite convenient because the approach is simple and easy to implement. However, the simplicity of the approach prevents the formulation of more elaborate querying tasks. For instance, queries which refer to both the structure and the content of the text cannot be formulated. To overcome this deficiency, more sophisticated query languages are required. Chapter 4 discusses various types of *query languages*. Since now the user might refer to the structure of a document in his query, this structure has to be defined. This is done by embedding the description of a document content and of its structure in a text language such as the Standard Generalized Markup Language (SGML). As illustrated in Figure 1.5, Chapter 6 is dedicated to the discussion of *text languages*.

Retrieval based on keywords might be of fairly low quality. Two possible reasons are as follows. First, the user query might be composed of too few terms which usually implies that the query context is poorly characterized. This is frequently the case, for instance, in the Web. This problem is dealt with through transformations in the query such as query expansion and user relevance feedback. Such *query operations* are covered in Chapter 5. Second, the set of keywords generated for a given document might fail to summarize its semantic content properly. This problem is dealt with through transformations in the text such as identification of noun groups to be used as keywords, stemming, and the use of a thesaurus. Additionally, for reasons of efficiency, text compression can be employed. Chapter 7 is dedicated to *text operations*.

Given the user query, the information system has to retrieve the documents which are related to that query. The potentially large size of the document collection (e.g., the Web is composed of millions of documents) implies that specialized indexing techniques must be used if efficient retrieval is to be achieved. Thus, to speed up the task of matching documents to queries, proper *indexing* and *search-*

ing techniques are used as discussed in Chapter 8. Additionally, query processing can be further accelerated through the adoption of *parallel and distributed IR* techniques as discussed in Chapter 9.

As illustrated in Figure 1.5, all the key issues regarding Text IR, from modeling to fast query processing, are covered in this book.

Modern user interfaces implement strategies which assist the user to form a query. The main objective is to allow him to define more precisely the context associated to his information need. The importance of query contextualization is a consequence of the difficulty normally faced by users during the querying process. Consider, for instance, the problem of quickly finding useful information in the Web. Navigation in hyperspace is not a good solution due to the absence of a logical and semantically well defined structure (the Web has no underlying logical model). A popular approach for specifying a user query in the Web consists of providing a set of keywords which are searched for. Unfortunately, the number of terms provided by a common user is small (typically, fewer than four) which usually implies that the query is vague. This means that new user interface paradigms which assist the user with the query formation process are required. Further, since a vague user query usually retrieves hundreds of documents, the conventional approach of displaying these documents as items of a scrolling list is clearly inadequate. To deal with this problem, new data visualization paradigms have been proposed in recent years. The main trend is towards visualization of a large subset of the retrieved documents at once and direct manipulation of the whole subset. *User interfaces* for assisting the user to form his query and current approaches for *visualization* of large data sets are covered in Chapter 10.

Following this, we discuss the application of IR techniques to multimedia data. The key issue is how to model, index, and search structured documents which contain multimedia objects such as digitized voice, images, and other binary data. *Models and query languages* for office and medical information retrieval systems are covered in Chapter 11. *Efficient indexing and searching* of multimedia objects is covered in Chapter 12. Some readers may argue that the models and techniques for multimedia retrieval are rather different from those for classic text retrieval. However, we take into account that images and text are usually together and that with the Web, other media types (such as video and audio) can also be mixed in. Therefore, we believe that in the future, all the above will be treated in a unified and consistent manner. Our book is a first step in that direction.

The final three chapters of the book are dedicated to applications of modern information retrieval: the Web, bibliographic systems, and digital libraries. As illustrated in Figure 1.5, Chapter 13 presents the Web and discusses the main problems related to the issue of searching the Web for useful information. Also, our discussion covers briefly the most popular search engines in the Web presenting particularities of their organization. Chapter 14 covers commercial document databases and online public access catalogs. Commercial document databases are still the largest information retrieval systems nowadays. LEXIS-NEXIS, for instance, has a database with 1.3 billion documents and attends to over 120 million query requests annually. Finally, Chapter 15 discusses modern digital

libraries. Architectural issues, models, prototypes, and standards are all covered. The discussion also introduces the '5S' model (streams, structures, spaces, scenarios and societies) as a framework for providing theoretical and practical unification of digital libraries.

1.6 How to Use this Book

Although several people have contributed chapters for this book, it is really a textbook. The contents and the structure of the book have been carefully designed by the two main authors who also authored or coauthored nine of the 15 chapters in the book. Further, all the contributed chapters have been judiciously edited and integrated into a unifying framework that provides uniformity in structure and style, a common glossary, a common bibliography, and appropriate cross-references. At the end of each chapter, a discussion on research issues, trends, and selected bibliography is included. This discussion should be useful for graduate students as well as for researchers. Furthermore, the book is complemented by a Web page with additional information and resources.

1.6.1 Teaching Suggestions

This textbook can be used in many different areas including computer science (CS), information systems, and library science. The following list gives suggested contents for different courses at the undergraduate and graduate level, based on syllabuses of many universities around the world:

- **Information Retrieval** (Computer Science, undergraduate): this is the standard course for many CS programs. The minimum content should include Chapters 1 to 8 and Chapter 10, that is, most of the part on Text IR complemented with the chapter on user interfaces. Some specific topics of those chapters, such as more advanced models for IR and sophisticated algorithms for indexing and searching, can be omitted to fit a one term course. The chapters on Applications of IR can be mentioned briefly at the end.
- **Advanced Information Retrieval** (Computer Science, graduate): similar to the previous course but with more detailed coverage of the various chapters particularly modeling and searching (assuming the previous course as a requirement). In addition, Chapter 9 and Chapters 13 to 15 should be covered completely. Emphasis on research problems and new results is a must.
- **Information Retrieval** (Information Systems, undergraduate): this course is similar to the CS course, but with a different emphasis. It should include Chapters 1 to 7 and Chapter 10. Some notions from Chapter 8 are

useful but not crucial. At the end, the system-oriented parts of the chapters on Applications of IR, in particular those on Bibliographic Systems and Digital Libraries, must be covered (this material can be complemented with topics from [501]).

- **Information Retrieval** (Library Science, undergraduate): similar to the previous course, but removing the more technical and advanced material of Chapters 2, 5, and 7. Also, greater emphasis should be put on the chapters on Bibliographic Systems and Digital Libraries. The course should be complemented with a thorough discussion of the user-centered view of the IR problem (for example, using the book by Allen [13]).
- **Multimedia Retrieval** (Computer Science, undergraduate or graduate): this course should include Chapters 1 to 3, 6, and 11 to 15. The emphasis could be on multimedia itself or on the integration of classical IR with multimedia. The course can be complemented with one of the many books on this topic, which are usually more broad and technical.
- **Topics in IR** (Computer Science, graduate): many chapters of the book can be used for this course. It can emphasize modeling and evaluation or user interfaces and visualization. It can also be focused on algorithms and data structures (in that case, [275] and [825] are good complements). A multimedia focus is also possible, starting with Chapters 11 and 12 and using more specific books later on.
- **Topics in IR** (Information Systems or Library Science, graduate) similar to the above but with emphasis on non-technical parts. For example, the course could cover modeling and evaluation, query languages, user interfaces, and visualization. The chapters on applications can also be considered.
- **Web Retrieval and Information Access** (generic, undergraduate or graduate): this course should emphasize hypertext, concepts coming from networks and distributed systems and multimedia. The kernel should be the basic models of Chapter 2 followed by Chapters 3, 4, and 6. Also, Chapters 11 and 13 to 15 should be discussed.
- **Digital Libraries** (generic, undergraduate or graduate): This course could start with part of Chapters 2 to 4 and 6, followed by Chapters 10, 14, and 15. The kernel of the course could be based on the book by Lesk [501].

More bibliography useful for many of the courses above is discussed in the last section of this chapter.

1.6.2 The Book's Web Page

As IR is a very dynamic area nowadays, a book by itself is not enough. For that reason (and many others), the book has a Web home page located and mirrored in the following places (mirrors in USA and Europe are also planned):

- Brazil: <http://www.dcc.ufmg.br/irbook>
- Chile: <http://sunsite.dcc.uchile.cl/irbook>

Comments, suggestions, contributions, or mistakes found are welcome through email to the contact authors given on the Web page.

The Web page contains the Table of Contents, Preface, Acknowledgements, Introduction, Glossary, and other appendices to the book. It also includes exercises and teaching materials that will be increasing in volume and changing with time. In addition, a reference collection (containing 1239 documents on Cystic Fibrosis and 100 information requests with extensive relevance evaluation [721]) is available for experimental purposes. Furthermore, the page includes useful pointers to IR syllabuses in different universities, IR research groups, IR publications, and other resources related to IR and this book. Finally, any new important results or additions to the book as well as an errata will be made publicly available there.

1.7 Bibliographic Discussion

Many other books have been written on information retrieval, and due to the current widespread interest in the subject, new books have appeared recently. In the following, we briefly compare our book with these previously published works.

Classic references in the field of information retrieval are the books by van Rijsbergen [785] and Salton and McGill [698]. Our distinction between data and information retrieval is borrowed from the former. Our definition of the information retrieval process is influenced by the latter. However, almost 20 years later, both books are now outdated and do not cover many of the new developments in information retrieval.

Three more recent and also well known references in information retrieval are the book edited by Frakes and Baeza-Yates [275], the book by Witten, Moffat, and Bell [825], and the book by Lesk [501]. All these three books are complementary to this book. The first is focused on data structures and algorithms for information retrieval and is useful whenever quick prototyping of a known algorithm is desired. The second is focused on indexing and compression, and covers images besides text. For instance, our definition of a textual image is borrowed from it. The third is focused on digital libraries and practical issues such as history, distribution, usability, economics, and property rights. On the issue of computer-centered and user-centered retrieval, a generic book on information systems that takes the latter view is due to Allen [13].

There are other complementary books for specific chapters. For example, there are many books on IR and hypertext. The same is true for generic or specific multimedia retrieval, as images, audio or video. Although not an information retrieval title, the book by Rosenfeld and Morville [682] on information architecture of the Web, is a good complement to our chapter on searching the

Web. The book by Menasce and Almeida [554] demonstrates how to use queueing theory for predicting Web server performance. In addition, there are many books that explain how to find information on the Web and how to use search engines.

The reference edited by Sparck Jones and Willet [414], which was long awaited, is really a collection of papers rather than an edited book. The coherence and breadth of coverage in our book makes it more appropriate as a textbook in a formal discipline. Nevertheless, this collection is a valuable research tool. A collection of papers on cross-language information retrieval was recently edited by Grefenstette [323]. This book is a good complement to ours for people interested in this particular topic. Additionally, a collection focused on intelligent IR was edited recently by Maybury [550], and another collection on natural language IR edited by Strzalkowski will appear soon [748].

The book by Korfhage [451] covers a lot less material and its coverage is not as detailed as ours. For instance, it includes no detailed discussion of digital libraries, the Web, multimedia, or parallel processing. Similarly, the books by Kowalski [459] and Shapiro *et al.* [719] do not cover these topics in detail, and have a different orientation. Finally, the recent book by Grossman and Frieder [326] does not discuss the Web, digital libraries, or visual interfaces.

For people interested in research results, the main journals on IR are: *Journal of the American Society of Information Sciences* (JASIS) published by Wiley and Sons, *ACM Transactions on Information Systems*, *Information Processing & Management* (IP&M, Elsevier), *Information Systems* (Elsevier), *Information Retrieval* (Kluwer), and *Knowledge and Information Systems* (Springer). The main conferences are: ACM SIGIR International Conference on Information Retrieval, ACM International Conference on Digital Libraries (ACM DL), ACM Conference on Information Knowledge and Management (CIKM), and Text REtrieval Conference (TREC). Regarding events of regional influence, we would like to acknowledge the SPIRE (South American Symposium on String Processing and Information Retrieval) symposium.

Chapter 13

Searching the Web

13.1 Introduction

The World Wide Web dates from the end of the 1980s [85] and no one could have imagined its current impact. The boom in the use of the Web and its exponential growth are now well known. Just the amount of textual data available is estimated to be in the order of one terabyte. In addition, other media, such as images, audio, and video, are also available. Thus, the Web can be seen as a very large, unstructured but ubiquitous database. This triggers the need for efficient tools to manage, retrieve, and filter information from this database. This problem is also becoming important in large intranets, where we want to extract or infer new information to support a decision process, a task called data mining. As mentioned in Chapter 1, we make the important distinction between data and information retrieval. We are interested in the latter case, in which the user searches for data that fulfills his information need.

We focus on text, because although there are techniques to search for images and other non-textual data, they cannot be applied (yet) on a large scale. We also emphasize syntactic search. That is, we search for Web documents that have user-specified words or patterns in their text. As discussed in Chapter 2, such words or patterns may or may not reflect the intrinsic semantics of the text. An alternative approach to syntactic search is to do a natural language analysis of the text. Although the techniques to preprocess natural language and extract the text semantics are not new, they are not yet very effective and they are also too costly for large amounts of data. In addition, in most cases they are only effective with well structured text, a thesaurus, and other contextual information.

There are basically three different forms of searching the Web. Two of them are well known and are frequently used. The first is to use search engines that index a portion of the Web documents as a full-text database. The second is to use Web directories, which classify selected Web documents by subject. The third and not yet fully available, is to search the Web exploiting its hyperlink†

† We will use hyperlink or link to denote a pointer (anchor) from a Web page to another Web page.

structure. We cover all three forms of Web search here.

We first discuss the challenges of searching the Web, followed by some Web statistics and models which can be used to understand the complexity of the problem. Next, we discuss in detail the main tools used today to search the Web. The discussion includes search engines, Web directories, hybrid systems, user interfaces, and searching examples. We continue with new query languages that exploit the graphical structure of the Web. Finally, we survey current trends and research issues. As Web research is a very dynamic field, we may have missed some important work, for which we apologize in advance.

13.2 Challenges

We now mention the main problems posed by the Web. We can divide them in two classes: problems with the data itself and problems regarding the user and his interaction with the retrieval system. The problems related to the data are:

- **Distributed data:** due to the intrinsic nature of the Web, data spans over many computers and platforms. These computers are interconnected with no predefined topology and the available bandwidth and reliability on the network interconnections varies widely.
- **High percentage of volatile data:** due to Internet dynamics, new computers and data can be added or removed easily (it is estimated that 40% of the Web changes every month [424]). We also have dangling links and relocation problems when domain or file names change or disappear.
- **Large volume:** the exponential growth of the Web poses scaling issues that are difficult to cope with.
- **Unstructured and redundant data:** most people say that the Web is a distributed hypertext. However, this is not exactly so. Any hypertext has a conceptual model behind it, which organizes and adds consistency to the data and the hyperlinks. That is hardly true in the Web, even for individual documents. In addition, each HTML page is not well structured and some people use the term *semi-structured data*. Moreover, much Web data is repeated (mirrored or copied) or very similar. Approximately 30% of Web pages are (near) duplicates [120, 723]. Semantic redundancy can be even larger.
- **Quality of data:** the Web can be considered as a new publishing medium. However, there is, in most cases, no editorial process. So, data can be false, invalid (for example, because it is too old), poorly written or, typically, with many errors from different sources (typos, grammatical mistakes, OCR errors, etc.). Preliminary studies show that the number of words with typos can range from 1 in 200 for common words to 1 in 3 for foreign surnames [588].

- **Heterogeneous data:** in addition to having to deal with multiple media types and hence with multiple formats, we also have different languages and, what is worse, different alphabets, some of them very large (for example, Chinese or Japanese Kanji).

Most of these problems (such as the variety of data types and poor data quality) are not solvable simply by software improvements. In fact, many of them will not change (and they should not, as in the case of language diversity!) because they are problems (also features) intrinsic to human nature.

The second class of problems are those faced by the user during the interaction with the retrieval system. There are basically two problems: (1) how to specify a query and (2) how to interpret the answer provided by the system. Without taking into account the semantic content of a document, it is not easy to precisely specify a query, unless it is very simple. Further, even if the user is able to pose the query, the answer might be a thousand Web pages. How do we handle a large answer? How do we rank the documents? How do we select the documents that really are of interest to the user? In addition, a single document could be large. How do we browse efficiently in large documents?

So, the overall challenge, in spite of the intrinsic problems posed by the Web, is to submit a good query to the search system, and obtain a manageable and relevant answer. Moreover, in practice we should try to achieve the latter goal even for poorly formulated queries. In the rest of this chapter, we use the term Web pages for HTML documents (HTML is described in Chapter 6). To denote all possible data types available on the Web, we use the term Web documents.

13.3 Characterizing the Web

13.3.1 Measuring the Web

Measuring the Internet and in particular the Web, is a difficult task due to its highly dynamic nature. Nowadays, there are more than 40 million computers in more than 200 countries connected to the Internet, many of them hosting Web servers. The estimated number of Web servers ranges from 2.4 million according to NetSizer [597] (November 1998) to over three million according to the Netcraft Web survey [596] (October 1998). This wide range might be explained when we consider that there are many Web sites that share the same Web server using virtual hosts, that not all of them are fully accessible, that many of them are provisional, etc. Other estimations were made by sampling 0.1% of all Internet numeric addresses obtaining about 2 million unique Web sites [619] or by counting domain names starting with `www` which in July 1998 were 780,000 according to the Internet Domain survey [599]. However, since not all Web servers have this prefix, the real number is even higher. Considering that in July 1998 the number of Internet hosts was estimated at 36.7 million [599], there is about one Web server per every ten computers connected to the

Internet. The characterization of the Web is a new task of the Web Consortium [797].

In two interesting articles, already (sadly) outdated, Bray [114] and Woodruff *et al.* [834] studied different statistical measures of the Web. The first study uses 11 million pages while the second uses 2.6 million pages, with both sets gathered in November 1995. Their characterization of Web pages is partially reproduced in the following paragraphs. A first question is how many different institutions (not Web servers) maintain Web data. This number is smaller than the number of servers, because many places have multiple servers. The exact number is unknown, but should be more than 40% of the number of Web servers (this percentage was the value back in 1995). The exact number of Web pages is also not known. Estimates at the beginning of 1998 ranged from 200 to 320 million, with 350 million as the best current estimate (July 1998 [91]). The latter study used 20,000 random queries based on a lexicon of 400,000 words extracted from Yahoo!. Those queries were submitted to four search engines and the union of all the answers covered about 70% of the Web. Figure 13.1 gives an approximation of how the number of Web servers and the number of pages have changed in recent years. Between 1997 and 1998, the size of the Web doubled in nine months and is currently growing at a rate of 20 million pages per month. On the other hand, it is estimated that the 30,000 largest Web sites (about 1% of the Web) account for approximately 50% of all Web pages [619].

The most popular formats for Web documents are HTML, followed by GIF and JPG (both for images), ASCII text, and Postscript, in that order. The most popular compression tools used are GNU zip, Zip, and Compress. What is a typical HTML page? First, most HTML pages are not standard, meaning that they do not comply with all the HTML specifications. In ad-

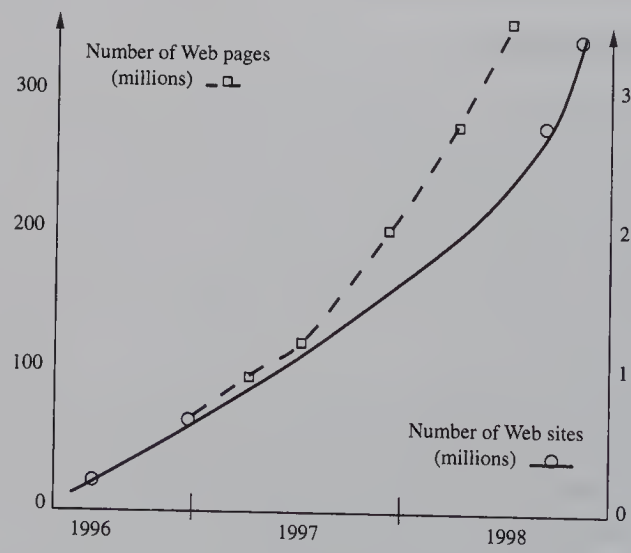


Figure 13.1 Approximate growth of the Web.

dition, although HTML is an instance of SGML, HTML documents seldom start with a formal document type definition. Second, they are small (around 5 Kbs on average with a median of 2 Kbs) and usually contain few images (between one and two on average with an average size of 14 Kb). The pages that have images use them for presentation issues such as colored bullets and lines. An average page has between five and 15 hyperlinks (more than eight links on average) and most of them are local (that is, they point to pages in their own Web server hierarchy). On average, no external server points to any given page (typically, there are only local links pointing to a given page). This is true even for home pages of Web sites. In fact, in 1995, around 80% of these home pages had fewer than ten external links pointing to each of them.

The top ten most referenced sites are Microsoft, Netscape, Yahoo!, and top US universities. In these cases we are talking about sites which are referenced by at least 100,000 places. On the other hand, the site with most links to outside sites is Yahoo!. In some sense, Yahoo! and other directories are the glue of the Web. Without them we would have many isolated portions (which is the case with many personal Web pages). If we assume that the average HTML page has 5 Kb and that there are 300 million Web pages, we have at least 1.5 terabytes of text. This is consistent with other measures obtained from search engines. Note that this volume does not include non-textual documents.

Regarding the languages used in Web pages, there have been three studies made. The first study was done by Funredes [637] from 1996 to 1998. It uses the AltaVista search engine and is based on searching different words in different languages. This technique might not be significant statistically, but the results are consistent with the second study which was carried out by Alis Technology [11] and is based on automatic software that can detect the language used. One of the goals of the study was to test such software (done in 8000 Web servers). The last study was done by OCLC in June of 1998 [619] by sampling Internet numeric addresses and using the SILC language identification software. Table 13.1 gives the percentages of Web pages written in each language (with the exception of the OCLC data that counts Web sites), as well as the number of people (millions) who speak the language. The variations for Japanese might be due to an inability to detect pages written in Kanji. Some languages, in particular Spanish and Portuguese, are growing fast and will surpass French in the near future. The total number of languages exceeds 100.

13.3.2 Modeling the Web

Can we model the document characteristics of the whole Web? Yes, as has already been discussed partially in Chapter 6. The Heaps' and Zipf's laws are also valid in the Web. In particular, the vocabulary grows faster (larger β) and the word distribution should be more biased (larger θ). However, there are no experiments on large Web collections to measure these parameters.

Language	Funredes (1998, %)	Alis Tech. (June 1997, %)	OCLC (June 1998, %)	Spoken by (millions)
English	76.4	82.3	71	450
Japanese	4.8	1.6	4	126
German	4.4	4.0	7	118
French	2.9	1.5	3	122
Spanish	2.6	1.1	3	266
Italian	1.5	0.8	1	63
Portuguese	0.8	0.7	2	175

Table 13.1 Languages of the Web.

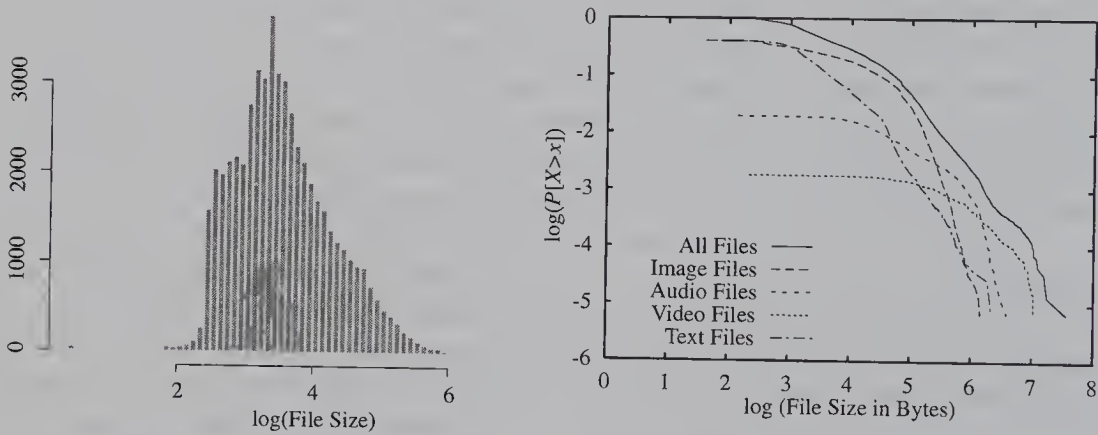


Figure 13.2 Left: Distribution for all file sizes (courtesy of M. Crovella, 1998). Right: Right tail distribution for different file types (from Crovella and Bestavros, 1996). All logarithms are in base 10.

An additional model is related to the distribution of document sizes. According to this model, the document sizes are self-similar [201], that is, they have a large variance (a similar behavior appears in Web traffic). This can be modeled by two different distributions. The main body of the distribution follows a logarithmic normal distribution, such that the probability of finding a document of size x bytes is given by

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp -(\ln x - \mu)^2/2\sigma^2$$

where the average (μ) and standard deviation (σ) are 9.357 and 1.318, respectively [59]. Figure 13.2 (left) shows the size distribution of the experimental data.

The right tail of the distribution is ‘heavy-tailed.’ That is, the majority of documents are small, but there is a non-trivial number of large documents. This is intuitive for image or video files, but it is also true for HTML pages. A good fit is obtained with the Pareto distribution

$$p(x) = \frac{\alpha k^\alpha}{x^{1+\alpha}}$$

where x is measured in bytes and k and α are parameters of the distribution [59] (see Figure 13.2 (right)). For text files, α is about 1.36, being smaller for images and other binary formats [201, 819]. Taking all Web documents into account, we get $\alpha = 1.1$ and $k = 9.3$ Kb [58]. That is, 9.3 Kb is the cut point between both distributions, and 93% of all the files have a size below this value. In fact, for less than 50 Kb, images are the typical files, from 50 to 300 Kb we have an increasing number of audio files, and over that to several megabytes, video files are more frequent. The parameters of these distributions were obtained from a sample of more than 54,000 Web pages requested by several users in a period of two months of 1995. Recent data collected in 1998 show that the size distributions have the same form, but parameters change [58]. Related information can be found on Web benchmarks such as WebSpec96 and the Sun/Inktomi Inkbench [395].

13.4 Search Engines

In this section we cover different architectures of retrieval systems that model the Web as a full-text database. One main difference between standard IR systems and the Web is that, in the Web, all queries must be answered without accessing the text (that is, only the indices are available). Otherwise, that would require either storing locally a copy of the Web pages (too expensive) or accessing remote pages through the network at query time (too slow). This difference has an impact on the indexing and searching algorithms, as well as on the query languages made available.

13.4.1 Centralized Architecture

Most search engines use a centralized crawler-indexer architecture. Crawlers are programs (software agents) that traverse the Web sending new or updated pages to a main server where they are indexed. Crawlers are also called robots, spiders, wanderers, walkers, and knowbots. In spite of their name, a crawler does not actually move to and run on remote machines, rather the crawler runs on a local system and sends requests to remote Web servers. The index is used in a centralized fashion to answer queries submitted from different places in the Web. Figure 13.3 shows the software architecture of a search engine based on the AltaVista architecture [17]. It has two parts: one that deals with the users,

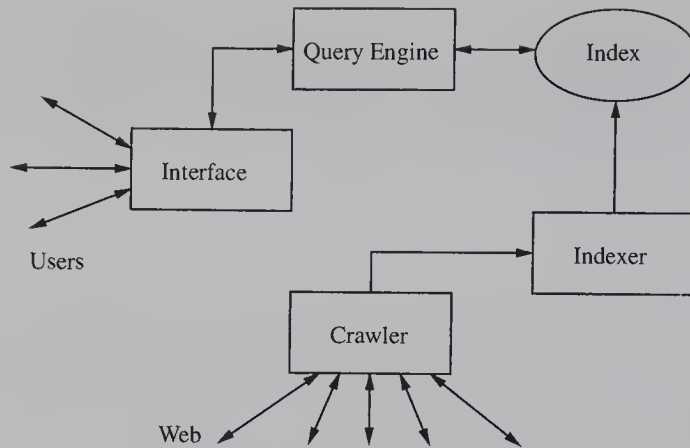


Figure 13.3 Typical crawler-indexer architecture.

consisting of the user interface and the query engine and another that consists of the crawler and indexer modules. In 1998, the overall AltaVista system was running on 20 multi-processor machines, all of them having more than 130 Gb of RAM and over 500 Gb of disk space. Only the query engine uses more than 75% of these resources.

The main problem faced by this architecture is the gathering of the data, because of the highly dynamic nature of the Web, the saturated communication links, and the high load at Web servers. Another important problem is the volume of the data. In fact, the crawler-indexer architecture may not be able to cope with Web growth in the near future. Particularly important is good load balancing between the different activities of a search engine, internally (answering queries and indexing) and externally (crawling).

The largest search engines, considering Web coverage in June 1998, were AltaVista [17], HotBot [380], Northern Light [608], and Excite [240], in that order. According to recent studies, these engines cover 28–55% [749] or 14–34% [490] of all Web pages, whose number was estimated at over 300 million in 1998. Table 13.2 lists the most important search engines and their estimated sizes along with their corresponding URLs. Beware that some search engines are powered by the same internal engine. For example, HotBot, GoTo, and Microsoft are powered by Inktomi [395] and Magellan by Excite's internal engine. Up to date information can be found in [749, 609].

Most search engines are based in the United States and focus on documents in English. Nevertheless, there are search engines specialized in different countries and/or languages, which are able, for instance, to query and retrieve documents written in Kanji (Chinese, Japanese, and Korean). Also there are search engines that take other approaches, like Ask Jeeves! which simulates an interview [34] or DirectHit [215] which ranks the Web pages in the answer in order of their popularity. We should also mention those search engines aimed at specific topics, for example the Search Broker [537] which allows us to search in many specific topics and DejaNews [212] which searches the USENET archives.

Search engine	URL	Web pages indexed
AltaVista	www.altavista.com	140
AOL Netfind	www.aol.com/netfind/	–
Excite	www.excite.com	55
Google	google.stanford.edu	25
GoTo	goto.com	–
HotBot	www.hotbot.com	110
Infoseek	www.infoseek.com	30
Lycos	www.lycos.com	30
Magellan	www.mckinley.com	55
Microsoft	search.msn.com	–
NorthernLight	www.nlsearch.com	67
WebCrawler	www.webcrawler.com	2

Table 13.2 URLs and estimated size (millions) of the largest search engines (May 1998).

There are also engines to retrieve specific Web pages such as personal or institutional home pages or specific objects such as electronic mail addresses, images, or software applets.

13.4.2 Distributed Architecture

There are several variants of the crawler-indexer architecture. Among them, the most important is Harvest [108]. Harvest uses a distributed architecture to gather and distribute data, which is more efficient than the crawler architecture. The main drawback is that Harvest requires the coordination of several Web servers.

The Harvest distributed approach addresses several of the problems of the crawler-indexer architecture, such as: (1) Web servers receive requests from different crawlers, increasing their load; (2) Web traffic increases because crawlers retrieve entire objects, but most of their content is discarded; and (3) information is gathered independently by each crawler, without coordination between all the search engines.

To solve these problems, Harvest introduces two main elements: gatherers and brokers. A gatherer collects and extracts indexing information from one or more Web servers. Gathering times are defined by the system and are periodic (i.e. there are harvesting times as the name of the system suggests). A broker provides the indexing mechanism and the query interface to the data gathered. Brokers retrieve information from one or more gatherers or other brokers, updating incrementally their indices. Depending on the configuration of gatherers and brokers, different improvements on server load and network traffic can be

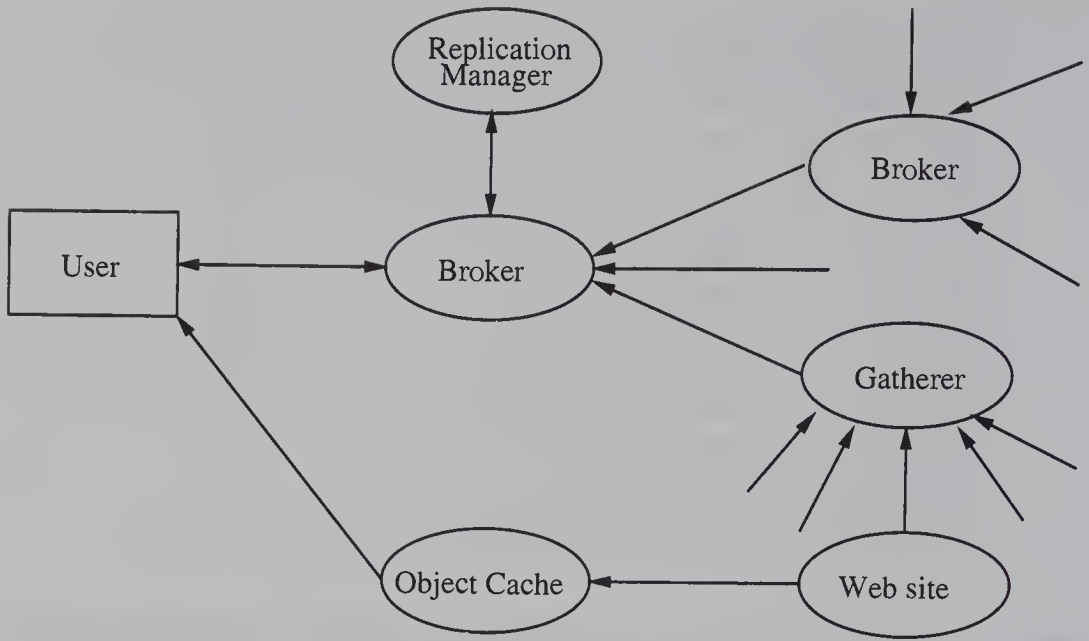


Figure 13.4 Harvest architecture.

achieved. For example, a gatherer can run on a Web server, generating no external traffic for that server. Also, a gatherer can send information to several brokers, avoiding work repetition. Brokers can also filter information and send it to other brokers. This design allows the sharing of work and information in a very flexible and generic manner. An example of the Harvest architecture is shown in Figure 13.4 [108].

One of the goals of Harvest is to build topic-specific brokers, focusing the index contents and avoiding many of the vocabulary and scaling problems of generic indices. Harvest includes a distinguished broker that allows other brokers to register information about gatherers and brokers. This is useful to search for an appropriate broker or gatherer when building a new system. The Harvest architecture also provides replicators and object caches. A replicator can be used to replicate servers, enhancing user-base scalability. For example, the registration broker can be replicated in different geographic regions to allow faster access. Replication can also be used to divide the gathering process between many Web servers. Finally, the object cache reduces network and server load, as well as response latency when accessing Web pages. More details on the system can be found in [108].

Currently, there are hundreds of Harvest applications on the Web (for example, the CIA, NASA, the US National Academy of Sciences, and the US Government Printing Office), as this software is on the public domain.‡ Netscape's Catalog Server is a commercial version of Harvest and Network Appliances' cache is a commercial version of the Harvest Cache.

‡ Information is available at harvest.transarc.com.

Enter ranking keywords in any language [Simple Search](#)

Enter boolean expression:

Range of dates:

From:

To:

e.g.: 21/Mar/96

☐ Count documents matching the boolean expression.

To take advantage of advanced search features, please consult the Help section.

Figure 13.5 Query interface for complex queries in AltaVista.

13.4.3 User Interfaces

There are two important aspects of the user interface of search engines: the query interface and the answer interface (see also Chapter 10). The basic query interface is a box where one or more words can be typed. Although a user would expect that a given sequence of words represents the same query in all search engines, it does not. For example, in AltaVista a sequence of words is a reference to the union of all the Web pages having at least one of those words, while in HotBot it is a reference to the Web pages having all the words. Another problem is that the logical view of the text is not known, that is, some search engines use stopwords, some do stemming, and some are not case sensitive (see Chapter 7).

All search engines also provide a query interface for complex queries as well as a command language including Boolean operators and other features, such as phrase search, proximity search, and wild cards. Figures 13.5 and 13.6 show the query interfaces for complex queries for the three largest search engines. They provide several filtering functions. The results can be filtered by additional words that must be present or absent from the answer or in a particular field such as the URL or title, language, geographic region or Internet domain, date range, or inclusion of specific data types such as images or audio.

The answer usually consists of a list of the ten top ranked Web pages. Figure 13.7 shows the three top documents for the main four search engines for the query `searching and Web and engine`. Each entry in this list includes some information about the document it represents. Typically, the information includes the URL, size, the date when the page was indexed, and a couple of lines with its content (title plus first lines or selected headings or sentences). Some search engines allow the user to change the number of pages returned in the list and the amount of information per page, but in most cases this is fixed or limited to a few choices. The order of the list is typically by relevance, but sorting by URL or date is also available in some engines. In addition, most search engines also have an option to find documents similar to each Web page in the answer.

 the #1 rated search engine

[Return to *Power Options*](#)

Look For
Search for pages containing the search term(s) in the form specified.

Language
Limits results to a specific language.

Word Filter
Limit results to pages containing/not containing the words specified.

[more terms](#)

Date
Limit results to pages published within a specified period of time.

☒ After , 19

☐ or on

Pages Must include
Return only pages containing the specified media types or technologies.

☐ image ☐ audio ☐ video ☐ Shockwave
☐ Java ☐ JavaScript ☐ ActiveX ☐ VRML
☐ Acrobat ☐ VB Script
☐ extension: (gif, txt,...)

Location/Domain
Return only pages in specific domains (wired.com, dol.gov), top-level domains (.edu, .com), and/or specific continents or countries

☒ Continent

☒ Domain
 (.com, .edu) website: (wired.com, etc.)
country code: (uk, fr, jp)

(See the domain and country code index for more detailed info.)

Page Depth
Control what type of pages are searched within each Web site

☒ Any Page
☒ Top Page
☒ Personal Page
☒ Page Depth

Word Stemming
Searches for grammatic variations of your search term. Ex: Searches for "thought" will also find "think" and "thinking".

☐ Enable Word Stemming

Return Results
Choose the number of results to be displayed and the length of each description

Words anywhere:

Words in title:

Publication name: See our [list o...](#)

Words in URL:

Select: ☒ Special Collection
☒ World Wide Web
☒ All Sources

➔ **SELECT DATE RANGE**
Fill in one date field or both to narrow your results by date.

Start date: End date: mm/dd/yy

☐ Sort results by date

➔ **SELECT SOURCES** ☐ All Sources below

☐ Journals & Magazines ☐ Non-profit web sites
☐ News archives ☐ Educational web sites
☐ Personal pages ☐ Military web sites
☐ Commercial web sites ☐ Government web sites

Documents written in:

Web sites from:

➔ **SELECT SUBJECTS** ☐ All Subjects

☐ Arts ☐ Gov't, Law & Politics ☐ Social sciences
☐ Business & Investing ☐ Health & Medicine ☐ Sports & Recreation
☐ Computing & Internet ☐ Humanities ☐ Technology
☐ Contemporary life ☐ Products & Services ☐ Travel
☐ Education ☐ Reference
☐ Entertainment ☐ Science & Mathematics

Figure 13.6 Query interface for complex queries for HotBot (left) and NorthernLight (right).

► AltaVista found 3,156,580 Web pages for you. [Refine your search](#)

1. **Welcome to PCfriend USA Searching Engine Web Site**

URL: www.pcfriend.net/menu1.htm

Last modified 23-Feb-98 - page size 826 bytes - in English [[Translate](#)]

2. **Searching Engine**

Home| TYP Databank| TYP Homepages| Net Trade Center| Fair News| Leading Firms. Business & Finance Database| New Media Database| World Trade Promotion...

URL: top1.twm-online.com.tw/search/index.htm

Last modified 22-Sep-98 - page size 4K - in English [[Translate](#)]

3. **Searching Engine**

Welcome] ~ [Contact] ~ [Map] ~ [Search] Searching Engine - Here are some popl complete. substring. Infoseek. the...

URL: violet.tele.pitt.edu/search.html

Last modified 23-Jun-97 - page size 12K - in English [[Translate](#)]

Web Matches: 49,690

1 - 10 [next](#) ►

Get the **Top 10 Most Visited Sites for "Searching Web Engine"**

1. **Sybil Web**

welcome to sybilweb overview | about sybilweb | site map | search | help | contact Search Tips Answers to Frequently Asked Questions (FAQ) Search Corner Web Compass Canada Sybil's Search Engine overview sybilweb, the Web component of Sybil's Search

99% <http://www.sybilweb.com/>

See results from [this site only](#).

2. **Welcome to Complatinos S.A., Costa Rica, Webdesign, Maintenance, Computers**

Aha, you have found the Complatinos S.A., Costa Rica, Design, Submitting, Hosting, Maintenance, Translation, Links, Logo

99% <http://www.complatinos.com/>

See results from [this site only](#).

3. **Yu Search Engine - Yu Internet Pretrazivac**

Srpski info Add URL Add E-mail Business Open Site Daily News Guide YuSearch Promo: Advertising Web Hosting Click! Web Search Enter keywords for searching Yu Web: E-mail Search Enter keywords for searching E-mail Web Index Arts

97% <http://www.yusearch.com/>

See results from [this site only](#).

Power Search found 113,731 items for:

☐ Special Collection ☐ World Wide Web ☐ All Sources

 Documents that best match your search

1. **Internet Search Mechanisms**

79% - **Directories & Lists:** Internet Search Mechanisms Internet Search Mechanisms Harold Goldstein - dcbieter@goldray... - Visit the Goldpages See Fossilized Insects, get your beading supplies and help save the... Date Not Available

WWW

Commercial site: <http://goldray.95.net/searches.htm>

2. **Internet Search Mechanisms**

79% - **Directories & Lists:** Internet Search Mechanisms Internet Search Mechanisms Harold Goldstein - dcbieter@goldray... - Visit the Goldpages See Fossilized Insects, get your beading supplies and help save the... Date Not Available

WWW

Commercial site: <http://goldray.com/searches.htm>

3. **NetVet Web Searching Web Picks**

79% - **Directories & Lists:** NetVetWeb Searching Web Picks Search Tools This Site Other Veterinary WWW Search Forms Other Search Engines Search NetVet and the Electronic Zoo! Other

WWW

Veterinary. 01/07/98

Educational site: <http://netvet.wustl.edu/search.htm>

Top 10 matches. [12760 hits About Your Results]

[Show Titles only](#) [List by Web site](#)

74% **W3 Search Engines** - This documents collects some of the most useful search engines available on the WWW. Omissions are the fault of the maintainer. Suggestions for additions are welcome! Some interesting information sources are available only through specialized software.
<http://cuwww.unige.ch/meta-index.html>
[Search for more documents like this one](#)

73% **webtaxi.com : the search engine and database navigation interface/guid...** - webtaxi.com is a breakthrough navigation service designed to help Internet users conveniently search the World Wide Web. webtaxi.com enhances the existing capabilities of current versions of Netscape Navigator (2.0 and higher). This free service was developed to offer efficient point and click access to search engines, newsgroups and thousands of hard-to-reach databases. webtaxi.com provides...
<http://www.webtaxi.com/>
[Search for more documents like this one](#)

71% **Free Software from AOL and PLS** - The industry's leading search software products are now free! nbsp; PLS's powerful search engine and products, accompanied by complete documentation, are available for download from this Web site free of charge. Check it out. And check back frequently for updates on product and service offerings.
<http://www.pls.com/>

Figure 13.7 Output for the query **searching** and **Web** and **engine** for the four main search engines; from top to bottom: AltaVista, HotBot, NorthernLight, and Excite.

The user can also refine the query by constructing more complex queries based on the previous answer.

The Web pages retrieved by the search engine in response to a user query are ranked, usually using statistics related to the terms in the query. In some cases this may not have any meaning, because relevance is not fully correlated with statistics about term occurrence within the collection. Some search engines also taking into account terms included in metatags or the title, or the popularity of a Web page to improve the ranking. This topic is covered next.

13.4.4 Ranking

Most search engines use variations of the Boolean or vector model (see Chapter 2) to do ranking. As with searching, ranking has to be performed without accessing the text, just the index. There is not much public information about the specific ranking algorithms used by current search engines. Further, it is difficult to compare fairly different search engines given their differences, and continuous improvements. More important, it is almost impossible to measure recall, as the number of relevant pages can be quite large for simple queries. Some inconclusive studies include [327, 498].

Yuwono and Lee [844] propose three ranking algorithms in addition to the classical tf-idf scheme (see Chapter 2). They are called Boolean spread, vector spread, and most-cited. The first two are the normal ranking algorithms of the Boolean and vector model extended to include pages pointed to by a page in the answer or pages that point to a page in the answer. The third, most-cited, is based only on the terms included in pages having a link to the pages in the answer. A comparison of these techniques considering 56 queries over a collection of 2400 Web pages indicates that the vector model yields a better recall-precision curve, with an average precision of 75%.

Some of the new ranking algorithms also use hyperlink information. This is an important difference between the Web and normal IR databases. The number of hyperlinks that point to a page provides a measure of its popularity and quality. Also, many links in common between pages or pages referenced by the same page often indicates a relationship between those pages. We now present three examples of ranking techniques that exploit these facts, but they differ in that two of them depend on the query and the last does not.

The first is WebQuery [148], which also allows visual browsing of Web pages. WebQuery takes a set of Web pages (for example, the answer to a query) and ranks them based on how connected each Web page is. Additionally, it extends the set by finding Web pages that are highly connected to the original set. A related approach is presented by Li [512].

A better idea is due to Kleinberg [444] and used in HITS (Hypertext Induced Topic Search). This ranking scheme depends on the query and considers the set of pages S that point to or are pointed by pages in the answer. Pages that have many links pointing to them in S are called authorities (that is, they should have relevant content). Pages that have many outgoing links are called hubs (they should point to similar content). A positive two-way feedback exists:

better authority pages come from incoming edges from good hubs and better hub pages come from outgoing edges to good authorities. Let $H(p)$ and $A(p)$ be the hub and authority value of page p . These values are defined such that the following equations are satisfied for all pages p :

$$H(p) = \sum_{u \in S \mid p \rightarrow u} A(u), \quad A(p) = \sum_{v \in S \mid v \rightarrow p} H(v)$$

where $H(p)$ and $A(p)$ for all pages are normalized (in the original paper, the sum of the squares of each measure is set to one). These values can be determined through an iterative algorithm, and they converge to the principal eigenvector of the link matrix of S . In the case of the Web, to avoid an explosion of the size of S , a maximal number of pages pointing to the answer can be defined. This technique does not work with non-existent, repeated, or automatically generated links. One solution is to weight each link based on the surrounding content. A second problem is that the topic of the result can become diffused. For example, a particular query is enlarged by a more general topic that contains the original answer. One solution to this problem is to analyze the content of each page and assign a score to it, as in traditional IR ranking. The link weight and the page score can be included on the previous formula multiplying each term of the summation [154, 93, 153]. Experiments show that the recall and precision on the first ten answers increases significantly [93]. The order of the links can also be used by dividing the links into subgroups and using the HITS algorithm on those subgroups instead of the original Web pages [153].

The last example is PageRank, which is part of the ranking algorithm used by Google [117]. PageRank simulates a user navigating randomly in the Web who jumps to a random page with probability q or follows a random hyperlink (on the current page) with probability $1 - q$. It is further assumed that this user never goes back to a previously visited page following an already traversed hyperlink backwards. This process can be modeled with a Markov chain, from where the stationary probability of being in each page can be computed. This value is then used as part of the ranking mechanism of Google. Let $C(a)$ be the number of outgoing links of page a and suppose that page a is pointed to by pages p_1 to p_n . Then, the PageRank, $PR(a)$ of a is defined as

$$PR(a) = q + (1 - q) \sum_{i=1}^n PR(p_i) / C(p_i)$$

where q must be set by the system (a typical value is 0.15). Notice that the ranking (weight) of other pages is normalized by the number of links in the page. PageRank can be computed using an iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web (which is the transition matrix of the Markov chain). Crawling the Web using this ordering has been shown to be better than other crawling schemes [168] (see next section).

Therefore, to help ranking algorithms, page designers should include informative titles, headings, and meta fields, as well as good links. However, keywords should not be repeated as some search engines penalize repeating words (spamming). Using full terms instead of indirect ways to refer to subjects should also be considered.

13.4.5 Crawling the Web

In this section we discuss how to crawl the Web, as there are several techniques. The simplest is to start with a set of URLs and from there extract other URLs which are followed recursively in a breadth-first or depth-first fashion. For that reason, search engines allow users to submit top Web sites that will be added to the URL set. A variation is to start with a set of popular URLs, because we can expect that they have information frequently requested. Both cases work well for one crawler, but it is difficult to coordinate several crawlers to avoid visiting the same page more than once. Another technique is to partition the Web using country codes or Internet names, and assign one or more robots to each partition, and explore each partition exhaustively.

Considering how the Web is traversed, the index of a search engine can be thought of as analogous to the stars in an sky. What we see has never existed, as the light has traveled different distances to reach our eye. Similarly, Web pages referenced in an index were also explored at different dates and they may not exist any more. Nevertheless, when we retrieve a page, we obtain its actual content. How fresh are the Web pages referenced in an index? The pages will be from one day to two months old. For that reason, most search engines show in the answer the date when the page was indexed. The percentage of invalid links stored in search engines vary from 2 to 9%. User submitted pages are usually crawled after a few days or weeks. Starting there, some engines traverse the whole Web site, while others select just a sample of pages or pages up to a certain depth. Non-submitted pages will wait from weeks up to a couple of months to be detected. There are some engines that learn the change frequency of a page and visit it accordingly [175]. They may also crawl more frequently popular pages (for example, pages having many links pointing to them). Overall, the current fastest crawlers are able to traverse up to 10 million Web pages per day.

The order in which the URLs are traversed is important. As already mentioned, the links in a Web page can be traversed breadth first or depth first. Using a breadth first policy, we first look at all the pages linked by the current page, and so on. This matches well Web sites that are structured by related topics. On the other hand, the coverage will be wide but shallow and a Web server can be bombarded with many rapid requests. In the depth first case, we follow the first link of a page and we do the same on that page until we cannot go deeper, returning recursively. This provides a narrow but deep traversal. Only recently, some research on this problem has appeared [168], showing that good ordering schemes can make a difference if crawling better pages first (using the PageRank scheme mentioned above).

Due to the fact that robots can overwhelm a server with rapid requests and can use significant Internet bandwidth (in particular the whole bandwidth of small domains can be saturated), a set of guidelines for robot behavior has been developed [457]. For this purpose, a special file is placed at the root of every Web server indicating the restrictions at that site, in particular the pages that should not be indexed. Crawlers can also have problems with HTML pages that use frames (a mechanism to divide a page in two or more parts) or image maps (hyperlinks associated to images). In addition, dynamically generated pages cannot be indexed as well as password protected pages.

13.4.6 Indices

Most indices use variants of the inverted file (see Chapter 8). In short, an inverted file is a list of sorted words (vocabulary), each one having a set of pointers to the pages where it occurs. Some search engines use elimination of stopwords to reduce the size of the index. Also, it is important to remember that a logical view of the text is indexed. Normalization operations may include removal of punctuation and multiple spaces to just one space between each word, uppercase to lowercase letters, etc. (see Chapter 7). To give the user some idea about each document retrieved, the index is complemented with a short description of each Web page (creation date, size, the title and the first lines or a few headings are typical). Assuming that 500 bytes are required to store the URL and the description of each Web page, we need 50 Gb to store the description for 100 million pages. As the user initially receives only a subset of the complete answer to each query, the search engine usually keeps the whole answer set in memory, to avoid having to recompute it if the user asks for more documents.

State of the art indexing techniques can reduce the size of an inverted file to about 30% of the size of the text (less if stopwords are used). For 100 million pages, this implies about 150 Gb of disk space. By using compression techniques, the index size can be reduced to 10% of the text [825]. A query is answered by doing a binary search on the sorted list of words of the inverted file. If we are searching multiple words, the results have to be combined to generate the final answer. This step will be efficient if each word is not too frequent. Another possibility is to compute the complete answer while the user requests more Web pages, using a lazy evaluation scheme. More details on searching over an inverted file can be found in Chapter 8.

Inverted files can also point to the actual occurrences of a word within a document (full inversion). However, that is too costly in space for the Web, because each pointer has to specify a page and a position inside the page (word numbers can be used instead of actual bytes). On the other hand, having the positions of the words in a page, we can answer phrase searches or proximity queries by finding words that are near each other in a page. Currently, some search engines are providing phrase searches, but the actual implementation is not known.

Finding words which start with a given prefix requires two binary searches in the sorted list of words. More complex searches, like words with errors,

arbitrary wild cards or, in general, any regular expression on a word, can be performed by doing a sequential scan over the vocabulary (see Chapter 8). This may seem slow, but the best sequential algorithms for this type of query can search around 20 Mb of text stored in RAM in one second (5 Mb is more or less the vocabulary size for 1 Gb of text). Thus, for several gigabytes we can answer those queries in a few seconds. For the Web this is still too slow but not completely out of the question. In fact, using Heaps' law and assuming $\beta = 0.7$ for the Web, the vocabulary size for 1 Tb is 630 Mb which implies a searching time of half a minute.

Pointing to pages or to word positions is an indication of the granularity of the index. The index can be less dense if we point to logical blocks instead of pages. In this way we reduce the variance of the different document sizes, by making all blocks roughly the same size. This not only reduces the size of the pointers (because there are fewer blocks than documents) but also reduces the number of pointers because words have locality of reference (that is, all the occurrences of a non-frequent word will tend to be clustered in the same block). This idea was used in Glimpse [540] which is at the core of Harvest [108]. Queries are resolved as for inverted files, obtaining a list of blocks that are then searched sequentially (exact sequential search can be done over 30 Mb per second in RAM). Glimpse originally used only 256 blocks, which was efficient up to 200 Mb for searching words that were not too frequent, obtaining an index of only 2% of the text. By tuning the number of blocks and the block size, reasonable space-time trade-offs can be achieved for larger document collections (for more details see Chapter 8). These ideas cannot be used (yet) for the Web because sequential search cannot be afforded, as it implies a network access. However, in a distributed architecture where the index is also distributed, logical blocks make sense.

13.5 Browsing

In this section we cover Web tools which are based on browsing and searching, in particular Web directories. Although the Web coverage provided by directories is very low (less than 1% of all Web pages), the answers returned to the user are usually much more relevant.

13.5.1 Web Directories

The best and oldest example of a Web directory is Yahoo! [839], which is likely the most used searching tool. Other large Web directories include eBLAST, LookSmart, Magellan, and NewHoo. Some of them are hybrids, because they also provide searches in the whole Web. Most search engines also provide subject categories nowadays, including AltaVista Categories, AOL Netfind, Excite Channels, HotBot, Infoseek, Lycos Subjects, and WebCrawler Select. are specific to some areas. For example, there are Web sites focused on business, news,

Web directory	URL	Web sites	Categories
eBLAST	www.eblast.com	125	–
LookSmart	www.looksmart.com	300	24
Lycos Subjects	a2z.lycos.com	50	–
Magellan	www.mckinley.com	60	–
NewHoo	www.newhoo.com	100	23
Netscape	www.netscape.com	–	–
Search.com	www.search.com	–	–
Snap	www.snap.com	–	–
Yahoo!	www.yahoo.com	750	–

Table 13.3 URLs, Web pages indexed and categories (both in thousands) of some Web directories (beginning of 1998).

Arts & Humanities	Local
Automotive	News
Business & Economy	Oddities
Computers & Internet	People
Education	Philosophy & Religion
Employment	Politics
Entertainment & Leisure	Recreation
Games	Reference
Government	Regional
Health & Fitness	Science & Technology
Hobbies & Interests	Shopping & Services
Home	Social Science
Investing	Society & Culture
Kids & Family	Sports
Life & Style	Travel & Tourism
Living	World

Table 13.4 The first level categories in Web directories.

and, in particular, research bibliography. Web directories are also called catalogs, yellow pages, or subject directories. Table 13.3 gives the URLs of the most important Web directories (not including the search engines already listed in section 13.4).

Directories are hierarchical taxonomies that classify human knowledge. Table 13.4 shows the first level of the taxonomies used by Web directories (the number of first level categories ranges from 12 to 26). Some subcategories are also available in the main page of Web directories, adding around 70 more topics. The largest directory, Yahoo!, has close to one million pages classified, followed by LookSmart, which has about 24,000 categories in total. Yahoo! also offers

14 regional or country specialized directories in other languages including Chinese, Danish, French, German, Italian, Japanese, Korean, Norwegian, Spanish, and Swedish. In most cases, pages have to be submitted to the Web directory, where they are reviewed, and, if accepted, classified in one or more categories of the hierarchy. Although the taxonomy can be seen as a tree, there are cross references, so it is really a directed acyclic graph.

The main advantage of this technique is that if we find what we are looking for, the answer will be useful in most cases. On the other hand, the main disadvantage is that the classification is not specialized enough and that not all Web pages are classified. The last problem becomes worse every day as the Web grows. The efforts to do automatic classification, by using clustering or other techniques, are very old. However, up to now, natural language processing is not 100% effective in extracting relevant terms from a document. Thus, classification is done manually by a limited number of people. This is a potential problem with users having a different notion of categories than the manmade categorization.

Web directories also allow the user to perform a search on the taxonomy descriptors or in the Web pages pointed to by the taxonomy. In fact, as the number of classified Web pages is small, we can even afford to have a copy of all pages. In that case they must be updated frequently, which may pose performance and temporal validity problems. In addition, most Web directories also send the query to a search engine (through a strategic alliance) and allow the whole Web to be searched.

13.5.2 Combining Searching with Browsing

Usually, users either browse following hypertext links or they search a Web site (or the whole Web). Currently, in Web directories, a search can be reduced to a subtree of the taxonomy. However, the search may miss related pages that are not in that part of the taxonomy. Some search engines find similar pages using common words, but often this is not effective.

WebGlimpse is a tool that tries to solve these problems by combining browsing with searching [539]. WebGlimpse attaches a small search box to the bottom of every HTML page, and allows the search to cover the neighborhood of that page or the whole site, without having to stop browsing. This is equivalent to following hypertext links that are constructed on the fly through a neighborhood search. WebGlimpse can be useful in building indices for personal Web pages or collections of favorite URLs.

First, WebGlimpse indexes a Web site (or a collection of specific documents) and computes neighborhoods according to user specifications. As a result, WebGlimpse adds the search boxes to selected pages, collects remote pages that are relevant, and caches those pages locally. Later, the users can search in the neighborhood of a page using the search boxes. As the name suggests, WebGlimpse uses Glimpse as its search engine [540].

The neighborhood of a Web page is defined as the set of Web pages that are reachable by a path of hypertext links within a maximum predefined distance. This distance can be set differently for local and remote pages. For example, it

can be unlimited locally, but be only three at any remote site. The neighborhood can also include all the subdirectories of the directory where the Web page is. The result is a graph of all the neighborhoods of the Web site or collection, and for each Web page, a file with all the Web pages in its neighborhood. When searching, any query in the whole index can be intersected with a neighborhood list, obtaining the relevant Web pages. A nice addition to WebGlimpse would be to visualize the neighborhoods. This problem is the topic of the next section.

13.5.3 Helpful Tools

There are many software tools to help browsing and searching. Some of them are add-ons to browsers, such as Alexa [10]. Alexa is a free Web navigation service that can be attached as a toolbar at the bottom of any browser and accompanies the user in his surfing. It provides useful information about the sites that are visited, including their popularity, speed of access, freshness, and overall quality (obtained from votes of Alexa users). Alexa also suggests related sites helping one's navigation. Another navigation service and searching guide is WebTaxi [805].

There are other tools that use visual metaphors, which can be broadly classified into two types: tools designed to visualize a subset of the Web and tools designed to visualize large answers. Both cases need to represent a large graph in a meaningful way. Specific commercial examples of tools to visualize Web subsets are Microsoft's SiteAnalyst (formerly from NetCarta), MAPA from Dynamic Diagrams, IBM's Mapuccino (formerly WebCutter [527], shown in Figure 10.22), SurfSerf, Merzscope from Merzcom, CLEARweb, Astra SiteManager, WebAnalyzer from InContext, HistoryTree from SmartBrowser, etc. Non-commercial works include WebMap [220], Sitemap, Ptolomeaus, and many earlier research [234, 578, 564, 20]. We have not included more generic visualization software, where Web visualization is just a particular case, or other related visualization tools such as Web usage analysis [642, 294, 737].

Metaphors to visualize large answers are covered in Chapter 10. Visual tools are not yet deployed in the whole Web because there is no standard way of communicating visualizers and search engines. One possible approach is to use a markup language based on XML (see Chapter 6), as proposed in [15].

13.6 Metasearchers

Metasearchers are Web servers that send a given query to several search engines, Web directories and other databases, collect the answers and unify them. Examples are Metacrawler [715] and SavvySearch [383, 223]. The main advantages of metasearchers are the ability to combine the results of many sources and the fact that the user can pose the same query to various sources through a single common interface. Metasearchers differ from each other in how ranking

Metasearcher	URL	Sources used
Cyber 411	www.cyber411.com	14
Dogpile	www.dogpile.com	25
Highway61	www.highway61.com	5
Inference Find	www.infind.com	6
Mamma	www.mamma.com	7
MetaCrawler	www.metacrawler.com	7
MetaFind	www.metafind.com	7
MetaMiner	www.miner.uol.com.br	13
MetaSearch	www.metasearch.com	—
SavvySearch	savvy.cs.colostate.edu:2000	>13

Table 13.5 URLs of metasearchers and number of sources that they use (October 1998).

is performed in the unified result (in some cases no ranking is done), and how well they translate the user query to the specific query language of each search engine or Web directory (the query language common to all of them could be small). Table 13.5 shows the URLs of the main metasearch engines as well as the number of search engines, Web directories and other databases that they search. Metasearchers can also run on the client, for example, Copernic, EchoSearch, WebFerret, WebCompass, and WebSeeker. There are others that search several sources and show the different answers in separate windows, such as All4One, OneSeek, Proteus, and Search Spaniel.

The advantages of metasearchers are that the results can be sorted by different attributes such as host, keyword, date, etc; which can be more informative than the output of a single search engine. Therefore browsing the results should be simpler. On the other hand, the result is not necessarily all the Web pages matching the query, as the number of results per search engine retrieved by the metasearcher is limited (it can be changed by the user, but there is an upper limit). Nevertheless, pages returned by more than one search engine should be more relevant.

We expect that new metasearchers will do better ranking. A first step in this direction is the NEC Research Institute metasearch engine, Inquirus [488, 489]. The main difference is that Inquirus actually downloads and analyzes each Web page obtained and then displays each page, highlighting the places where the query terms were found. The results are displayed as soon as they are available in a progressive manner, otherwise the waiting time would be too long. This technique also allows non-existent pages or pages that have changed and do not contain the query any more to be discarded, and, more important, provides for better ranking than normal search engines. On the other hand, this metasearcher is not available to the general public.

Measure	Average value	Range
Number of words	2.35	0 to 393
Number of operators	0.41	0 to 958
Repetitions of each query	3.97	1–1.5 million
Queries per user session	2.02	1–173,325
Screens per query	1.39	1–78,496

Table 13.6 Queries on the Web: average values.

The use of metasearchers is justified by coverage studies that show that a small percentage of Web pages are in all search engines [91]. In fact, fewer than 1% of the Web pages indexed by AltaVista, HotBot, Excite, and Infoseek are in all of those search engines. This fact is quite surprising and has not been explained (yet). Metasearchers for specific topics can be considered as software agents and are covered in section 13.8.2.

13.7 Finding the Needle in the Haystack

13.7.1 User Problems

We have already glanced at some of the problems faced by the user when interacting with the query interfaces currently provided by search engines. First, the user does not exactly understand the meaning of searching using a set of words, as discussed in Chapter 10. Second, the user may get unexpected answers because he is not aware of the logical view of the text adopted by the system. An example is the use of uppercase letters when the search engine is not case sensitive. Hence, a word like ‘Bank’ loses part of its semantics if we search for ‘bank.’ Simple experiments also show that due to typos or variations of a word, even if correctly capitalized, 10–20% of the matches can be lost. Similarly, foreign names or words that are difficult to spell may appear incorrectly which may result in a loss of up to 50% of the relevant answers, as mentioned in section 13.2.

Another problem is that most users have trouble with Boolean logic. In natural language, sometimes we use ‘and’ and ‘or’ with different meaning depending on the context. For example, when choosing between two things, we use an exclusive ‘or,’ which does not match the Boolean interpretation. Because of this, several studies show that around 80% of the queries do not use any Boolean or other operation. For these reasons many people have trouble using command query languages, and query forms should clearly specify which words must or must not be contained in a document that belongs to the answer.

There are a few surveys and analyses of query logs with respect to the usage of search engines [647, 403, 728]. The latter reference is based on 285 million user sessions containing 575 million queries. Table 13.6 gives the main results

of that study, carried out in September 1998. Some of the strange results might be due to queries done by mechanized search agents. The number of queries submitted per day to AltaVista is over 13 million. Users select a search engine mainly based on ease of use, speed, coverage, relevance of the answer, and habit. The main purposes are research, leisure, business, and education. The main problems found are that novice users do not know how to start and lack the general knowledge that would help in finding better answers. Other problems are that search engines are slow, that the answer is too large, not very relevant, and not always up to date. Also, most people do not care about advertising, which is one of the main sources of funding for search engines. When searching, 25% of the users use a single keyword, and on average their queries have only two or three terms. In addition, about 15% of the users restrict the search to a predefined topic and most of them (nearly 80%) do not modify the query. In addition, most users (about 85%) only look at the first screen with results and 64% of the queries are unique. Also, many words appear in the same sentence, suggesting that proximity search should be used. There are also studies about users' demographics and software and hardware used.

13.7.2 Some Examples

Now we give a couple of search examples. One problem with full-text retrieval is that although many queries can be effective, many others are a total deception. The main reason is that a set of words does not capture all the semantics of a document. There is too much contextual information (that can be explicit or even implicit) lost at indexing time, which is essential for proper understanding. For example, suppose that we want to learn an oriental game such as Shogi or Go. For the first case, searching for Shogi will quickly give us good Web pages where we can find what Shogi is (a variant of chess) and its rules. However, for Go the task is complicated, because unlike Shogi, Go is not a unique word in English (in particular, because uppercase letters are converted to lowercase letters, see Chapter 7). The problem of having more than one meaning for a word is called *polysemy*. We can add more terms to the query, such as **game** and **Japanese** but still we are out of luck, as the pages found are almost all about Japanese games written in English where the common verb *go* is used. Another common problem comes from synonyms. If we are searching for a certain word, but a relevant page uses a synonym, we will not find it.

The following example (taken from [152]) better explains the polysemy problem, where the ambiguity comes from the same language. Suppose that we want to find the running speed of the jaguar, a big South American cat. A first naive search in AltaVista would be **jaguar speed**. The results are pages that talk about the Jaguar car, an Atari video game, a US football team, a local network server, etc. The first page about the animal is ranked 183 and is a fable, without information about the speed. In a second try, we add the term **cat**. The answers are about the Clans Nova Cat and Smoke Jaguar, LMG Enterprises, fine cars, etc. Only the page ranked

25 has some information on jaguars but not the speed. Suppose we try Yahoo!. We look at 'Science:Biology:Zoology:Animals:Cats:Wild_Cats' and 'Science:Biology:Animal_Behavior.' No information about jaguars there.

13.7.3 Teaching the User

Interfaces are slowly improving in assisting the user with the task of acquiring a better grasp of what Web pages are being retrieved. Query forms must specify clearly if one or all the words must be in a page, which words should not be in a page, etc., without using a written Boolean query language. Second, users should try to give as many terms as possible, in particular terms that must be or should not be in the pages. In particular, a user should include all possible synonyms of a word. If the user can restrict the search to a field (for example, the page title) or limit some attribute (date, country), this will certainly reduce the size of the answer. In case of doubt, the user should remember to look at the help information provided by the search engine. If he cannot find where one of the relevant terms is in a page, he can use the Find option of the browser.

Even if we are able to pose a good query, the answer can still be quite large. Considering that the visual tools mentioned before are not yet available for the general public, the user must learn from experience. There are many strategies for quickly finding relevant answers. If the user is looking for an institution, he can always try to guess the corresponding URL by using the `www` prefix followed by a guessed institution acronym or brief name and ending with a top level domain (country code or `com`, `edu`, `org`, `gov` for the US). If this does not work, the user can search the institution name in a Web directory.

If we are looking for work related to a specific topic, a possible strategy is: (1) select an article relevant to the topic, if possible with non-common author surnames or title keywords (if it is not available, try any bibliographic database or a Web directory search for a first reference); and (2) use a search engine to find all Web pages that have all those surnames and keywords. Many of the results are likely to be relevant, because we can find: (a) newer papers that reference the initial reference, (b) personal Web pages of the authors, and most important, (c) pages about the topic that already contain many relevant references. This strategy can be iterated by changing the reference used as better references appear during the search.

As mentioned at the beginning of this chapter, the Web poses so many problems, that it is easier and more effective to teach the user how to properly profit from search engines and Web directories, rather than trying to guess what the user really wants. Given that the coverage of the search engines is low, use several engines or a metasearcher. Also, remember that you have to evaluate the quality of each answer, even if it appears to be relevant. Remember that anybody can publish in the Web, and that does not mean that the data is correct or still valid. The lessons learned in the examples shown above are: (1) search engines still return too much hay together with the needle; and (2) Web directories do not have enough depth to find the needle. So, we can use the following rules of thumb:

- **Specific queries** Look in an encyclopedia, that is the reason that they exist. In other words, do not forget libraries.
- **Broad queries** Use Web directories to find good starting points.
- **Vague queries** Use Web search engines and improve the query formulation based on relevant answers.

13.8 Searching using Hyperlinks

In this section we cover other paradigms to search the Web, which are based on exploiting its hyperlinks. They include Web query languages and dynamic searching. These ideas are still not widely used due to several reasons, including performance limitations and lack of commercial products.

13.8.1 Web Query Languages

Up to this point, queries have been based on the content of each page. However, queries can also include the link structure connecting Web pages. For example, we would like to search for all the Web pages that contain at least one image and are reachable from a given site following at most three links. To be able to pose this type of query, different data models have been used. The most important are a labeled graph model to represent Web pages (nodes) and hyperlinks (edges) between Web pages, and a semi-structured data model to represent the content of Web pages. In the latter model, the data schema is not usually known, may change over time, may be large and descriptive, etc. [2, 129].

Although some models and languages for querying hypertext were proposed before the Web appeared [563, 72, 184], the first generation of Web query languages were aimed at combining content with structure (see also Chapter 4). These languages combine patterns that appear within the documents with graph queries describing link structure (using path regular expressions). They include W3QL [450], WebSQL [556, 33], WebLog [476], and WQL [511]. The second generation of languages, called Web data manipulation languages, maintain the emphasis on semi-structured data. However, they extend the previous languages by providing access to the structure of Web pages (the model also includes the internal structure) and by allowing the creation of new structures as a result of a query. Languages in this category include STRUQL [253], FLORID [373], and WebOQL [32]. All the languages mentioned are meant to be used by programs, not final users. Nevertheless, there are some examples of query interfaces for these languages.

Web query languages have been extended to other Web tasks, such as extracting and integrating information from Web pages, and constructing and restructuring Web sites. More details about Web query languages can be found in the excellent survey by Florescu, Levy, and Mendelzon [258].

13.8.2 Dynamic Search and Software Agents

Dynamic search in the Web is equivalent to sequential text searching. The idea is to use an online search to discover relevant information by following links. The main advantage is that you are searching in the current structure of the Web, and not in what is stored in the index of a search engine. While this approach is slow for the entire Web, it might be used in small and dynamic subsets of the Web. The first heuristic devised was the *fish search* [113], which exploits the intuition that relevant documents often have neighbors that are relevant. Hence, the search is guided by following links in relevant documents. This was improved by *shark search* [366], which does a better relevance assessment of neighboring pages. This algorithm has been embedded in Mapuccino (see section 13.5.3), and Figure 10.22 shows a Web subset generated by this type of search. The main idea of these algorithms is to follow links in some priority, starting from a single page and a given query. At each step, the page with highest priority is analyzed. If it is found to be relevant, a heuristic decides to follow or not to follow the links on that page. If so, new pages are added to the priority list in the appropriate positions.

Related work includes software agents for searching specific information on the Web [602, 477]. This implies dealing with heterogeneous sources of information which have to be combined. Important issues in this case are how to determine relevant sources (see also Chapters 9 and 15, as well as section 10.4.4) and how to merge the results retrieved (the fusion problem). Examples are shopping robots such as Jango [401], Junglee [180], and Express [241].

13.9 Trends and Research Issues

The future of the Web might surprise us, considering that its massive use started less than five years ago. There are many distinct trends and each one opens up new and particular research problems. What follows is a compilation of the major trends as we have perceived them.

- **Modeling:** Special IR models tailored for the Web are needed [308, 155, 652]. As we have seen, Web user queries are different. We also have the pull/push dichotomy: Will we search for information or will the information reach us? In both cases we need better search paradigms and better information filtering [782].
- **Querying:** Further work on combining structure and content in the queries is needed as well as new visual metaphors to pose those queries and visualize the answers [44]. Future query languages may include concept-based search and natural language processing, as well as searching by example (this implies document clustering and categorization on the Web [810, 120, 157]).
- **Distributed architectures:** New distributed schemes to traverse and search the Web must be devised to cope with its growth. This will have an impact on current crawling and indexing techniques, as well as caching

techniques for the Web. Which will be the bottleneck in the future? Server capacity or network bandwidth?

- **Ranking:** Better ranking schemes are needed, exploiting both content and structure (internal to a page and hyperlinks); in particular, combining and comparing query-dependent and independent techniques. One problem related to advertisements is that search engines may rank some pages higher due to reasons that are not based on the real relevance of a page (this is called the search engine persuasion problem in [543]).
- **Indexing:** Which is the best logical view for the text? What should be indexed? How to exploit better text compression schemes to achieve fast searching and get lower network traffic? How to compress efficiently word lists, URL tables, etc. and update them without significant run-time penalty? Many implementation details must be improved.
- **Dynamic pages:** A large number of Web pages are created on demand and current techniques are not able to search on those dynamic pages. This is called the hidden Web.
- **Duplicated data:** Better mechanisms to detect and eliminate repeated Web pages (or pages that are syntactically very similar) are needed. Initial approaches are based on resemblance measures using document fingerprints [121, 120]. This is related to an important problem in databases: finding similar objects.
- **Multimedia:** Searching for non-textual objects will gain importance in the near future. There are already some research results in the literature [579, 80, 136].
- **User interfaces:** Better user interfaces are clearly needed. The output should also be improved, for example allowing better extraction of the main content of a page or the formulation of content-based queries [766].
- **Browsing:** More tools will appear, exploiting links, popularity of Web pages, content similarity, collaboration, 3D, and virtual reality [384, 638, 385, 421]. An important trend would be to unify further searching with browsing.

An important issue to be settled in the future is a standard protocol to query search engines. One proposal for such a protocol is STARTS [316], which could allow us to choose the best sources for querying, evaluate the query at these sources, and merge the query results. This protocol would make it easier to build metasearchers, but at the same time that is one of the reasons for not having a standard. In that way, metasearchers cannot profit from the work done by search engines and Web directories. This is a particular case of the federated searching problem from heterogeneous sources as it is called in the database community [656]. This is a problem already studied in the case of the Web, including discovery and ranking of sources [161, 845, 319]. These issues are also very important for digital libraries [649] (see also Chapter 15) and visualization issues [15]. A related topic is metadata standards for the Web (see Chapter 6)

and their limitations [544]. XML helps [436, 213, 306], but semantic integration is still needed.

Hyperlinks can also be used to infer information about the Web. Although this is not exactly searching the Web, this is an important trend called Web mining. Traditionally, Web mining had been focused on text mining, that is, extracting information from Web pages. However, the hyperlink structure can be exploited to obtain useful information. For example, the ParaSite system [736] uses hyperlink information to find pages that have moved, related pages, and personal Web pages. HITS, already mentioned in Section 13.4.4, has also been used to find communities and similar pages [444, 298]. Other results on exploiting hyperlink structure can be found in [639, 543, 154]. Further improvements in this problem include Web document clustering [810, 120, 162] (already mentioned), connectivity services (for example, asking which Web pages point to a given page [92]), automatic link generation [320], extracting information [100, 115], etc.

Another trend is intranet applications. Many companies do not want their private networks to be public. However, for business reasons they want to allow Web users to search inside their intranets obtaining partial information. This idea leads to the concept of *portals* for which there are already several commercial products. New models to see Web sites as databases and/or information systems are also important.

13.10 Bibliographic Discussion

There are hundreds of books about the Web. Many of them include some information about searching the Web and tips for users. A recent book edited by Abrams includes a chapter on searching the Web [3]. Other sources are [682], the special numbers of *Scientific American* on the Internet (March 1997) and IEEE's *Internet Computing on Search Technologies* (July/August 1998). For details about crawlers and other software agents see [166, 817].

In addition, the best source for references to the Web is the Web itself. To start with, there are many Web sites devoted to inform and rate search engines and Web directories. Among them we can distinguish Search Engine Watch [749] and Search Engine Showdown [609]. A survey about Web characterizations is given by Pitkow [641] and a good directory to Web characteristics is [217]. Other Web pages provide pointers and references related to searching the Web, in particular the World Wide Web Consortium (www.w3.org), the World Wide Web journal (w3j.com) and WWW conferences. These and other pointers are available in the Web page of this book (see Chapter 1).

Acknowledgements

We would like to thank the following for their helpful comments: Omar Alonso, Eric Brown, Pablo de la Fuente, Monika Henzinger and Gonzalo Navarro.

Glossary

Ad hoc retrieval:

standard retrieval task in which the user specifies his information need through a query which initiates a search (executed by the information system) for documents which are likely to be relevant to the user.

All-pairs or spatial-join query:

a query that requests all the *pairs* of objects that are within the specified distance from their partner.

Amdahl's law:

Using N processors, the maximal speedup S obtainable for a given problem is related to f , the fraction of the problem that must be computed sequentially. The relationship is given by: $S_N \leq \frac{1}{f + (1-f)/N} \leq \frac{1}{f}$

ASCII:

Standard binary codes to represent characters in one byte for most European languages.

Belief network:

a probabilistic model of document retrieval based on interpreting documents, user queries, and index terms as nodes of a Bayesian network. This model is distinct from the inference network model.

Bit-parallelism:

a speed-up technique based on exploiting the fact that the processor performs some operations in parallel over all the bits of the computer word.

Block addressing:

a technique used to reduce the size of the lists of occurrences by pointing to text blocks instead of exact positions.

Boolean model:

a classic model of document retrieval based on classic set theory.

Browsing:

interactive task in which the user is more interested in exploring the document collection than in retrieving documents which satisfy a specific information need.

CACM collection:

a reference collection composed of all the 3204 articles published in the Communications of the ACM from 1958 to 1979.

CISI collection:

a reference collection composed of 1460 documents selected from previous collections assembled at ISI (Institute of Scientific Information).

Clustering:

the grouping of documents which satisfy a set of common properties. The aim is to assemble together documents which are related among themselves. Clustering can be used, for instance, to expand a user query with new and related index terms.

Coding:

the substitution of text symbols by numeric codes with the aim of encrypting or compressing text.

Collection:

a group of items, often documents. In (digital) libraries this designates all the works included, usually selected based on a collection management plan.

Compression of text:

the study of techniques for representing text in fewer bytes or bits.

Content-based query:

query exploiting data content.

Conversion:

changing from one form to another, as in converting from analog to digital (also called 'digitization'), or paper to online (as in 'retrospective conversion' of a card catalog to an online catalog, or old books to scanned images).

Cystic Fibrosis collection:

a reference collection composed of 1239 documents indexed with the term *cystic fibrosis* in the National Library of Medicine's MEDLINE database.

Data cartridge:

data structure and associated methods to represent and query a particular multimedia data type.

Data mining:

extraction of new data, relations, or partial information from any type of data.

Data retrieval:

the retrieval of items (tuples, objects, Web pages, documents) whose contents satisfy the conditions specified in a (regular expression like) user query.

Database industry:

the organizations, including commercial, government, and not-for-profit sectors, who produce, provide access to, and market databases of bibliographic, reference, and full-text information.

Database producers:

the organizations, including commercial, government, and not-for-profit sectors, who create electronic abstracting and indexing tools as well as other reference and full-text databases.

Database vendors:

the organizations, primarily in the commercial and not-for-profit sectors, who license databases from their producers and provide search software and a front end for consumer access to the information contained.

Digital library:

the combination of a collection of digital objects (repository); descriptions of those objects (metadata); a set of users (patrons or target audience or users); and systems that offer a variety of services such as capture, indexing, cataloging, search, browsing, retrieval, delivery, archiving, and preservation.

Digital object:

a string of bits that is viewed as an entity in its own right (e.g., a full-text document) though it may be a part of another digital object (e.g., an image that is part of a book), often with associated 'metadata' and sometimes with terms and conditions (especially on access).

Digital preservation:

ensuring that a digital object continues to be accessible and useful over a long period of time, which usually requires both media conversion (copying from one old tape format to a new tape format before the old tapes are no longer readable) and format conversion (changing from some file structure or encoding to a newer one that will continue to be used and understood).

Directory:

a usually hierarchical categorization of concepts in a domain of knowledge.

Distributed computing:

the application of multiple computers connected by a network to solve a single problem.

Distributed information retrieval:

the application of distributed computing techniques to solve information retrieval problems.

DLI:

Digital Libraries Initiative, a program of the US National Science Foundation, for research and development related to digital libraries, which began with \$24M of funding split across six universities for 1994-98, and which will continue from 1998 onward with roughly double that amount of support

Document:

a unit of retrieval. It might be a paragraph, a section, a chapter, a Web page, an article, or a whole book.

Document surrogate:

a representation of a document such as the title and a short summary. Surrogates are commonly used to display the answers to a user query.

DTD:

Document Type Definition: SGML definition for a markup language.

E measure:

an information retrieval performance measure, distinct from the harmonic mean, which combines recall and precision.

Edit distance:

(between two strings): minimum number of insertions, deletions, and replacements of characters necessary to make two strings equal.

Efficiency:

a measure of parallel algorithm performance given by: $\phi = \frac{S}{N}$, where S is speedup and N is the number of processors.

Entropy:

measure of information defined on the statistics on the characters of a text.

Exact match:

mechanism by which only the objects satisfying some well specified criteria, against object attributes, are returned to the user as a query answer.

Extended Boolean model:

a set theoretic model of document retrieval based on an extension of the classic Boolean model. The idea is to interpret partial matches as Euclidean distances represented in a vectorial space of index terms.

Extended pattern:

a general pattern allowing rich expressions such as wild cards, classes of characters, and others.

Faceted query:

a query which is divided into topics and/or facets, each of which should be present in documents in the answer.

Feature:

information extracted from an object and used during query processing.

Federated search:

support for finding items that are scattered among a distributed collection of information sources or services, typically involving sending queries to a number of servers and then merging the results to present in an integrated, consistent, coordinated format.

Filtering:

retrieval task in which the information need of the user is relatively static while *new* documents constantly enter the system. Typical examples are news wire services and electronic mail lists.

Full text:

a logical view of the documents in which all the words which compose the text of the document are used as indexing terms.

Fuzzy model:

a set theoretic model of document retrieval based on fuzzy theory.

Generalized vector space model:

a generalization of the classic vector model based on a less restrictive interpretation of term-to-term independence.

Global analysis:

a reference to techniques of identifying document and term relationships through the analysis of all the documents in a collection. Global analysis is used, for instance, to build thesauri.

Granularity:

the amount of computation relative to the amount of communication performed by a parallel program.

Guided tour:

a sequence of navigational choices (usually in a hypertext) aimed at presenting the nodes in a logical order for some goal.

Harmonic mean:

an information retrieval performance measure which combines recall and precision.

Heaps' Law:

an empirical rule which describes the vocabulary growth as a function of the text size. It establishes that a text of n words has a vocabulary of size $O(n^\beta)$ for $0 < \beta < 1$.

HTML:

hypertext markup language of the Web, based on SGML.

Huffman coding:

an algorithm for coding text in which the most frequent symbols are represented by the shortest codes.

Human-Computer Interfaces (HCI):

the study of interfaces which assist the user with information seeking related tasks such as: query formulation, selection of information sources, understanding of search results, and tracking of the retrieval task.

Hypertext model:

a model of information retrieval based on representing document relationships as edges of a generic graph in which the documents are the nodes.

Independence of index terms:

see *Term-to-term independence*.

Index:

a data structure built on the text to speed up searching.

Index point:

the initial position of a text element which can be searched for, for example a word.

Index term (or keyword):

a pre-selected term which can be used to refer to the content of a document. Usually, index terms are nouns or noun groups. In the Web, however, some search engines use all the words in a document as index terms.

Inference network:

a probabilistic model of document retrieval based on interpreting documents, user queries, and index terms as nodes of a Bayesian network.

Information retrieval (IR):

part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a *user information need* usually expressed in natural language.

Information retrieval performance:

an evaluation of an information system in terms of the quality of the answers it generates with regard to a set of test queries. The quality of the answer set is evaluated by comparing the documents in it with those in a set of documents (provided by specialists) known to be relevant to the test query in focus.

Informative feedback:

information to the user about the relationship between the query specification and the documents retrieved.

Interoperability:

the working together of a number of computer systems, typically for a common purpose, such as when a number of digital libraries 'support federated searching,' often enabled by standards and agreed-upon conventions including data formats and protocols.

Intranet:

an Internet type network built inside an organization, which may or may not be connected to the Internet itself.

Inverted file:

a text index composed of a vocabulary and a list of occurrences.

Keyword:

see *Index term*.

Kohonen's feature map:

a bi-dimensional map whose regions represent the main themes in a document or in a collection.

KWIC:

KeyWords In Context, a technique that displays the occurrences of query terms within the context of the documents retrieved.

Latent semantic indexing:

an algebraic model of document retrieval based on a singular value decomposition of the vectorial space of index terms.

Levenshtein distance:

see *Edit distance*.

Lexicographical order:

order in which the words are listed in a dictionary or telephone guide.

Local analysis:

a reference to techniques of identifying document and term relationships through the analysis of the documents retrieved by a given user query.

Local context analysis:

a technique of query expansion which combines local and global analysis.

Logical view of documents:

the representation of documents and Web pages adopted by the system. The most common form is to represent the text of the document by a set of indexing terms or keywords.

MARC:

a standardized record format used by libraries and bibliographic utilities for sharing and storing cataloging information about bibliographic materials.

Metasearch:

a search technique common on the World Wide Web where a single point of entry is provided to multiple heterogeneous back-end search engines. A meta search system sends a user's query to the back-end search engines, combines the results, and returns a single, unified hit-list to the user.

Metadata:

attributes of data or a document, usually descriptive as author or content, often broken up into categories or facets, typically maintained in a catalog, sometimes recorded according to a scheme like the Dublin Core or MARC.

MIMD:

a parallel computer architecture consisting of multiple instruction streams and multiple data streams.

MISD:

a parallel computer architecture consisting of multiple instruction streams and a single data stream.

Model for IR:

a set of premisses and an algorithm for ranking documents with regard to a user query. More formally, an IR model is a quadruple $[\mathbf{D}, \mathbf{Q}, \mathcal{F}, R(q_i, d_j)]$ where $\mathbf{D}$ is a set of logical views of documents, $\mathbf{Q}$ is a set of user queries, $\mathcal{F}$ is a framework for modeling documents and queries, and $R(q_i, d_j)$ is a ranking function which associates a numeric ranking to the query q_i and the document d_j .

Modeling:

the part of IR which studies the algorithms (or *models*) used for ranking documents according to a system assigned likelihood of relevance to a given user query.

Multidimensional Scaling (MDS):

a method to map objects into points, trying to preserve distances.

Multimedia data:

data combining several different media, such as text, images, sound, and video.

Multimedia Database Management System:

system able to represent, manage, and store multimedia data.

Multimedia Information Retrieval System:

Information Retrieval system handling multimedia data.

Multitasking:

the simultaneous execution of multiple, independent tasks. On a single processor machine the tasks share the processor and their execution is interleaved sequentially. On a machine with multiple processors the tasks may execute concurrently.

Nearest-neighbor query:

a query that requests the spatial object closest to the specified object.

Neural networks:

an algebraic model of document retrieval based on representing query, index terms, and documents as a neural network.

 n -gram:

any substring of length n .

Non-overlapping lists model:

a model of retrieving structured documents through indexing structures implemented as non-overlapping lists.

Object-relational database technology:

technology that extends the relational model with the main features of the object-oriented one.

Occurrence list:

a data structure which assigns to each text word the list of its positions in the text.

OCR (Optical Character Recognition):

software that takes as input a bit-mapped image (for example, produced by a scanner) and outputs any text present in the image as ASCII text. The quality of the result depends on the text type and fonts and can be as high as 99%.

Online Public Access Catalog:

see *OPAC*.

OPAC (Online Public Access Catalog):

library management system software which provides user access to information contained in a library collection.

Optical Character Recognition:

see *OCR*.

Panning (see also *zooming*):

action of a movie camera that scans sideways across a scene. This effect can be simulated in a screen window even in the absence of the camera.

Parallel computing:

the simultaneous application of multiple processors to solve a single problem. Each processor works on a different part of the problem.

Parallel information retrieval:

the application of parallel computing to solve information retrieval problems.

Pattern:

set of syntactic features that describe the text segments to be matched, ranging from simple words to regular expressions.

Precision:

an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant.

Probabilistic model:

a classic model of document retrieval based on a probabilistic interpretation of document relevance (to a given user query).

Proximal nodes model:

a model of retrieving structured documents through a hierarchical indexing structure.

Query:

the expression of the user information need in the input language provided by the information system. The most common type of input language simply allows the specification of keywords and of a few Boolean connectives.

Query expansion:

a process of adding new terms to a given user query in an attempt to provide better contextualization (and hopefully retrieve documents which are more useful to the user).

Query preview:

a low-cost, rapid-turnaround visualization of what the results of many variations on a query might be.

Range query:

a query that requests all the spatial objects that intersect the given range.

Recall:

an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved.

Reference collection:

a collection of documents used for testing information retrieval models and algorithms. A reference collection usually includes a set of documents, a set of test queries, and a set of documents known to be relevant to each query.

Regular expression:

general pattern that allows the expression of alternative strings, repetitions and concatenations of substrings.

Repository:

a physical or digital place where objects are stored for a period of time, from which individual objects can be obtained if they are requested and their terms and conditions are satisfied.

Retrieval task:

the task executed by the information system in response to a user request. It is basically of two types: *ad hoc* and *filtering*.

Retrieval unit:

the type of objects returned by an information retrieval system as the response to a query, e.g. documents, files, Web pages, etc.

Scatter/Gather:

a browsing strategy which clusters the local documents in the answer set dynamically into topically-coherent groups and presents the user with descriptions of such groups.

Search history:

a mechanism for tracking the history of a user session or of a collection of user sessions. The search history should show what the available choices are at any given point, what moves have been made in the past, short-term tactics, and annotations on choices made along the way.

Search starting point:

‘where’ or ‘how’ an information-seeking task is initiated. Search interfaces should provide the users with good ways to get started.

Semi-static text:

a text collection which does not change too frequently.

Semi-structured data:

data whose structure may not match, or only partially match, the structure prescribed by the data schema.

Sequential or on-line text searching:

the problem of finding a pattern in a text without using any precomputed information on the text.

SGML:

standard markup metalanguage. HTML is a markup language based on SGML.

Signature file:

a text index based on storing a signature for each text block to be able to filter out some blocks quickly.

SIMD:

a parallel computer architecture consisting of a single instruction stream and multiple data streams.

SISD:

a traditional sequential computer architecture consisting of a single instruction stream and a single data stream.

SMP (Symmetric Multiprocessor):

a common MIMD computer architecture where all of the parallel processors have equal access to the same system resources, and any processor can execute any task. The processors may operate independently, or they may cooperate to execute a parallel task concurrently.

Source selection:

the process of selecting one or more document collections from a set of document collections where the selected collections are most likely to contain documents relevant to the query. Commonly used in distributed IR systems.

SQL3:

the upcoming SQL standard for relational databases.

Spatial Access Methods:

file structures for fast storage and retrieval of spatial objects like points, lines, polygons.

Speedup:

a measure of parallel algorithm performance, defined as:

$$S = \frac{\text{Running time of best available sequential algorithm}}{\text{Running time of parallel algorithm}}$$

STARTS (Stanford Proposal for Internet Meta-Searching):

a protocol for distributed, heterogeneous search developed at Stanford University in cooperation with a consortium of search product and search service vendors.

Statistical text compression:

a reference to techniques and methods of text compression based on probability estimates for the rates of occurrence of each symbol.

Stemming:

a technique for reducing words to their grammatical roots.

Stopwords:

words which occur frequently in the text of a document. Examples of stopwords are articles, prepositions, and conjunctions.

Suffix tree and suffix array:

text indices based on a lexicographical arrangement of all the text suffixes.

Syntax tree:

structural interpretation of a query, where the nodes are the operators and the subtrees are the operands.

Tag:

a string which is used to mark the beginning or ending of structural elements in the text.

Term-to-term independence:

a fundamental assumption underlying the classic vector model which states that the vectors used for representing index terms form an orthonormal basis. Such an assumption is usually interpreted to mean pairwise orthogonality (but not in the generalized vector space model).

Text structure:

information present in a text apart from its content, which relates its different portions in a semantically meaningful way.

Textual Image:

image that is mostly text and then can be compressed more than conventional images. In addition, by using OCR, keywords can be extracted and used for retrieving the image.

Thesaurus:

a data structure composed of (1) a pre-compiled list of important words in a given domain of knowledge and (2) for each word in this list, a list of related (synonym) words.

TREC collection:

a reference collection which contains over a million documents and which has been used extensively in the TREC conferences. The TREC collection has been organized by NIST and is becoming a standard for comparing IR models and algorithms.

User information need:

a natural language declaration of the informational need of a user. For instance, *'find documents which discuss the political implications of the Monica Lewinsky scandal in the results of the 1998 elections for the U.S. Congress'*.

User relevance feedback:

an interactive process of obtaining information from the user about the relevance and the non-relevance of retrieved documents.

URL:

stands for Uniform Resource Locator and is used to address any Internet resource, including Web pages.

Vector model:

a classic model of document retrieval based on representing documents and queries as vectors of index terms. The model adopts as its foundation the notion of term-to-term independence.

Vocabulary:

the set of all the different words in a text.

WAIS:

(Wide Area Information Service): suite of protocols designed to publish information and to allow querying of databases in the Internet.

Working memory load:

information about the decisions and choices made during an interactive user session. A system with a well designed interface should keep this information around to facilitate the task of the user (for instance, allowing the user to easily resume temporarily abandoned strategies).

XML:

a subset of SGML, defined for the Web. In XML it is easier to define new markup languages.

Z39.50:

a national standard that has become an international standard. It is a protocol for client/server communication with retrieval systems, allowing a single client to interact with one or more systems, or retrieval systems (or gateways) to communicate with other retrieval systems. It supports connection-oriented sessions, having a system EXPLAIN itself, submission of queries, obtaining information, the results lists, and obtaining retrieved documents.

Zipf's Law:

an empirical rule that describes the frequency of the text words. It states that the i -th most frequent word appears as many times as the most frequent one divided by i^θ , for some $\theta \geq 1$.

Zooming (see also *panning*):

action of a movie camera that can either move in for a close-up or back away to get a wider view. This effect can be simulated in a screen window even in the absence of the camera.

References

- [1] Ijsbrand Jan Aalbersberg. Incremental relevance feedback. In *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–22, Copenhagen, Denmark, 1992.
- [2] Serge Abiteboul. Querying semi-structured data. In Foto N. Afrati and Phokion Kolaitis, editors, *Int. Conf. on Database Theory (ICDT)*, number 1186 in LNCS, pages 1–18, Delphi, Greece, 1997. Springer-Verlag.
- [3] Marc Abrams, editor. *World Wide Web: Beyond the Basics*. Prentice Hall, 1998.
- [4] Brent Agnew, Christos Faloutsos, Zhengyu Wang, Don Welch, and Xiaogang Xue. Multimedia indexing over the Web. In *Proc. of the SPIE Conference on Storage and Retrieval for Image and Video Databases*, pages 72–83, San Jose, CA, Feb 1997.
- [5] M. Agosti, G. Gradenigo, and P. G. Marchetti. A hypertext environment for interacting with large textual databases. *Information Processing & Management*, 28(3):371–387, 1992.
- [6] Rakesh Agrawal, Christos Faloutsos, and Arun Swami. Efficient similarity search in sequence databases. In *Fourth Int. Conf. on Foundations of Data Organization and Algorithms*, pages 69–84, Evanston, Illinois, Oct 1993. <ftp://olympus.cs.umd.edu/pub/TechReports/fodo.ps>.
- [7] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, and Kyuseok Shim. Fast similarity search in the presence of noise, scaling and translation in time-series databases. In *Proc. of VLDB Conf.*, pages 490–501, Zurich, Switzerland, September 1995.
- [8] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of VLDB Conf.*, pages 487–499, Santiago, Chile, Sept 1994.
- [9] A. Aho and M. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, June 1975.
- [10] Alexa: Main Page. <http://www.alexa.com>, 1998.
- [11] Alis Technologies, Web languages hit parade. <http://babel.alis.com:8080/palmares.html>, 1997.
- [12] James Allan. Relevance feedback with too much data. In *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 337–343, Seattle, WA, USA, 1995.
- [13] Bryce L. Allen. *Information Tasks: Toward a User-Centered Approach to Information Systems*. Academic Press, San Diego, CA, 1996.

- [14] Robert B. Allen, Pascal Obry, and Michael Littman. An interface for navigating clustered document sets returned by queries. In *Proc. of ACM COCS: Conference on Organizational Computing Systems*, pages 66–171, Milpitis, CA, November 1993.
- [15] O. Alonso and R. Baeza-Yates. A model for visualizing large answers in WWW. In *XVIII Int. Conf. of the Chilean CS Society*, pages 2–7, Antofagasta, Chile, 1998. IEEE CS Press.
- [16] T. M. Alrashid, J. A. Barker, B. S. Christian, S. C. Cox, M. W. Rabne, E. A. Slotta, and L. R. Upthegrove. Safeguarding Copyrighted Contents, Digital Libraries and Intellectual Property Management, CWRU's Rights Management System. *D-Lib Magazine*, April 1998. <http://www.dlib.org/dlib/april98/04barker.html>.
- [17] AltaVista: Main page. <http://www.altavista.com>, 1996.
- [18] G. M. Amdahl. Validity of the single-processor approach to achieving large scale computing capabilities. In *Proc. AFIPS 1967 Spring Joint Computer Conf.*, volume 30, pages 483–485, Atlantic City, N.J., April 1967.
- [19] A. Amir, G. Benson, and M. Farach. Let sleeping files lie: pattern matching in z-compressed files. *Journal of Computer and Systems Sciences*, 52(2):299–307, 1996.
- [20] Keith Andrews. Visualizing cyberspace: Information visualization in the Harmony Internet browser. In *Proceedings '95 Information Visualization*, pages 97–104, Atlanta, USA, October 1995.
- [21] P. Anick, J. Brennan, R. Flynn, D. Hanssen, B. Alvey, and J. Robbins. A direct manipulation interface for Boolean information retrieval via natural language query. In *Proc. of the 13th Annual International ACM/SIGIR Conference*, pages 135–150, Brussels, Belgium, 1990.
- [22] Peter G. Anick. Adapting a full-text information retrieval system to the computer troubleshooting domain. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–358, 1994.
- [23] ANSI/NISO Standards, Z39.50-Information Retrieval: Application Service Definition and Protocol Specification, 1995. See <http://lcweb.loc.gov/z3950/agency>.
- [24] P. Apers, H. Blanken, and M. Houtsma. *Multimedia Databases in Perspective*. Springer-Verlag, 1997.
- [25] A. Apostolico and Z. Galil. *Combinatorial Algorithms on Words*. Springer-Verlag, New York, 1985.
- [26] M. Araújo, G. Navarro, and N. Ziviani. Large text searching allowing errors. In *Proc. WSP'97*, pages 2–20, Valparaíso, Chile, 1997. Carleton University Press.
- [27] Y. Arens, C. Chee, C. Hsu, and C. Knoblock. Retrieving and integrating data from multiple information sources. *Journal on Intelligent and Cooperative Information Systems*, 2(2):127–158, 1993.
- [28] W. Y. Arms. Key concepts in the architecture of the digital library. *D-Lib Magazine*, July 1995. <http://www.dlib.org/dlib/July95/07arms.html>.
- [29] W. Y. Arms. Relaxing assumptions about the future of digital libraries: The hare and the tortoise. *D-Lib Magazine*, April 1997. <http://www.dlib.org/dlib/april97/04arms.html>.
- [30] W. Y. Arms. Implementing Policies for Access Management. *D-Lib Magazine*, February 1998. <http://www.dlib.org/dlib/february98/arms/02arms.html>.
- [31] W. Y. Arms, C. Blanchi, and E. A. Overly. An Architecture for Information in digital libraries. *D-Lib Magazine*, February 1997. <http://www.dlib.org/dlib/february97/cnri/02arms1.html>.

- [32] Gustavo Arocena and Alberto Mendelzon. WebOQL: Restructuring documents, databases and Webs. In *Int. Conf. on Data Engineering*, pages 24–33, Orlando Florida, 1998.
- [33] Gustavo O. Arocena, Alberto O. Mendelzon, and George A. Mihaila. Applications of a Web query language. In *6th Int. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [34] AskJeeves: Main Page. <http://www.askjeeves.com>, 1998.
- [35] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the ACM*, 24(3):397–417, 1977.
- [36] R. Baeza-Yates. An extended model for full-text databases. *Journal of Brazilian CS Society*, 3(2):57–64, April 1996.
- [37] R. Baeza-Yates, E. Barbosa, and N. Ziviani. Hierarchies of indices for text searching. In *Proceedings of RIAO'94 Intelligent Multimedia Information Retrieval Systems and Management*, pages 11–13, 1994.
- [38] R. Baeza-Yates and G. Gonnet. Efficient Text Searching of Regular Expressions. In G. Ausiello, M. Dezani-Ciancaglini, and S. Ronchi Della Rocca, editors, *ICALP'89*, number 372 in LNCS, pages 46–62, Stresa, Italy, 1989. Springer-Verlag.
- [39] R. Baeza-Yates and G. Gonnet. A new approach to text searching. *Communications of the ACM*, 35(10):74–82, October 1992.
- [40] R. Baeza-Yates and G. Gonnet. Fast text searching for regular expressions or automaton searching on a trie. *J. of the ACM*, 43(6):915–936, Nov 1996.
- [41] R. Baeza-Yates and G. Navarro. Integrating contents and structure in text retrieval. *ACM SIGMOD Record*, 25(1):67–79, March 1996.
- [42] R. Baeza-Yates and G. Navarro. Block-addressing indices for approximate text retrieval. In *Proc. of the 6th CIKM Conference*, pages 1–8, Las Vegas, Nevada, 1997.
- [43] R. Baeza-Yates and G. Navarro. Faster approximate string matching. *Algorithmica*, 23(2):127–158, 1999.
- [44] R. Baeza-Yates, G. Navarro, J. Vegas, and P. de la Fuente. A model and a visual query language for structured text. In Berthier Ribeiro-Neto, editor, *Proc. of the 5th Symposium on String Processing and Information Retrieval*, pages 7–13, Santa Cruz, Bolivia, Sept 1998. IEEE CS Press.
- [45] R. Baeza-Yates and C. Perleberg. Fast and practical approximate pattern matching. *Information Processing Letters*, 59:21–27, 1996.
- [46] Ricardo Baeza-Yates. Visualizing large answers in text databases. In *Int. Workshop on Advanced User Interfaces*, pages 101–107, Gubbio, Italy, May 1996. ACM Press.
- [47] Ricardo A. Baeza-Yates, Walter Cunto, Udi Manber, and Sun Wu. Proximity matching using fixed queries trees. In M. Crochemore and D. Gusfield, editors, *5th Combinatorial Pattern Matching*, number 807 in LNCS, pages 198–212. Springer-Verlag, Asilomar, CA, June 1994.
- [48] J. Baker. UCLA-NSF Social Aspects of Digital Libraries Workshop, January 1996. <http://www.gslis.ucla.edu/DL/>.
- [49] Y. Bakos and E. Brynjolfsson. Bundling Information Goods: Pricing, Profits, and Efficiency. Technical report, MIT Center for Coordination Science, 1997.
- [50] M. Q. W. Baldonado, C.-C. K. Chang, L. Gravano, and A. Paepcke. Metadata for digital libraries: architecture and design rationale. In *Proc. of the 2nd ACM International Conference on Digital Libraries*, pages 47–56, Philadelphia, PA, USA, 1997.
- [51] Michelle Q. Wang Baldonado and Terry Winograd. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proc. of*

- the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 11–18, Atlanta, GA, USA, May 1997.
- [52] Michelle Q. Wang Baldonado and Terry Winograd. Hi-cites: Dynamically-created citations with active highlighting. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 408–415, Los Angeles, CA, March 1998.
 - [53] D. Ballard and C. Brown. *Computer Vision*. Prentice Hall, 1982.
 - [54] Liam Bannon, Allen Cypher, Steven Greenspan, and Melissa L. Monty. Evaluation and analysis of users' activity organization. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 54–57, 1983.
 - [55] D. Barber. OhioLINK: A Consortial Approach to Digital Library Management. *D-Lib Magazine*, April 1997. <http://www.dlib.org/dlib/april97/04barber.html>.
 - [56] E. Barbosa, G. Navarro, R. Baeza-Yates, C. Perleberg, and N. Ziviani. Optimized binary search and text retrieval. In P. Spirakis, editor, *Proc. of European Symposium on Algorithms*, number 979 in LNCS, pages 311–326. Springer-Verlag, 1995. Improved version to appear in *Algorithmica* including W. Cunto as author.
 - [57] R. Barbosa. *Query Performance on Distributed Digital Libraries*. CS Department, Federal University of Minas Gerais, Brazil, 1998. Master Thesis. In Portuguese.
 - [58] P. Barford, A. Bestavros, A. Bradley, and M. E. Crovella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web*, 1999. Special Issue on Characterization and Performance Evaluation, to appear.
 - [59] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 151–160, July 1998.
 - [60] J. Barnett. Computational methods for a mathematical theory of evidence. In *Proc. of the 7th Int. Joint Conference on Artificial Intelligence*, pages 868–875, Vancouver, Canada, 1981.
 - [61] Jeff Barry, Dania Bilal, and W. David Penniman. Automated system marketplace: the competitive struggle. *Library Journal*, 123(6):43–52, 1998.
 - [62] Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. Latent semantic indexing is an optimal special case of multidimensional scaling. In *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 161–167, Copenhagen, Denmark, 1992.
 - [63] Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. Automatic combination of multiple ranked retrieval systems. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 173–181, Dublin, Ireland, 1994.
 - [64] Chumki Basu, Haym Hirsh, and William Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proc. of AAAI*, pages 714–720, Madison, WI, July 1998.
 - [65] Marcia J. Bates. The design of browsing and berrypicking techniques for the on-line search interface. *Online Review*, 13(5):407–431, 1989.
 - [66] Marcia J. Bates. Where should the person stop and the information search interfaces start? *Information Processing & Management*, 26(5):575–591, 1990.
 - [67] Marcia J. Bates. Document familiarity, relevance, and bradford's law: The Getty online searching project report no. 5. *Information Processing & Management*, 32:697–707, 1996.
 - [68] Micheline Beaulieu and Christine L. Borgman, editors. *Journal of the American Society for Information Science*, volume 47(7). John Wiley & Sons, 1996.

- [69] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An efficient and robust access method for points and rectangles. *ACM SIGMOD*, pages 322–331, May 1990.
- [70] B. Bederson, L. Stead, and J. Hollan. Pad++: Advances in multiscale interfaces. In *Proc. of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 315–316, Boston, MA, USA, 1994. See this and other papers at <http://www.cs.umd.edu/hcil/pad++/papers/>.
- [71] Benjamin B. Bederson, James D. Hollan, Ken Perlin, Jonathan Meyer, David Bacon, and George Furnas. Pad++: A zoomable graphical sketchpad for exploring alternate interface physics. *Journal of Visual Languages and Computing*, 7(1):3–31, 1996.
- [72] C. Beeri and Y. Kornatzky. A logical query language for hypertext systems. In *Proc. of the European Conference on Hypertext*, pages 67–80. Cambridge University Press, 1990.
- [73] Rick Belew. Rave reviews: Acquiring relevance assessments from multiple users. In Marti A. Hearst and Haym Hirsh, editors, *Working Notes of the AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, CA, March 1996.
- [74] N. Belkin and B. Croft. Information filtering and information retrieval: Two sides of a same coin? *Communications of the ACM*, 35(12):29–38, December 1992.
- [75] N. Belkin, P. G. Marchetti, and C. Cool. Braque – design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.
- [76] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3):431–448, May–June 1995.
- [77] T. C. Bell, A. Moffat, C. Nevill-Manning, I. H. Witten, and J. Zobel. Data compression in full-text retrieval systems. *Journal of the American Society for Information Science*, 44:508–531, 1993.
- [78] T. C. Bell, J. Cleary, and I. H. Witten. *Text Compression*. Prentice-Hall, 1990.
- [79] Alberto Belussi and Christos Faloutsos. Estimating the selectivity of spatial queries using the ‘correlation’ fractal dimension. In *Proc. of VLDB Conf.*, pages 299–310, Zurich, Switzerland, September 1995.
- [80] A. B. Benítez, M. Beigi, and S-F. Chang. Using relevance feedback in content-based image metasearch. *IEEE Internet Computing*, 2(4):59–69, 1998.
- [81] J. Bentley, D. Sleator, R. Tarjan, and V. Wei. A locally adaptive data compression scheme. *Communications of the ACM*, 29:320–330, 1986.
- [82] S. Berchtold, C. Boehm, B. Braunmueller, D. A. Keim, and H. P. Kriegel. Fast similarity search in multimedia databases. In *Proc. of the ACM SIGMOD Int’l Conf. on Management of Data*, pages 1–12, 1997.
- [83] Stefan Berchtold, Daniel A. Keim, and Hans-Peter Kriegel. The X-tree : An index structure for high-dimensional data. *VLDB*, pages 28–39, 1996.
- [84] T. Berners-Lee. The World Wide Web Consortium. <http://www.w3.org>.
- [85] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The World-Wide Web. *Communication of the ACM*, 37(8):76–82, 1994.
- [86] J. W. Berry. Digital libraries: new initiatives with world wide implications. *IFLA Journal*, 22(1):9–17, 1995. <http://www.nlc-bnc.ca/ifla/IV/ifla61/61-berjo.htm>.
- [87] E. Bertino and E. Ferrari. Temporal synchronization models for multimedia data. *IEEE Trans. on Knowledge and Data Eng.*, 10(4):612–631, July/August 1998.

- [88] E. Bertino, E. Ferrari, and M. Stolf. MPGS: An interactive tool for the specification and generation of multimedia presentations. *IEEE Transactions on Multimedia Systems*, 1999. To appear.
- [89] E. Bertino and L. D. Martino. *Object-Oriented Database Systems — Concepts and Architecture*. Addison-Wesley, 1993.
- [90] E. Bertino and F. Rabitti. Query Processing in MULTOS. In C. Thanos, editor, *Multimedia Office Filing: The MULTOS Approach*, pages 273–295. Elsevier Science Publishers B.V. (North Holland), 1990.
- [91] K. Bharat and A. Z. Broder. A technique for measuring the relative size and overlap of public Web search engines. In *7th WWW Conference*, pages 379–388, Brisbane, Australia, 1998.
- [92] Krishna Bharat, Andrei Broder, Monika Henzinger, Puneet Kumar, and Suresh Venkatasubramanian. The connectivity server: fast access to linkage information on the Web. In *7th WWW Conf.*, Brisbane, Australia, April 1998.
- [93] Krishna Bharat and Monika Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *21st ACM SIGIR*, pages 104–111, Melbourne, Australia, 1998.
- [94] J. Bichteler and E. A. Eaton III. The combined use of bibliographic coupling and co-citation for document retrieval. *Journal of the American Society for Information Science*, 31(4), July 1980.
- [95] Eric A. Bier, Maureen C. Stone, Ken Pier, Ken Fishkin, Thomas Baudel, Matt Conway, William Buxton, and Tony DeRose. Toolglass and magic lenses: The see-through interface. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 2 of *Videos: Part II*, pages 445–446, Boston, MA, USA, 1994.
- [96] Patricia A. Billingsley. Taking panes: Issues in the design of windowing systems. In Martin Helander, editor, *Handbook of Human-Computer Interaction*, pages 413–436. Springer Verlag, 1988.
- [97] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the University of Michigan Digital Library. *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995. Stanford, CA, AAAI Press.
- [98] A. P. Bishop. Working towards an understanding of digital library use: A report on the user research efforts of the NSF/ARPA/NASA DLI projects. *D-Lib Magazine*, October 1995. <http://www.dlib.org/dlib/october95/10bishop.html>.
- [99] K. Bohm and T. C. Rakov. Metadata for multimedia documents. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 21–26, Minneapolis, MN, USA, May 1994.
- [100] Philippe Bonnet and Anthony Tomasic. Partial answers for unavailable data sources. In *Workshop on Flexible Query-Answering Systems*, pages 43–54, 1998.
- [101] A. Bookstein. On the perils of merging Boolean and weighted retrieval systems. *Journal of the American Society for Information Sciences*, 29(3):156–158, 1978.
- [102] A. Bookstein. Fuzzy requests: An approach to weighted Boolean searches. *Journal of the American Society for Information Sciences*, 31:240–247, 1980.
- [103] A. Bookstein. Implication of Boolean structure for probabilistic retrieval. In *Proc. of the 8th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 11–17, Montreal, Canada, 1985.
- [104] Christine L. Borgman. Why are online catalogs hard to use? Lessons learned from information retrieval studies. *Journal of the American Society for Information Science*, 37(6):387–400, 1986.

- [105] Christine L. Borgman. Why are online catalogs *still* hard to use? *Journal of the American Society for Information Science*, 47(7):493–503, 1996.
- [106] J. Bosak. XML, Java, and the future of the Web. Technical report, Sun Microsystems, 1997. <http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm>.
- [107] Charles P. Bourne. On-line systems: history, technology, and economics. *Journal of the American Society for Information Science*, 31:155–160, 1980.
- [108] C. Mic Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber, and Michael F. Schwartz. The Harvest information discovery and access system. In *Proc. 2nd Int. WWW Conf.*, pages 763–771, October 1994.
- [109] George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, USA, 3rd edition, 1994.
- [110] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Communications of the ACM*, 20(10):762–772, 1977.
- [111] James Boyle, William Ogden, Steven Uhler, and Patricia Wilson. QMF usability: How it really happened. In *Proc. of IFIP INTERACT'84: Human-Computer Interaction*, pages 877–882, 1984.
- [112] Tolga Bozkaya and Z. Meral Ozsoyoglu. Distance-based indexing for high-dimensional metric spaces. In *Proc. of ACM SIGMOD Conference*, pages 357–368, Tucson, AZ, USA, 1997.
- [113] P. M. E. De Bra and R. D. J. Post. Searching for arbitrary information in the WWW: The fish search for Mosaic. In *Proc. of the 2nd Int. WWW Conference*, Chicago, October 1994. <http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/www-fall94.html>.
- [114] T. Bray. Measuring the web. In *Proc. of the 5th Int. WWW Conference*, Paris, May 1996. http://www5conf.inria.fr/fich_html/papers/P9/Overview.html.
- [115] S. Brin. Extracting patterns and relations from the World Wide Web. In *Workshop on Web Databases*, Valencia, Spain, March 1998.
- [116] Sergei Brin. Near neighbor search in large metric spaces. In *Proc. of VLDB Conf.*, pages 574–584, Zurich, Switzerland, Sept 1995.
- [117] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. of the 7th Int. WWW Conference*, Brisbane, Australia, April 1998.
- [118] T. Brinkhoff, H. P. Kriegel, and R. Schneider. Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems. In *Proc. of the IEEE Conference on Data Engineering*, pages 40–49, Vienna, Austria, 1993.
- [119] Thomas Brinkhoff, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. Multi-step processing of spatial joins. In *Proc. of ACM SIGMOD*, pages 197–208, Minneapolis, MN, USA, May 1994.
- [120] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the Web. In *6th Int. WWW Conference*, pages 391–404, Santa Clara, CA, USA, April 1997.
- [121] A. Z. Broder. On the resemblance and containment of documents. In *Conf. on Compression and Complexity of Sequences*, pages 21–29, Salerno, Italy, 1997. IEEE Computer Society.
- [122] J. Broglio, J. P. Callan, W. B. Croft, and D. W. Nachbar. Document retrieval and routing using the INQUERY system. In D. K. Harman, editor, *Overview of the Third Retrieval Conference (TREC-3)*, pages 29–38. NIST Special Publication 500-225, 1995.
- [123] Eric W. Brown. *Execution Performance Issues in Full-Text Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, 1996. Available as UMass Comp. Sci. Tech. Rep. TR95-81.

- [124] J. S. Brown and P. Duguid. The social life of documents. *First Monday*, May 1996. <http://www.firstmonday.dk/issues/issue1/documents/>.
- [125] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.
- [126] M. Buckland and C. Plaunt. On the Construction of Selection Systems. *Library Hi Tech*, 12:15–28, 1994.
- [127] Chris Buckley, Gerard Salton, and James Allan. The effect of adding relevance information in a relevance feedback environment. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, Dublin, Ireland, 1994.
- [128] P. Buneman, S. B. Davidson, K. Hart, C. Overton, and L. Wong. A data transformation system for biological data sources. In *Proc. of the 21st International Conference on Very Large Data Bases*, pages 158–169, Zurich, Switzerland, September 1995.
- [129] Peter Buneman. Semistructured data. In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 117–121, Tucson, Arizona, 1997.
- [130] Robin Burke, Kristian Hammond, Vladimir Kulukin, Steven Lytinen, Noriko Tomuro, and Scott Schoenberg. Experiences with the FAQ system. *AI Magazine*, 18(2):57–66, 1997.
- [131] W. A. Burkhard and R. M. Keller. Some approaches to best-match file searching. *Communications of the ACM*, 16(4):230–236, April 1973.
- [132] F. Burkowski. An algebra for hierarchically organized text-dominated databases. *Information Processing & Management*, 28(3):333–348, 1992.
- [133] F. Burkowski. Retrieval activities in a database consisting of heterogeneous collections of structured text. In *Proc. of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 112–125, Copenhagen, Denmark, 1992.
- [134] Arthur W. Burks, Herman H. Goldstine, and John von Neumann. Preliminary discussion of the logical design of an electronic computing instrument. In William Aspray and Arthur Burks, editors, *Papers of John von Neumann on Computers and Computer Theory*, pages 97–142. The MIT Press, Cambridge, MA, 1987 (originally appeared in 1946).
- [135] V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, July 1945.
- [136] D. Byers. Full-text indexing of non-textual resources. In *7th WWW Conf.*, Brisbane, Australia, April 1998.
- [137] Brendon Cahoon and Kathryn McKinley. Performance evaluation of a distributed architecture for information retrieval. In *Proc. 19th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 110–118, Zurich, Switzerland, August 1996.
- [138] J. Callan. Document filtering with inference networks. In *Proc. of the 19th ACM SIGIR Conference*, pages 262–269, Zurich, Switzerland, August 1996.
- [139] J. P. Callan, Z. Lu, and W. Bruce Croft. Searching Distributed Collections with Inference Networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, July 1995. ACM Press.
- [140] S. K. Card, M. Pavel, and J. E. Farrell. Window-based computer dialogues. In *Proc. of Interact '84, First IFIP Conference on Human-Computer Interaction*, pages 355–359, 1984.
- [141] Stuart K. Card, Jock D. MacKinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, 1998.

- [142] Stuart K. Card, Thomas P. Moran, and Allen Newell. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates, 1983.
- [143] Stuart K. Card, George G. Robertson, and Jock D. Mackinlay. The information visualizer, an information workspace. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 181–188, 1991.
- [144] Stuart K. Card, George G. Robertson, and William York. The WebBook and the Web Forager: An information workspace for the World-Wide Web. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 111–117, Zurich, Switzerland, April 1996.
- [145] S. Carliner. Designing wizards. *Training & Development*, 52(7):62–63, 1998.
- [146] Carnegie Mellon University DLI Team, Informedia. <http://www.informedia.cs.cmu.edu/>.
- [147] Claudio Carpineto and Giovanni Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45(5):553–578, 1996.
- [148] J. Carriere and R. Kazman. WebQuery: Searching and visualizing the Web through connectivity. In *6th Int. WWW Conf.*, Santa Clara, CA, USA, April 1997.
- [149] M. Castagli and S. Eubank. *Nonlinear Modeling and Forecasting*. Addison-Wesley, 1992.
- [150] R. Cattell. *The Object Database Standard: ODMG-93*. Morgan Kauffman, 1993.
- [151] W. B. Cavnar and A. M. Gillies. Data retrieval and the realities of document conversion. In *Proc. of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, College Station, TX, USA, June 1994.
- [152] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *VLDB Journal*, 7(3):163–178, 1998.
- [153] S. Chakrabarti, B. Dom, D. Gibson, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Experiments in topic distillation. In *ACM-SIGIR'98 Post Conference Workshop on Hypertext Information Retrieval for the Web*, Melbourne, Australia, 1998.
- [154] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *7th Int. WWW Conference*, pages 65–74, Brisbane, Australia, April 1998.
- [155] M. Chalmers, K. Rodden, and D. Brodbeck. The order of things: activity-centred information access. In *7th Int. WWW Conference*, Brisbane, Australia, April 1998.
- [156] Matthew Chalmers and Paul Chitson. Bead: Exploration in information visualization. In *Proc. of the 15th Annual International ACM/SIGIR Conference*, pages 330–337, Copenhagen, Denmark, 1992.
- [157] C-H. Chang and C-C. Hsu. Customizable multi-engine search tool with clustering. In *6th. Int. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [158] S.-F. Chang, J. R. Smith, H. J. Meng, H. Wang, and D. Zhong. Finding Images/Video in Large Archives: Columbia's Content-Based Visual Query Project. *D-Lib Magazine*, February 1997. <http://www.dlib.org/dlib/february97/columbia/02chang.html>.
- [159] Shan-Ju Chang and Ronald E. Rice. Browsing: A multidimensional framework. *Annual Review of Information Science and Technology*, 28:231–276, 1993.
- [160] S. Chapman and A. R. Kenney. Digital conversion of library research materials: A case for full informational capture. *D-Lib Magazine*, October 1996. <http://www.dlib.org/dlib/october96/cornell/10chapman.html>.

- [161] S. Chawathe, H. García-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *Proc. of IPSJ Conference*, pages 7–18, October 1994.
- [162] C. Chen. Structuring and visualizing the WWW by generalized similarity analysis. In *Proc. of the 8th ACM Conference on Hypertext and Hypermedia*, pages 177–186, Southampton, England, 1997.
- [163] Hsinchun Chen, Andrea L. Houston, Robin R. Sewell, and Bruce R. Schatz. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7):582–608, 1998.
- [164] Michael Chen and Marti A. Hearst. Presenting Web site search results in context: A demonstration. In *Proc. of the 20th Annual International ACM/SIGIR Conference*, page 381, Melbourne, Australia, 1998.
- [165] Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, and David A. Pater-son. RAID: High-performance, reliable secondary storage. *ACM Computing Surveys*, 26(2):145–185, June 1994.
- [166] F. Cheong. *Internet Agents, Spiders, Wanderers, Brokers and Bots*. New Riders, 1996.
- [167] D. R. Chestnutt. The model editions partnership: ‘Smart Text’ and beyond. *D-Lib Magazine*, July/August 1997. <http://www.dlib.org/dlib/july97/07chesnutt.html>.
- [168] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. In *7th WWW Conference*, Brisbane, Australia, April 1998.
- [169] S. Christodoulakis, M. Theodoridou, F. Ho, M. Papa, and A. Pathria. Multimedia document presentation, information extraction, and document formation in MINOS: a model and a system. *IEEE Transactions on Office Information Systems*, 4(4):345–383, October 1986.
- [170] W. W. Chu, I. T. Jeong, R. K. Taira, and C. M. Breant. A Temporal Evolutionary Object-Oriented Model and its Query Languages for Medical Image Management. In *Proc. of the Int. Conference on Very Large Data Bases*, pages 53–64, Vancouver, Canada, Aug 1992.
- [171] K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.
- [172] Paolo Ciaccia, Marco Patella, and Pavel Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proc. of VLDB Conf.*, pages 426–435, Athens, Greece, August 1997.
- [173] C. Clarke, G. Cormack, and F. Burkowski. An algebra for structured text search and a framework for its implementation. *The Computer Journal*, 38(1):43–56, 1995.
- [174] Charles L. A. Clarke, Gordon V. Cormack, and Forbes J. Burkowski. Shortest substring ranking (multitext experiments for TREC-4). In Donna Harman, editor, *Proc. of the Fourth Text Retrieval Conference TREC-4*. National Institute of Standards and Technology Special Publication, 1996.
- [175] E. G. Coffman, Z. Liu, and R. R. Weber. Optimal robot scheduling for Web search engines. Technical Report 3317, INRIA, France, December 1997.
- [176] Douglas E. Comer and David L. Stevens. *Internetworking with TCP/IP Vol III: Client-Server Programming and Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1993.
- [177] *Communications of the ACM*, 37(2), February 1994.
- [178] *Communications of the ACM*, 38(8), August 1995.
- [179] B. Commentz-Walter. A string matching algorithm fast on the average. In *Proc. ICALP’79*, pages 118–132. Springer-Verlag, 1979.

- [180] Compaq: Junglee shopping guide. <http://www.compaq.junglee/compaq/top.html>, 1998.
- [181] Jeff Conklin. Hypertext: An introduction and survey. *IEEE Computer*, 20(9):17–41, September 1987.
- [182] Dan Connolly on the architecture of the Web: Let a thousand flowers bloom. *IEEE Internet Computing*, 2(2):22–31, 1998.
- [183] M. Consens and T. Milo. Algebras for querying text regions. In *Proc. of the ACM Conference on Principles of Distributed Systems*, pages 11–22, San Jose, California, USA, 1995.
- [184] M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in Graphlog. In *Hypertext'89*, pages 269–292, 1989.
- [185] Alan Cooper. *About Face: The Essentials of User Interface Design*. IDG Books, 1995.
- [186] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 198–210, Copenhagen, Denmark, 1992.
- [187] Wm. S. Cooper. The formalism of probability theory in IR: A foundation or an encumbrance? In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 242–247, Dublin, Ireland, 1994. Triennial ACM-SIGIR Award Paper.
- [188] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press/McGraw-Hill, Cambridge, MA, 1990.
- [189] Corporation for National Research Initiatives, The Handle System, May 1998. <http://www.handle.net/>.
- [190] P. Cotton, editor. *ISO Working Draft SQL Multimedia and Application Packages (SQL/MM), Part II Full text*. ISO/IEC, 1993. SC21/WG3 N1631, SQL/MM MUN-004.
- [191] S. B. Cousins, S. P. Ketchpel, A. Paepcke, H. García-Molina, S. W. Hassan, and M. Roscheisen. InterPay: Managing multiple payment mechanisms in digital libraries. In *Proc. of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*, Austin, Texas, June 1995.
- [192] S. B. Cousins, A. Paepcke, T. Winograd, E. A. Bier, and K. Pier. The digital library integrated task environment (DLITE). In *Proc. of the 2nd ACM International Conference on Digital Libraries*, pages 142–151, Philadelphia, PA, USA, July 1997.
- [193] Steve B. Cousins. *Reification and Affordances in a User Interface for Interacting with Heterogeneous Distributed Applications*. PhD thesis, Stanford University, May 1997.
- [194] B. Cox, J. D. Tygar, and M. Sirbu. NetBill Security and Transaction Protocol. In *Proc. of the 1st USENIX Workshop on Electronic Commerce*, New York, USA, July 1995.
- [195] S. Y. Crawford, J. M. Hurd, and A. C. Weller. *From Print to Electronic: the Transformation of Scientific Communication*. American Society for Information Science Monograph Series, Medford, New Jersey, 1996.
- [196] M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, Oxford, UK, 1994.
- [197] W. B. Croft. What do people want from information retrieval? (The top 10 research issues for companies that use and sell IR systems). *D-Lib Magazine*, November 1995. <http://www.dlib.org/dlib/november95/11croft.html>.
- [198] W. B. Croft. Experiments with representation in a document retrieval system. *Information Technology: Research and Development*, 2(1):1–21, 1983.
- [199] W. B. Croft and D. J. Harper. Using probabilistic models of retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.

- [200] Carolyn J. Crouch and Bokyung Yang. Experiments in automatic statistical thesaurus construction. In *Proc. of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88, Copenhagen, Denmark, 1992.
- [201] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. In *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 160–169, May 1996.
- [202] Douglass R. Cutting, David Karger, and Jan Pedersen. Constant interaction-time Scatter/Gather browsing of very large document collections. In *Proc. of the 16th Annual International ACM/SIGIR Conference*, pages 126–135, Pittsburgh, PA, 1993.
- [203] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. of the 15th Annual International ACM/SIGIR Conference*, pages 318–329, Copenhagen, Denmark, 1992.
- [204] Allen Cypher, editor. *Watch What I Do: Programming by Demonstration*. MIT Press, Cambridge, 1993.
- [205] A. Czumaj, M. Crochemore, L. Gasieniec, S. Jarominek, Thierry Lecroq, W. Plandowski, and W. Rytter. Speeding up two string-matching algorithms. *Algorithmica*, 12:247–267, 1994.
- [206] T. Dao, R. Sacks-Davis, and J. Thom. Indexing structured text for queries on containment relationships. In *Proc. of 7th Australasian Database Conference*, Melbourne, Australia, Jan/Feb 1996.
- [207] Rick Darnell. *HTML 4.0 Unleashed, Professional Reference Edition*. Samms.net Publishing, 1998.
- [208] M. Dartois, A. Maeda, T. Sakaguchi, T. Fujita, S. Sugimoto, and K. Tabata. A multilingual electronic text collection of folk tales for casual users using off-the-shelf browsers. *D-Lib Magazine*, October 1997. <http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html>.
- [209] S. Davidson, C. Overton, V. Tannen, and L. Wong. BioKleisli: A digital library for biomedical researchers. *International Journal of Digital Libraries*, 1(1):35–53, April 1997.
- [210] J. R. Davis and C. Lagoze. A protocol and server for a distributed technical report library. Technical report, Cornell University Computer Science Department, June 1994.
- [211] Y. F. Day, S. Dagtas, M. Iino, A. Khokhar, and A. Ghafoor. Object-oriented conceptual modeling of video data. In *Proc. of the IEEE Conference on Data Engineering*, pages 401–405, Taipei, Taiwan, 1995.
- [212] DejaNews: Main Page. <http://www.dejanews.com>, 1998.
- [213] Alin Deutsch, Mary Fernández, Daniela Florescu, Alon Levy, and Dan Suciu. A query language for XML. <http://www.research.att.com/~mff/xml/w3cnote.html>, 1998.
- [214] A. Dillon. What is the shape of information? Human factors in the development and use of digital libraries. Allerton discussion document submitted for the 1995 Allerton Institute. <http://edfu.lis.uiuc.edu/allerton/95/s4/dillon.html>.
- [215] DirectHit: Main Page. <http://www.directhit.com>, 1998.
- [216] Division of Information and Intelligent Systems, National Science Foundation. Digital Libraries Initiative – Phase 2. Announcement Number NSF 98-63, 1998. <http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm>.
- [217] M. Dodge. The geography of cyberspace directory: Main page. http://www.geog.ucl.ac.uk/casa/martin/geography_of_cyberspace.html, 1997.

- [218] R. Dolin, D. Agrawal, A. El Abbadi, and J. Pearlman. Using automated classification for summarizing and selecting heterogeneous information sources. *D-Lib Magazine*, January 1998. <http://www.dlib.org/dlib/january98/dolin/01dolin.html>.
- [219] R. Dolin, D. Agrawal, L. Dillon, and A. El Abbadi. Pharos: a scalable distributed architecture for locating heterogeneous information sources. In *Proc. of the 6th CIKM Conference*, pages 348–355, Las Vegas, Nevada, 1997.
- [220] P. Dömel. Webmap: A graphical hypertext navigation tool. In *Proc. of the 2nd. WWW Conference*, 1994. <http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/-WWW2.Proceedings.html>.
- [221] P. Doty and A. P. Bishop. The national information infrastructure and electronic publishing: A reflective essay. *Journal of the American Society for Information Science*, 45(10):785–799, 1994.
- [222] Karen M. Drabenstott and Marjorie S. Weller. The exact-display approach for online catalog subject searching. *Information Processing and Management*, 32(6):719–745, 1996.
- [223] D. Dreilinger. Savvy Search, 1996. <http://savvy.cs.colostate.edu:2000/form?beta>.
- [224] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.
- [225] Mark D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *ACM Transactions on Information Systems*, 15(2):137–153, 1997.
- [226] Deborah M. Edwards and Lynda Hardman. ‘Lost in hyperspace’: Cognitive mapping and navigation in a hypertext environment. In Ray McAleese, editor, *Hypertext I: Theory into Practice*, pages 105–125. Ablex Publishing Corporation, 1988.
- [227] Dennis E. Egan. Individual differences in human-computer interaction. In Martin Helander, editor, *Handbook of Human-Computer Interaction*, pages 543–568. Springer Verlag, 1988.
- [228] Dennis E. Egan, Michael E. Lesk, R. Daniel Ketchum, Carol C. Lochbaum, Joel R. Remde, Louis M. Gómez, and Thomas K. Landauer. Hypertext for the electronic library? CORE sample results. In *Proc. of the ACM Hypertext Conference*, pages 299–312, May 1991.
- [229] Dennis E. Egan, Joel R. Remde, Louis M. Gomez, Thomas K. Landauer, Jennifer Eberhardt, and Carol C. Lochbaum. Formative design evaluation of SuperBook. *Transactions on Information Systems*, 7(1):30–57, 1989.
- [230] Dennis E. Egan, Joel R. Remde, Thomas K. Landauer, Carol C. Lochbaum, and Louis M. Gomez. Behavioral evaluation and analysis of a hypertext browser. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 205–210, May 1989.
- [231] M. J. Egenhofer. Extending SQL for Graphical Display. *Cartography and Geographic Information Systems*, 18(4):230–245, 1991.
- [232] Stephen G. Eick. Graphically displaying text. *Journal of Computational and Graphical Statistics*, 3(2):127–142, June 1994.
- [233] Stephen G. Eick and Graham J. Wills. High interaction graphics. *European Journal of Operations Research*, 81(3):445–459, March 1995.
- [234] Stephen G. Eick and Graham J. Wills. Navigating large networks with hierarchies. In *Proc. of the Conference on Visualization '93*, pages 204–209, San Jose, October 1993.
- [235] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21:194–203, 1975.
- [236] D. Ellis. A behavioural model for information retrieval system design. *Journal of Information Science*, 15:237–247, 1989.

- [237] R. Entlich, L. Garson, M. Lesk, L. Normore, J. Olsen, and S. Weibel. Making a digital library: The Chemistry Online Retrieval Experiment – A summary of the CORE project (1991-1995). *D-Lib Magazine*, December 1995. <http://www.dlib.org/dlib/december95/briefings/12core.html>.
- [238] M. L. Escobar-Molano and S. Ghandeharizadeh. A Framework for Conceptualizing Structured Video. In *Proc. First Int. Workshop on Multimedia Information Systems*, pages 95–110, Arlington, Virginia, September 1995.
- [239] ERCIM Digital Library Working Group, European Research Consortium for Informatics and Mathematics, June 1998. <http://www.area.pi.cnr.it/ErcimDL/>.
- [240] Excite: Main page. <http://www.excite.com>, 1995.
- [241] Express: Main Page. <http://www.express.infoseek.com>, 1998.
- [242] C. Faloutsos and R. Chan. Text access methods for optical and large magnetic disks: design and performance comparison. In *Proc. of VLDB'88*, pages 280–293, Los Angeles, CA, USA, 1988.
- [243] C. Faloutsos and S. Christodoulakis. Description and performance analysis of signature file methods. *ACM TOIS*, 5(3):237–257, 1987.
- [244] Christos Faloutsos, Ron Barber, Myron Flickner, J. Hafner, Wayne Niblack, Dragutin Petkovic, and William Equitz. Efficient and effective querying by image content. *J. of Intelligent Information Systems*, 3(3/4):231–262, July 1994.
- [245] Christos Faloutsos and Volker Gaede. Analysis of n -dimensional quadrees using the Hausdorff fractal dimension. In *Proc. of VLDB Conf.*, pages 40–50, Bombay, India, September 1996.
- [246] Christos Faloutsos, H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo. A signature technique for similarity-based queries. In *SEQUENCES'97*, Salerno, Italy, June 1997.
- [247] Christos Faloutsos and Ibrahim Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *Proc. ACM SIGACT-SIGMOD-SIGART PODS*, pages 4–13, Minneapolis, MN, May 1994.
- [248] Christos Faloutsos and King-Ip (David) Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *Proc. of ACM SIGMOD*, pages 163–174, May 1995.
- [249] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. In *Proc. ACM SIGMOD*, pages 419–429, Minneapolis, MN, USA, May 1994. ‘Best Paper’ award.
- [250] M. Farach and M. Thorup. String matching in Lempel-Ziv compressed strings. In *Proc. 27th ACM Annual Symposium on the Theory of Computing*, pages 703–712, Las Vegas, NE, USA, 1995.
- [251] I. A. Ferguson and M. J. Wooldridge. Paying their way: Commercial digital libraries for the 21st century. *D-Lib Magazine*, June 1997. <http://www.dlib.org/dlib/june97/zuno/06ferguson.html>.
- [252] Terry Ellen Ferl and Larry Millsap. The knuckle-cracker’s dilemma: a transaction log study of OPAC subject searching. *Information Technology and Libraries*, 15:81–98, 1996.
- [253] Mary Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu. A query language for a Web-site management system. *SIGMOD Record*, 26(3):4–11, September 1997.
- [254] E. Ferrari and G. Haus. Designing music objects for a multimedia database. In *Proc. 12th Colloquium on Musical Informatics*, pages 135–138, Gorizia, Italy, September 1998.
- [255] Gerhard Fischer and Helga Nieper-Lemke. Helgon: Extending the retrieval reformulation paradigm. In *Proc. of ACM CHI Conference on Human Factors in Computing Systems*, pages 357–362, 1989.

- [256] Ken Fishkin and Maureen C. Stone. Enhanced dynamic queries via movable filters. In *Proc. of ACM CHI Conference on Human Factors in Computing Systems*, volume 1 of *Papers: Information Visualization*, pages 415–420, Denver, CO, USA, 1995.
- [257] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jon Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, September 1995.
- [258] D. Florescu, A. Levy, and A. Mendelzon. Database techniques for the World-Wide Web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.
- [259] M. J. Flynn. Very high-speed computing systems. In *Proc. IEEE*, volume 54, pages 1901–1909, 1966.
- [260] Corporation for National Research Initiatives. Computer Science Technical Reports Project (CSTR), May 1996. <http://www.cnri.reston.va.us/home/cstr.html>.
- [261] D. J. Foskett. Thesaurus. In K. Sparck Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 111–134. Morgan Kaufmann Publishers, Inc., 1997.
- [262] Richard H. Fowler, Wendy A. L. Fowler, and Bradley A. Wilson. Integrating query, thesaurus, and documents through a common visual representation. In *Proc. of the 14th Annual International ACM/SIGIR Conference*, pages 142–151, Chicago, 1991.
- [263] C. Fox. Lexical analysis and stoplists. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 102–130. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [264] E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a Modern OPAC: From REVTOLC to MARIAN. In *Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 248–259, Pittsburgh, PA, June/July 1993.
- [265] E. Fox and L. Lunin. Introduction and overview to perspectives on digital libraries: Guest editor's introduction to special issue. *Journal of the American Society for Information Science*, 44(8):441–443, 1993.
- [266] E. A. Fox. Source book on digital libraries. Technical Report TR-93-35, Virginia Polytechnic Institute and State University, 1993.
- [267] E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett. Digital libraries. *Communications of the ACM*, 38(4):22–28, April 1995.
- [268] E. A. Fox and R. Gupta. Courseware on Digital Libraries. <http://ei.cs.vt.edu/~dlib/>.
- [269] E. A. Fox, L. S. Heath, and D. Hix. Project Envision Final Report: A user-centered database from the computer science literature, July 1995. <http://ei.cs.vt.edu/papers/ENVreport/final.html>.
- [270] E. A. Fox and G. Marchionini. Toward a worldwide digital library. *Communications of the ACM*, 41(4):29–32, April 1998. <http://purl.lib.vt.edu/dlib/pubs/CACM199804>.
- [271] E. A. Fox. *Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types*. PhD thesis, Cornell University, Ithaca, New York, <Http://www.ncstrl.org>, 1983.
- [272] Edward A. Fox. Characterization of two new experimental collections in computer and information science containing textual and bibliographical concepts. *Technical Report*, 83-561, 1983. <Http://www.ncstrl.org>.
- [273] Edward A. Fox, Deborah Hix, Lucy T. Nowell, Dennis J. Brueni, William C. Wake, Lenwood S. Heath, and Durgesh Rao. Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, 44(8):480–491, 1993.

- [274] W. Frakes. Stemming algorithms. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 131–160. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [275] W. B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures & Algorithms*. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [276] W. Francis and H. Kucera. *Frequency Analysis of English Usage*. Houghton Mifflin Co., 1982.
- [277] Eric Freeman and Scott Fertig. Lifestreams: Organizing your electronic life. In Robin Burke, editor, *Working Notes of the AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, Cambridge, MA, November 1995.
- [278] H. P. Frei and D. Stieger. The use of semantic links in hypertext information retrieval. *Information Processing & Management*, 31(1):1–13, 1994.
- [279] J. French, E. Fox, K. Maly, and A. Selman. Wide Area Technical Report Service: Technical Reports Online. *Communications of the ACM*, 38(4):47, April 1995.
- [280] A. Friedlander. D-lib Program: Research in Digital Libraries, May 1998. <http://www.dlib.org/>.
- [281] N. Fuhr. Models for retrieval with probabilistic indexing. *Information Processing & Management*, 25:55–72, 1989.
- [282] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.
- [283] N. Fuhr. DOLORES: A System for Logic-Based Retrieval of Multimedia Objects. In B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 257–265, Melbourne, Australia, 1998.
- [284] Norbert Fuhr. Optimal polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.
- [285] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2nd edition, 1990.
- [286] George W. Furnas and Jeff Zacks. Multitrees: Enriching and reusing hierarchical structure. In *Proc. of ACM CHI Conference on Human Factors in Computing Systems*, volume 2, pages 330–336, 1994.
- [287] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc. of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–480, 1988.
- [288] R. Furuta. Defining and Using Structure in Digital Documents. In *Proc. of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, College Station, TX, USA, 1994.
- [289] Richard Furuta, Frank M. Shipman III, Catherine C. Marshall, Donald Brenner, and Hao wei Hsieh. Hypertext paths and the World-Wide Web: Experiences with Walden's paths. In *Proc. of the Eighth ACM Conference on Hypertext*, pages 167–176, Southampton, England, 1997.
- [290] Volker Gaede and Oliver Günther. Multidimensional access methods. *ACM Computing Surveys*, 30(2):170–231, 1998.
- [291] R. G. Gallager. Variations on a theme by Huffman. *IEEE Transactions on Information Theory*, 24:668–674, 1978.

- [292] S. Gauch, G. Wang, and M. Gómez. ProFusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computing*, 2(9):637–649, September 1996.
- [293] E. Gaynor. From MARC to Markup: SGML and Online Library Systems. *ALCTS Newsletter*, 7(2), 1996.
- [294] Naum Gershon, Joshua LeVasseur, Joel Winstead, James Croall, Ari Pernick, and William Ruh. Visualizing Internet resources. In *Proc. of the '95 Information Visualization*, pages 122–128, Atlanta, October 1995.
- [295] Fredric C. Gey. Inferring probability of relevance using the method of logistic regression. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–231, Dublin, Ireland, 1994.
- [296] S. Gibbs. Composite Multimedia and Active Objects. In *Proc. Int. Conf. on Object-Oriented Programming: Systems, Languages, and Applications*, pages 97–112, October 1991.
- [297] S. Gibbs, C. Breiteneder, and D. Tschritzis. Audio/video databases: an object-oriented approach. In *Proc. 9th Int. Conf. on Data Engineering*, pages 381–390, 1993.
- [298] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web communities from link topologies. In *9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, USA, 1998.
- [299] H. Gladney, E. Fox, Z. Ahmed, R. Ashany, N. Belkin, and M. Zemankova. Digital library: Gross structure and requirements: Report from a March 1994 workshop. In *Proc. of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, pages 101–107, College Station, TX, USA, 1994.
- [300] H. Gladney, F. Mintzer, F. Schiattarella, J. Bescós, and M. Treu. Digital access to antiquities. *Communications of the ACM*, 41(4):49–57, April 1998.
- [301] H. M. Gladney. Safeguarding digital library contents and users: document access control. *D-Lib Magazine*, June 1997. <http://www.dlib.org/dlib/june97/ibm/06gladney.html>.
- [302] H. M. Gladney and J. B. Lotspiech. Safeguarding digital library contents and users: Assuring convenient security and data quality. *D-Lib Magazine*, May 1997. <http://www.dlib.org/dlib/may97/ibm/05gladney.html>.
- [303] Charles Goldfarb. *The SGML Handbook*. Oxford University Press, Oxford, 1990.
- [304] Charles Goldfarb and Paul Prescod. *The XML Handbook*. Prentice Hall, Oxford, 1998.
- [305] Dina Q. Goldin and Paris C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. *Int. Conf. on Principles and Practice of Constraint Programming*, pages 137–153, Sept 1995.
- [306] R. Goldman, J. McHugh, and J. Widom. Lore: A database management system for XML. Technical report, Stanford University Database Group, 1998.
- [307] S. W. Golomb. Run-length encodings. *IEEE Transactions on Information Theory*, 12(3):399–401, 1966.
- [308] G. Golovchinsky. What the query told the link: the integration of Hypertext and Information Retrieval. In *8th ACM Conference on Hypertext and Hypermedia*, pages 67–74, Southampton, England, 1997.
- [309] G. Gonnet. Examples of PAT applied to the Oxford English Dictionary. Technical Report OED-87-02, UW Centre for the New OED and Text Research, Univ. of Waterloo, 1987.
- [310] G. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures*. Addison-Wesley, Wokingham, England, 2nd edition, 1991.

- [311] G. Gonnet, R. Baeza-Yates, and T. Snider. New indices for text: Pat trees and Pat arrays. In William Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, pages 66–82. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [312] J. Gordon and E. Shortliffe. The Dempster-Shafer Theory of Evidence. In *Rule-Based Expert Systems*, pages 113–138. Addison-Wesley Publishing Company, 1984.
- [313] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 37th Allerton Institute 1995, January 1996. <http://edfu.lis.uiuc.edu/allerton/95/>.
- [314] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 38th Allerton Institute, January 1997. <http://edfu.lis.uiuc.edu/allerton/96/>.
- [315] P. Graham. Glossary on Digital Library Terminology, 1997. Informal file sent by electronic mail for comments, available with permission of the author if suitable attribution is made.
- [316] L. Gravano, K. Chang, H. García-Molina, C. Lagoze, and A. Paepcke. STARTS: Stanford protocol proposal for Internet retrieval and search. Technical report, Stanford University, Digital Library Project, 1997. <http://www-db.stanford.edu/~gravano/starts.html>.
- [317] Luis Gravano, Chen-Chuan K. Chang, and Hector García-Molina. STARTS: Stanford proposal for Internet meta-searching. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 207–218, Tucson, AZ, May 1997.
- [318] Luis Gravano, Hector García-Molina, and Anthony Tomasic. The effectiveness of GLOSS for the text-database discovery problem. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 126–137, Minneapolis, MN, May 1994.
- [319] Luis Gravano, Hector García-Molina, and Anthony Tomasic. GLOSS: Text-source discovery over the Internet. *ACM Transactions on Database Systems*, 1999. To appear.
- [320] S. J. Green. Automated link generation: can we do better than term repetition. In *7th WWW Conference*, Brisbane, Australia, 1998.
- [321] Stephan Green, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. Technical Report Department of Computer Science CS-TR-3838, University of Maryland, 1997.
- [322] Sharon L. Greene, Susan J. Devlin, Philip E. Cannata, and Louis M. Gomez. No IFs, ANDs, or ORs: A study of database querying. *International Journal of Man-Machine Studies*, 32(3):303–326, 1990.
- [323] Gregory Grefenstette. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Boston, USA, 1998.
- [324] S. M. Griffin. NSF/DARPA/NASA Digital Libraries Initiative: A program manager's perspective. *D-Lib Magazine*, July/August 1998. <http://www.dlib.org/dlib/july98/07griffin.html>.
- [325] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts. Integrating structured data and text: A relational approach. *Journal of the American Society for Information Science*, 48:122–132, 1997.
- [326] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers, 1998.
- [327] V. Gudivada, V. Raghavan, W. Grosky, and R. Kasanagottu. Information retrieval on the World Wide Web. *IEEE Internet Computing*, Oct-Nov:58–68, 1997.

- [328] J.-C. Guéron. The virtual library: An oxymoron? NLM and MLA 1998 Leiter Lecture, National Library of Medicine, Bethesda, MD, May 1998.
- [329] Catherine Guinan and Alan F. Smeaton. Information retrieval from hypertext using dynamically planned guided tours. In *Proc. of the 4th ACM Conference on Hypertext*, pages 122–130, 1992.
- [330] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *Proc. ACM SIGMOD*, pages 47–57, Boston, Mass, June 1984.
- [331] Trudi Bellardo Hahn. Pioneers of the online age. *Information Processing & Management*, 32:33–48, 1996.
- [332] David Haines and W. Bruce Croft. Relevance feedback and inference networks. In *Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, PA, USA, 1993.
- [333] J. Hakala. The 5th Dublin Core Metadata Workshop, October 1997. <http://linnea.helsinki.fi/meta/DC5.html>.
- [334] Frank G. Halasz, Thomas P. Moran, and Randall H. Trigg. Notecards in a nutshell. In *Proc. of ACM CHI+GI Conference on Human Factors in Computing Systems and Graphics Interface*, pages 45–52, 1987.
- [335] Micheline Hancock-Beaulieu. User friendliness and human-computer interaction in on-line library catalogs. *Program*, 46(1):29–37, 1992.
- [336] Micheline Hancock-Beaulieu. Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1):8–19, 1997.
- [337] Micheline Hancock-Beaulieu, Margaret Fieldhouse, and Thien Do. An evaluation of interactive query expansion in an online library catalogue with a graphical user interface. *Journal of Documentation*, 51(3):225–243, 1995.
- [338] Micheline Hancock-Beaulieu and Stephen Walker. An evaluation of automatic query expansion in an online library catalogue. *Journal of Documentation*, 48(4):406–421, 1992.
- [339] D. Harman. Ranking algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 363–392. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [340] D. Harman. Relevance feedback and other query modification techniques. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 241–263. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [341] D. Harman, E. Fox, R. Baeza-Yates, and W. Lee. Inverted files. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Algorithms and Data Structures*, chapter 3, pages 28–43. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [342] D. K. Harman. Overview of the third text retrieval conference. In *Proc. of the 3rd Text REtrieval Conference (TREC-3)*, pages 1–19, Gaithersburg, USA, 1995. National Institute of Standards and Technology Special Publication.
- [343] Donna Harman. Relevance feedback revisited. In *Proc. of the 5th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, Copenhagen, Denmark, 1992.
- [344] D. J. Harper. *Relevance Feedback in Document Retrieval Systems: An Evaluation of Probabilistic Strategies*. PhD thesis, Jesus College, Cambridge, England, 1980.
- [345] D. J. Harper and C. J. van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216, 1978.
- [346] G. Harper. The Conference on Fair Use (CONFU), September 1997. <http://www.utsystem.edu/ogc/intellectualproperty/confu.htm>.

- [347] S. P. Harter. What is a digital library? Definitions, content, and issues. In *Proc. of KOLISS DL '96: International Conference on Digital Libraries and Information Services for the 21st Century*, Seoul, Korea, September 1996. <http://php.indiana.edu/~harter/korea-paper.htm>.
- [348] Stephen P. Harter. *Online Information Retrieval*. Academic Press, 1986.
- [349] S. Harum. Digital Library Initiative, January 1998. <http://dli.grainger.uiuc.edu/national.htm>.
- [350] Alexander G. Hauptmann and Bert F. Green. A comparison of command, menu-selection and natural-language computer programs. *Behaviour and Information Technology*, 2(2):163–178, 1983.
- [351] Leigh Watson Healy. *Library Systems: Current Developments and Future Directions*. Council on Library and Information Resources, Washington DC, 1998.
- [352] J. Heaps. *Information Retrieval — Computational and Theoretical Aspects*. Academic Press, 1978.
- [353] M. Hearst, G. Kopec, and D. Brotsky. Research in support of digital libraries at Xerox PARC, Part II: Paper and digital documents. *D-Lib Magazine*, June 1996. <http://www.dlib.org/dlib/june96/hearst/06hearst.html>.
- [354] M. A. Hearst. Research in Support of Digital Libraries at Xerox PARC, Part I: The changing social roles of documents. *D-Lib Magazine*, May 1996. <http://www.dlib.org/dlib/may96/05hearst.html>.
- [355] Marti A. Hearst. TileBars: Visualization of term distribution information in full text information access. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, Denver, CO, May 1995.
- [356] Marti A. Hearst. Improving full-text precision using simple query constraints. In *Proc. of the 5th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, 1996.
- [357] Marti A. Hearst. The use of categories and clusters in organizing retrieval results. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999. To appear.
- [358] Marti A. Hearst and Chandu Karadi. Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proc. of the 20th Annual International ACM SIGIR Conference*, pages 246–255, Philadelphia, PA, 1997.
- [359] Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proc. of the 19th Annual International ACM SIGIR Conference*, pages 76–84, Zurich, Switzerland, 1996.
- [360] L. Heath, D. Hix, L. Nowell, W. Wake, G. Averbach, and E. Fox. Envision: A user-centered database from the computer science literature. *Communications of the ACM*, 38(4):52–53, April 1995.
- [361] J. H. Hellerstein and M. Stonebraker. Predicate migration: Optimizing queries with expensive predicates. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 267–276, Washington, DC, USA, 1993.
- [362] Joseph M. Hellerstein, Jeffrey F. Naughton, and Avi Pfeffer. Generalized search trees for database systems. In *Proc. of VLDB Conf.*, pages 562–573, Zurich, Switzerland, Sept 1995.
- [363] Matthias Hemmje, Clemens Kunkel, and Alexander Willett. LyberWorld – a visualization user interface supporting fulltext retrieval. In *Proc. of the 17th Annual International ACM SIGIR Conference*, pages 249–259, Dublin, Ireland, July 1994.

- [364] R. Hendly, N. Drew, A. Wood, and R. Beale. Narcissus: Visualizing information. In *Proc. of the IEEE Information Visualization Symposium*, pages 90–96, Atlanta, GA, USA, Oct 1995.
- [365] David G. Hendry and David J. Harper. An informal information-seeking environment. *Journal of the American Society for Information Science*, 48(11):1036–1048, 1997.
- [366] M. Hersovici, M. Jacobi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. The shark-search algorithm. An application: tailored Web site mapping. In *7th WWW Conference*, Brisbane, Australia, April 1998.
- [367] C. Hert. Information retrieval: A social informatics perspective. Allerton discussion document submitted for the 1996 Allerton Institute, 1996.
- [368] Morten Hertzum and Erik Frokjaer. Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction*, 3(2):136–161, 1996.
- [369] Eric van Herwijnen. *Practical SGML*. Kluwer Academic Publishers, 2nd edition, 1994.
- [370] Ron R. Hightower, Laura T. Ring, Jonathan I. Helfman, Benjamin B. Bederson, and James D. Hollan. Graphical multiscale Web histories: A study of padprints. In *Proc. of the Ninth ACM Conference on Hypertext*, pages 58–65, Pittsburgh, PA, USA, 1998.
- [371] Charles R. Hildreth. The use and understanding of keyword searching in an university online catalog. *Information Technology and Libraries*, 16:52–62, 1997.
- [372] C. R. Hildreth. Online library catalogues as IR systems: what can we learn from research? In P. A. Yates Mercer, editor, *Future Trends in Information Science and Technology*, pages 9–25. Taylor-Graham, Los Angeles, 1988.
- [373] Rainer Himmeroder, Georg Lausen, Bertram Ludascher, and Christian Schlepphorst. On a declarative semantics for Web queries. In *Proc. of the Int. Conf. on Deductive and Object-Oriented Database (DOOD)*, pages 386–398, Singapore, December 1997.
- [374] D. S. Hirschberg and D. A. Lelewer. Efficient decoding of prefix codes. *Communications of the ACM*, 33(4), 1990.
- [375] J. Hopcroft and J. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley Publishing Company, 1979.
- [376] R. N. Horspool. Practical fast searching in strings. *Software Practice and Experience*, 10:501–506, 1980.
- [377] R. N. Horspool and G. V. Cormack. Constructing word-based text compression algorithms. In *Proc. of IEEE Second Data Compression Conference*, pages 62–81, 1992.
- [378] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proc. of AAAI 98*, Madison, WI, July 1998.
- [379] P. Hoschka. Synchronized multimedia integration language. W3C Working Draft, February 1998. <http://www.w3.org/TR/WD-smil>.
- [380] Hotbot: Main page. <http://www.hotbot.com>, 1996.
- [381] T. Y. Hou, A. Hsu, P. Liu, and M. Y. Chiu. A content-based indexing technique using relative geometry features. In *Proc. of SPIE'92*, volume 1662, pages 59–68, 1992.
- [382] N. Van House, D. Levy, A. Bishop, and B. Battenfield. User needs assessment and evaluation: issues and methods (workshop). In *Proc. of the 1st ACM International Conference on Digital Libraries*, page 186, College Station, TX, USA, 1996.
- [383] Adele Howe and Danielle Dreilinger. Savvysearch: A metasearch engine that learns which search engines to query. *AI Magazine*, 18(2):19–25, 1997.
- [384] J-Y. Huang, C-T. Fang-Tsou, and J-L. Chang. Multiuser 3D Web browsing system. *IEEE Internet Computing*, 2(5):70–73, 1998.

- [385] M. L. Huang, P. Eades, and R. F. Cohen. WebOFDAV — navigating and visualizing the Web on-line with animated context swapping. In *7th WWW Conference*, Brisbane, Australia, April 1998.
- [386] D. Huffman. A method for the construction of minimum-redundancy codes. *Proc. of the I.R.E.*, 40(9):1090–1101, 1952.
- [387] D. A. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proc. of the 19th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, Zurich, Switzerland, 1996.
- [388] David A. Hull, Jan O. Pedersen, and Hinrich Schütze. Method combination for document filtering. In *Proc. of the 19th Annual Int. ACM SIGIR Conference*, pages 279–287, Zurich, Switzerland, 1996.
- [389] R. Iannella. Australian digital library initiatives. *D-Lib Magazine*, December 1996. <http://www.dlib.org/dlib/december96/12iannella.html>.
- [390] IBM Corporation, IBM Digital Library, 1998. <http://www.software.ibm.com/is/dig-lib/>.
- [391] E. Ide. New experiments in relevance feedback. In G. Salton, editor, *The SMART Retrieval System*, pages 337–354. Prentice Hall, 1971.
- [392] IEEE Standards Committee on Optical Disk and Multimedia Platforms (SC ODMF), IEEE SFQL. Technical report, IEEE, Washington, USA, 1992.
- [393] G. H. Brett II. An integrated system for distributed information services. *D-Lib Magazine*, December 1996. <http://www.dlib.org/dlib/december96/dipps/12brett.html>.
- [394] Infoseek Corporation, Distributed Search Patent. http://software.infoseek.com/-patents/dist_search/Default.htm.
- [395] Inktomi: Main Page. <http://www.inktomi.com>, 1998.
- [396] International DOI Foundation, Digital Object Identifier System, June 1998. <http://www.doi.org/index.html>.
- [397] H. Ishikawa, F. Suzuki, F. Kozakura, A. Makinouchi, M. Miyagishima, M. Aoshima Y. Izumida, and Y. Yamane. The model, language and implementation of an object-oriented multimedia knowledge base management system. *ACM Transactions on Database Systems*, 18:1–50, 1993.
- [398] ISO. *Office Document Architecture (ODA) an Interchange Format*, 1986.
- [399] Paul S. Jacobs and Lisa F. Rau. Innovations in text interpretation. *Artificial Intelligence*, 63(1-2):143–191, 1993.
- [400] H. V. Jagadish. A retrieval technique for similar shapes. In *Proc. ACM SIGMOD Conf.*, pages 208–217, Denver, Colorado, May 1991.
- [401] Jango: Main Page. <http://www.jango.com>, 1998.
- [402] M. Jarke and J. Koch. Query optimization in database systems. *ACM Computing Surveys*, 16(2):111–152, June 1984.
- [403] B. J. Jensen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the Web. *ACM SIGIR Forum*, 32(1):5–17, 1998.
- [404] Byeong-Soo Jeong and Edward Omiecinski. Inverted file partitioning schemes in multiple disk systems. *IEEE Trans. Par. and Dist. Syst.*, 6(2):142–153, February 1995.
- [405] H. Jiang, A. K. Elmagarmid, A. A. Helal, A. Joshi, and M. Ahmed. *Video Database Systems*. Kluwer Academic, San Francisco, CA, 1997.
- [406] Thorsten Joachims, Dayne Freitag, and Tom Mitchell. WebWatcher: A tour guide for the World Wide Web. In *Proc. of the 15th Int. Joint Conference on Artificial Intelligence*, Nagoya, Japan, August 1997.

- [407] Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *Proc. of the 2nd Int. IEEE Visualization Conference*, pages 284–291, San Diego, USA, 1991.
- [408] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.
- [409] K. Sparck Jones. A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28(1):11–20, 1972.
- [410] K. Sparck Jones. A statistical interpretation of term specificity and its application to retrieval. *Information Storage and Retrieval*, 9(11):619–633, 1973.
- [411] K. Sparck Jones. Experiments in relevance weighting of search terms. *Information Processing & Management*, 15(13):133–144, 1979.
- [412] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.
- [413] K. Sparck Jones and E. O. Barber. What makes an automatic keyword classification effective. *Journal of the American Society for Information Sciences*, 22(3):166–175, 1971.
- [414] K. Sparck Jones and P. Willet. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, Inc., 1997.
- [415] Karen Sparck Jones. The Cranfield tests. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, pages 256–284. Butterworths, London, 1981.
- [416] S. Jones, S. Walker, M. Gatford, and T. Do. Peeling the onion: Okapi system architecture and software design issues. *Journal of Documentation*, 53:58–68, 1997.
- [417] Steve Jones. Graphical query specification and dynamic result previews for a digital library. In *Proc. of UIST'98, ACM Symposium on User Interface Software and Technology*, San Francisco, USA, November 1998.
- [418] T. Joseph and A. F. Cárdenas. PICQUERY: A high level language for pictorial database management. *IEEE Transactions on Software Engineering*, 14(5):630–638, 1988.
- [419] T. Joyce and R. M. Needham. The thesaurus approach to information retrieval. In K. Sparck Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 15–20. Morgan Kaufmann Publishers, Inc., 1997.
- [420] D. Austin Henderson Jr and Stuart K. Card. Rooms: The use of multiple virtual workspaces to reduce space contention in a window-based graphical user interface. *ACM Transactions on Graphics*, 5(3):211–243, 1986.
- [421] E. J. Whitehead Jr and M. Wiggins. WebDAV: IETF standard for collaborative authoring on the Web. *IEEE Internet Computing*, 2(5):34–40, 1998.
- [422] W. M. Shaw Jr, R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Cluster-based retrieval models. *Information Processing & Management*, 33(1):1–14, 1997.
- [423] W. M. Shaw Jr., R. Burgin, and P. Howell. Performance standards and evaluations in IR test collections: Vector-space and other retrieval models. *Information Processing & Management*, 33(1):15–36, 1997.
- [424] B. Kahle. Archiving the Internet. http://www.alexandria.com/~brewster/essays/sciam_article.html, 1997.
- [425] B. Kahle and A. Medlar. An information server for corporate users: Wide Area Information Servers. *ConneXions — The Interoperability Report*, 5(11):2–9, 1991. <ftp://think.com/wais/wais-corporate-paper.text>.
- [426] R. Kahn and R. Wilensky. A framework for distributed digital object services. Technical Report cnri.dlib/tn95-01, CNRI, May 1995. <http://www.cnri.reston.va.us/k-w.html>.
- [427] Ibrahim Kamel and Christos Faloutsos. Hilbert R-tree: An improved R-tree using fractals. In *Proc. of VLDB Conference*, pages 500–509, Santiago, Chile, Sept 1994.

- [428] Eser Kandogan and Ben Shneiderman. Elastic windows: Evaluation of multi-window operations. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 1, pages 250–257, Atlanta, GA, USA, March 1997.
- [429] P. B. Kantor. Assessing the factors leading to adoption of digital libraries, and growth in their impacts: the Goldilocks principle. Allerton discussion document submitted for the 1996 Allerton Institute. <http://edfu.lis.uiuc.edu/allerton/96/kantor.html>.
- [430] J. Karkkainen and E. Ukkonen. Two and higher dimensional pattern matching in optimal expected time. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 715–723, San Francisco, USA, 1994.
- [431] Norio Katayama and Shin'ichi Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proc. of ACM SIGMOD*, pages 369–380, Tucson, AZ, 1997.
- [432] Henry Kautz, Bart Selman, and Mehul Shah. The hidden Web. *AI Magazine*, 18(2):27–36, 1997.
- [433] P. R. Keller and M. M. Keller. *Visual Cues: Practical Data Visualization*. IEEE Computer Society Press, 1993.
- [434] R. Kengeri, C. D. Seals, H. P. Reddy, H. D. Harley, and E. A. Fox. Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD, December 1997. <http://fox.cs.vt.edu/~fox/u/Usability.pdf>.
- [435] M. M. Kessler. Comparison of results of bibliographic coupling and analytic subject indexing. *American Documentation*, 16(3):223–233, July 1965.
- [436] R. Khare and A. Rifkin. XML: A door to automated Web applications. *IEEE Internet Computing*, 1(4):78–86, 1977.
- [437] R. Khare and A. Rifkin. The origin of (document) species. In *7th WWW Conference*, Brisbane, Australia, April 1998.
- [438] S. Khoshafian and A. B. Baker. *Multimedia and Imaging Databases*. Morgan Kauffman, 1996.
- [439] P. Kilpelainen and H. Mannila. Retrieval from hierarchical texts by partial patterns. In *Proc. of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–222, Pittsburgh, USA, 1993.
- [440] Hanhwe Kim and Stephen C. Hirtle. Spatial metaphors and disorientation in hypertext browsing. *Behaviour and Information Technology*, 14(4):239–250, 1995.
- [441] Steven T. Kirsch. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents. US Patent 5,659,732, August 1997.
- [442] W. Klas and A. Sheth. Special issue on meta-data for digital media. *SIGMOD Record*, 23(4):19–20, 1994.
- [443] Adrience J. Kleiboemer, Manette B. Lazear, and Jan O. Pedersen. Tailoring a retrieval system for naive users. In *Proc. of the 5th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, USA, 1996.
- [444] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, USA, Jan 1998.
- [445] D. E. Knuth. *The Art of Computer Programming*, volume 3: Searching and Sorting. Addison-Wesley, 1973.
- [446] D. E. Knuth. Dynamic Huffman coding. *Journal of Algorithms*, 6:163–180, 1985.
- [447] D. E. Knuth, J. H. Morris, Jr, and V. R. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6(1):323–350, 1977.

- [448] Jurgen Koenemann and Nicholas J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 1 of *Papers*, pages 205–212, Zurich, Switzerland, 1996.
- [449] Janet L. Kolodner. *Case-based Reasoning*. Morgan Kaufmann Publishers, 1993.
- [450] D. Konopnicki and O. Shmueli. W3QS: A query system for the World Wide Web. In *Proc. of VLDB'95*, pages 54–65, Zurich, Switzerland, September 1995.
- [451] Robert Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, Inc., 1997.
- [452] Robert R. Korfhage. To see or not to see – is that the query? In *Proc. of the 14th Annual Int. ACM SIGIR Conference*, pages 134–141, Chicago, USA, 1991.
- [453] Flip Korn and Ben Shneiderman. Navigating terminology hierarchies to access a digital library of medical images. Technical Report HCIL-TR-94-03, University of Maryland, USA, Korn & Shneiderman, 1995.
- [454] Flip Korn, Nikolaos Sidiropoulos, Christos Faloutsos, Eliot Siegel, and Zenon Protopapas. Fast nearest-neighbor search in medical image databases. *Proc. Conf. on Very Large Data Bases*, pages 215–226, September 1996.
- [455] J. Korpela. Lurching toward Babel: HTML, CSS, and XML. *IEEE Computer*, 31(7):103–106, 1998.
- [456] S. M. Kosslyn. Understanding charts and graphs. *Applied Cognitive Psychology*, 3:185–226, 1989.
- [457] M. Koster. Guidelines for robot writers, 1993. <http://info.webcrawler.com/mak/-projects/robots/guidelines.html>.
- [458] Nick Koudas, Christos Faloutsos, and Ibrahim Kamel. Declustering spatial databases on a multi-computer architecture. *EDBT Conf. Proc.*, pages 592–614, March 1996.
- [459] Gerald Kowalski. *Information Retrieval Systems, Theory and Implementation*. Kluwer Academic Publishers, Boston, USA, 1997.
- [460] Robyn Kozierok and Pattie Maes. A learning interface agent for scheduling meetings. In *Proc. of the 1993 Int. Workshop on Intelligent User Interfaces*, pages 81–88, New York, NY, 1993.
- [461] D. Kraft and D. A. Buel. Fuzzy sets and generalized Boolean retrieval systems. *International Journal of Man-Machine Studies*, 19:45–56, 1983.
- [462] Joseph B. Kruskal and Myron Wish. *Multidimensional scaling*. SAGE publications, Beverly Hills, 1978.
- [463] Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proc. of the 16th Annual Int. ACM SIGIR Conference*, pages 181–190, Pittsburgh, PA, 1993.
- [464] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proc. of the 18th Annual Int. ACM SIGIR Conference*, pages 68–73, Seattle, WA, 1995.
- [465] K. L. Kwok, L. Grunfeld, and D. D. Lewis. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In *Proc. of the Text REtrieval Conference*, pages 247–256, Gaithersburg, MD, USA, 1995.
- [466] K. L. Kwok. A neural network for probabilistic information retrieval. In *Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–30, Cambridge, USA, June 1989.
- [467] K. L. Kwok. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, 8(4):363–386, October 1990.

- [468] K. L. Kwok. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(3):324–353, July 1995.
- [469] K. L. Kwok, L. Papadopolous, and Y. Y. Kwan. Retrieval experiments with a large collection using PIRCS. In *Proc. of the 1st Text Retrieval Conference*, pages 153–172, Gaithersburg, MD, USA, 1993. Special Publication 500-267, National Institute of Standards and Technology (NIST).
- [470] Eric Lagergren and Paul Over. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In *Proc. of the 21st Annual Int. ACM SIGIR Conference*, pages 164–172, Melbourne, Australia, 1998.
- [471] C. Lagoze. Networked Computer Science Technical Reference Library. <http://www.ncstrl.org>.
- [472] C. Lagoze. The Warwick framework: a container architecture for diverse sets of metadata. *D-Lib Magazine*, July/August 1996. <http://www.dlib.org/dlib/july96/lagoze/07lagoze.html>.
- [473] C. Lagoze and D. Ely. Implementation issues in an open architecture framework for digital object services. Technical Report TR95-1540, Cornell University Computer Science Department, 1995.
- [474] C. Lagoze, D. Fielding, and S. Payette. Making global digital libraries work: collection services, connectivity regions, and collection views. In *Proc. of the 3rd ACM International Conference on Digital Libraries*, 1998.
- [475] C. Lagoze, R. McGrath, E. Overly, and N. Yeager. A design for inter-operable secure object stores (ISOS). Technical Report TR95-1558, Cornell University Computer Science Department, 1995.
- [476] Laks V. S. Lakshmanan, Fereidoon Sadri, and Iyer N. Subramanian. A declarative language for querying and restructuring the Web. In *6th Int. Workshop on Research Issues in Data Engineering, RIDE '96*, New Orleans, February 1996.
- [477] B. LaMacchia. The Internet fish construction kit. In *6th. Int'l. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [478] John Lamping, Ramana Rao, and Peter Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proc. of the ACM Conference on Human Factors in Computing Systems*, pages 401–408, Denver, CO, USA, May 1995.
- [479] F. W. Lancaster. Natural language in information retrieval. In *Indexing and Abstracting in Theory and Practice*, pages 193–218. University of Illinois, Champaign, IL, 1991.
- [480] G. Landau and U. Vishkin. Fast string matching with k differences. *Journal of Computer Systems Science*, 37:63–78, 1988.
- [481] Thomas K. Landauer, Dennis E. Egan, Joel R. Remde, Michael Lesk, Carol C. Lochbaum, and Daniel Ketchum. Enhancing the usability of text through computer delivery and formative evaluation: the SuperBook project. In C. McKnight, A. Dillon, and J. Richardson, editors, *Hypertext: A Psychological Perspective*, pages 71–136. Ellis Horwood, 1993.
- [482] Andrew Large and Jamshid Beheshti. OPACS: a research review. *LISR*, 19:111–133, 1997.
- [483] Jill H. Larkin and Herbert A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11:65–99, 1987.
- [484] Ray R. Larson. Between Scylla and Charybdis: Subject searching in the online catalog. *Advances in Librarianship*, 15:175–236, 1991.

- [485] Ray R. Larson. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. In *Proc. of the 1996 Annual ASIS Meeting*, pages 71–78, 1996.
- [486] Ray R. Larson, Jerome McDonough, Paul O’Leary, and Lucy Kuntz. Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science*, 47(7):555–567, 1996.
- [487] Ora Lassila. Web metadata: A matter of semantics. *IEEE Internet Computing*, 2(4):30–37, 1998.
- [488] S. Lawrence and C. L. Giles. Inquirus, the NECI meta search engine. In *7th WWW Conference*, pages 95–105, Brisbane, Australia, 1998.
- [489] S. Lawrence and C. Lee Giles. Context and page analysis for improved Web search. *IEEE Internet Computing*, 2(4):38–46, 1998.
- [490] S. Lawrence and C. L. Giles. Searching the World Wide Web (in reports). *Science*, 280(5360):98, April 3 1998.
- [491] Doug Lea. *Concurrent Programming in Java: Design Principles and Patterns*. The Java Series. Addison-Wesley, Reading, MA, 1997.
- [492] David B. Leake, editor. *Case-based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press, Menlo Park, CA, 1996.
- [493] J. H. Lee, W. Y. Kim, and Y. H. Lee. Ranking documents in thesaurus-based Boolean retrieval systems. *Information Processing & Management*, 30(1):79–91, 1993.
- [494] J. J. Lee and P. Kantor. A study of probabilistic information retrieval systems in the case of inconsistent expert judgements. *Journal of the American Society for Information Sciences*, 42(3):166–172, 1991.
- [495] Joon Ho Lee. Properties of extended Boolean models in information retrieval. In *Proc. of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 182–190, Dublin, Ireland, 1994.
- [496] Joon Ho Lee, Won Yong Kim, Myoung Ho Kim, and Yoon Joon Lee. On the evaluation of Boolean operators in the extended Boolean retrieval framework. In *Proc. of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–297, Pittsburgh, PA, USA, 1993.
- [497] O. Lehtinen, E. Sutinen, and J. Tarhio. Experiments on block indexing. In *Proc. of WSP’96*, pages 183–193, Recife, Brazil, 1996. Carleton University Press.
- [498] H. V. Leighton and J. Srivastava. Precision among World Wide Web search engines: AltaVista, Excite, Hotbot, Infoseek, and Lycos. <http://www.winona.msus.edu/library/webind2/webind2.htm>, 1997.
- [499] B. Leiner. D-Lib Working Group on Digital Library Metrics, May 1998. <http://www.dlib.org/metrics/public/metrics-home.html>.
- [500] M. E. Lesk. Word-word associations in document retrieval systems. *American Documentation*, 20(1):8–36, 1969.
- [501] Michael Lesk. *Practical Digital Libraries; Books, Bytes, and Bucks*. Morgan Kaufmann, 1997.
- [502] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction*, 1(2):126–160, 1994.
- [503] R. LeVan. Dublin Core and Z39.50. Draft version 1.2, February 1998. <http://cypress.dev.oclc.org:12345/~rrl/docs/dublincoreandz3950.html>.
- [504] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. Original in Russian in *Doklady Akademii Nauk SSSR*, 163(4):8 45–848, 1965.

- [505] D. M. Levy. Topics in document research. In *Proc. of the ACM Conference on Document Processing Systems (SIGDOC '88)*, pages 187–193, 1988.
- [506] D. M. Levy. Fixed or fluid?: document stability and new media. In *Proc. of the 1994 ACM European Conference on Hypermedia Technology*, pages 24–31, 1994.
- [507] D. M. Levy. I read the news today, oh boy: reading and attention in digital libraries. In *Proc. of the 2nd ACM Int. Conference on Digital Libraries*, pages 202–211, Philadelphia, PA, USA, July 1997.
- [508] D. M. Levy, D. C. Brotsky, and K. R. Olson. Formalizing the figural: aspects of a foundation for document manipulation. In *Proc. of the ACM Conference on Document Processing Systems (SIGDOC '88)*, pages 145–151, 1988.
- [509] D. M. Levy and C. C. Marshall. Going digital: a look at assumptions underlying digital libraries. *Communications of the ACM*, 38:77–84, April 1995.
- [510] The LEXIS-NEXIS Computing Complex. <http://www.lexis-nexis.com/lnc/about/datacenter.html>.
- [511] W-S. Li, J. Shim, K. S. Candan, and Y. Hara. WebDB: A Web query system and its modeling, language, and implementation. In *Proc. of Advances in Digital Libraries*, Santa Barbara, CA, USA, April 1998.
- [512] Y. Li. Toward a qualitative search engine. *IEEE Internet Computing*, 2(4):24–29, 1998.
- [513] Library of Congress, Metadata, Dublin Core and USMARC: A Review of Current Efforts, January 1997. <gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc>.
- [514] Library of Congress, MARC Standards, June 1998. <http://lcweb.loc.gov/marc/marc.html>.
- [515] Library of Congress, Z39.50 Maintenance Agency, June 1998. <http://lcweb.loc.gov/z3950/agency/>.
- [516] J. C. R. Licklider. *Libraries of the Future*. Cambridge, MA.: M.I.T. Press, 1965.
- [517] H. W. Lie and B. Bos. *Cascading Style Sheets: Designing for the Web*. Addison-Wesley, 1997.
- [518] Henry Lieberman. Letizia: an agent that assists Web browsing. In *Proc. of the 14th Int. Joint Conference on Artificial Intelligence*, pages 924–929, 1995.
- [519] King-Ip Lin, H. V. Jagadish, and Christos Faloutsos. The TV-tree — an index structure for high-dimensional data. *VLDB Journal*, 3:517–542, October 1994.
- [520] Xia Lin, Dagobert Soergel, and Gary Marchionini. A self-organizing semantic map for information retrieval. In *Proc. of the 14th Annual Int. ACM SIGIR Conference*, pages 262–269, Chicago, 1991.
- [521] Ming-Ling Lo and Chinya V. Ravishankar. Spatial joins using seeded trees. In *Proc. of ACM SIGMOD*, pages 209–220, Minneapolis, MN, USA, May 1994.
- [522] R. M. Losee and A. Bookstein. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing & Management*, 24(3):315–321, 1988.
- [523] Henry J. Lowe and G. Octo Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association*, 271(4):1103–1108, 1994.
- [524] Zhihong Lu, Kathryn S. McKinley, and Brendon Cahoon. The hardware/software balancing act for information retrieval on symmetric multiprocessors. Technical Report TR98-25, Dept. of Comp. Sci., Univ. of Mass., Amherst, MA, 1998.
- [525] C. Lynch and H. García-Molina. Interoperability, scaling, and the digital libraries research agenda: A report on the May 18-19, 1995, IITA Digital Libraries Workshop, August 1995. <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>.

- [526] Clifford Lynch. The next generation of public access information retrieval systems for research libraries: lessons from ten years of the MELVYL system. *Information Technology and Libraries*, 11(4):405–415, 1992.
- [527] Y. Maarek, M. Jacovi, M. Shtalhaim, S. Ur, D. Zernik, and I. Z. Ben Shaul. WebCutter: A system for dynamic and tailorable site mapping. In *6th WWW Conference*, pages 713–722, Santa Clara, CA, USA, 1997.
- [528] Y. S. Maarek. Organizing documents to support browsing in digital libraries. Allerton discussion document submitted for the 1995 Allerton Institute. <http://edfu.lis.uiuc.edu/allerton/95/s4/maarek.html>.
- [529] Y. S. Maarek and A. J. Wecker. The librarian’s assistant: Automatically assembling books into dynamic bookshelves. In *Proc. of RIAO '94: Intelligent Multimedia Information Retrieval Systems and Management*, New York, USA, October 1994.
- [530] Jock Mackinlay, George Robertson, and Stuart K. Card. The perspective wall: Detail and context smoothly integrated. In *Proc. of the ACM Conference on Human Factors in Computing Systems*, pages 173–179, 1991.
- [531] Jock D. Mackinlay, Ramana Rao, and Stuart K. Card. An organic user interface for searching citation links. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 1 of *Papers*, pages 67–73, Denver, CO, USA, 1995.
- [532] A. MacLean, P. J. Barnard, and M. D. Wilson. Evaluating the human interface of a data entry system: User choice and performance measures yield different tradeoff functions. In *Proc. of the HCI'85 Conference on People and Computers: Designing the Interface*, pages 172–185, 1985.
- [533] I. A. MacLeod. Storage and retrieval of structured documents. *Information Processing & Management*, 26(2):197–208, 1990.
- [534] I. A. MacLeod. A query language for retrieving information from hierarchic text structures. *The Computer Journal*, 34(3):254–264, 1991.
- [535] Ian A. Macleod, T. Patrick Martin, Brent Nordin, and John R. Phillips. Strategies for building distributed information retrieval systems. *Information Processing & Management*, 23(6):511–528, 1987.
- [536] Pattie Maes and Robyn Kozierok. Learning interface agents. In *Proc. of AAAI 93*, pages 459–465, Washington, DC, July 1993.
- [537] U. Manber and P. Bigot. Search Broker: Main page. <http://debussy.cs.arizona.edu/sb/>, 1997.
- [538] U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, pages 319–327, San Francisco, USA, 1990.
- [539] U. Manber, M. Smith, and B. Gopal. WebGlimpse: combining browsing and searching. In *Proc. of USENIX Technical Conference*, pages 195–206, Anaheim, USA, Jan 1997.
- [540] Udi Manber and Sun Wu. GLIMPSE: A tool to search through entire file systems. In *Proc. of USENIX Technical Conference*, pages 23–32, San Francisco, USA, January 1994. <ftp://cs.arizona.edu/glimpse/glimpse.ps.Z>.
- [541] B. Mandelbrot. *Fractal Geometry of Nature*. W. H. Freeman, New York, 1977.
- [542] Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, 1995.
- [543] M. Marchiori. The quest for correct information on the Web: Hyper search engines. In *6th WWW Conf.*, pages 265–274, Santa Clara, CA, USA, 1997.
- [544] M. Marchiori. The limits of Web metadata and beyond. In *7th WWW Conference*, Brisbane, Australia, April 1998.

- [545] S. Marcus and V. S. Subrahmanian. Foundations of Multimedia Database Systems. *Journal of the ACM*, 4(3):474–505, 1996.
- [546] Karen Markey, Pauline Atherton, and Claudia Newton. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review*, 4:225–236, 1982.
- [547] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Association for Computing Machinery*, 7(3):216–244, 1960.
- [548] C. C. Marshall. Annotation: from paper books to the digital library. In *Proc. of the 2nd ACM Int. Conference on Digital Libraries*, pages 131–141, Philadelphia, PA, USA, 1997.
- [549] Y. Masunaga. Design issues of OMEGA an object-oriented multimedia database management system. *Journal of Information Processing*, 14:60–74, 1991.
- [550] Mark T. Maybury. *Intelligent Multimedia Information Retrieval*. MIT Press, 1997.
- [551] Ray McAleese, editor. *Hypertext I: Theory into Practice*. Ablex Publishing Corporation, 1988.
- [552] B. McCune, R. Tong, J. S. Dean, and D. Shapiro. Rubric: A system for rule-based information retrieval. *IEEE Transactions on Software Engineering*, 11(9), 1985.
- [553] Charles T. Meadow, Barbara A. Cerny, Christine L. Borgman, and Donald O. Case. Online access to knowledge: System design. *Journal of the American Society for Information Science*, 40(2):86–98, 1989.
- [554] Daniel A. Menasce and Virgilio A. F. Almeida. *Capacity Planning for Web Performance: Metrics, Models, and Methods*. Prentice Hall, 1998.
- [555] Filippo Menczer and Richard K. Belew. Adaptive information agents in distributed textual environments. In Katia P. Sycara and Michael Wooldridge, editors, *Proc. of the 2nd Int. Conference on Autonomous Agents*, pages 157–164, May 1998.
- [556] A. Mendelzon, G. Mihaila, and T. Milo. Querying the World Wide Web. *International Journal on Digital Libraries*, 1(1):54–67, April 1997.
- [557] Beth Meyer, Richard A. Sit, Victoria A. Spaulding, Sherry E. Mead, and Neff Walker. Age group differences in World Wide Web navigation. In Katia P. Sycara and Michael Wooldridge, editors, *Proc. of ACM Conference on Human Factors in Computing Systems*, pages 157–164, Atlanta, GA, USA, March 1997.
- [558] A. Michard. Graphical presentation of Boolean expressions in a database query language: design notes and an ergonomic evaluation. *Behaviour and Information Technology*, 1(3):279–288, 1982.
- [559] F. Miksa and P. Doty. Intellectual Realities and the Digital Library. In *Proc. of the 1st Annual Conference on the Theory and Practice of Digital Libraries*, College Station, TX, USA, June 1994.
- [560] E. J. Miller. CNI/OCLC Metadata Workshop: Workshop on Metadata for Networked Images, September 1996. <http://purl.oclc.org/metadata/image>.
- [561] G. A. Miller, E. B. Newman, and E. A. Friedman. Length-frequency statistics for written english. *Information and Control*, 1:370–389, 1958.
- [562] J. Minker, G. A. Wilson, and B. H. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8(6):329–348, 1972.
- [563] T. Minohara and R. Watanabe. Queries on structure in hypertext. In D. Lomet, editor, *Proc. of Foundations of Data Organization and Algorithms*, number 730 in LNCS, pages 394–411, Chicago, IL, USA, 1993. Springer-Verlag.

- [564] Richard Mitchell, David Day, and Lynette Hirschman. Fishing for information on the Internet. In *Proc. of the '95 Information Visualization*, pages 105–111, Atlanta, USA, October 1995.
- [565] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [566] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. of 21st Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, Melbourne, Australia, 1998.
- [567] S. Miyamoto, T. Miyake, and K. Nakayama. Generation of a pseudothsaurus for information retrieval based on co-occurrences and fuzzy set operations. *IEEE Transactions on Systems and Man Cybernetics*, 13(1):62–70, 1983.
- [568] S. Miyamoto and K. Nakayama. Fuzzy information retrieval based on a fuzzy pseudothsaurus. *IEEE Transactions on Systems and Man Cybernetics*, 16(2):278–282, 1986.
- [569] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.
- [570] W. E. Moen. Accessing distributed cultural heritage information. *Communications of the ACM*, 41(4):45–48, April 1998.
- [571] A. Moffat. Word-based text compression. *Software Practice and Experience*, 19(2):185–198, 1989.
- [572] A. Moffat and T. Bell. In situ generation of compressed inverted files. *Journal of the American Society for Information Science*, 46(7):537–550, 1995.
- [573] A. Moffat and J. Zobel. Parameterized compression for sparse bitmaps. In *Proc. of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 274–285, Copenhagen, Denmark, 1992.
- [574] Alistair Moffat and Justin Zobel. Information retrieval systems for large document collections. In Donna K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 85–94, Gaithersburg, MD, USA, 1995. Dept of Commerce, National Institute of Standards and Technology. Special Publication 500-226.
- [575] E. S. Moura, G. Navarro, and N. Ziviani. Indexing compressed text. In R. Baeza-Yates, editor, *Proc. of the 4th South American Workshop on String Processing*, volume 8, pages 95–111, Valparaíso, Chile, 1997. Carleton University Press International Informatics Series.
- [576] E. S. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Direct pattern matching on compressed text. In *Proc. of the 5th Symposium on String Processing and Information Retrieval*, pages 90–95, Santa Cruz, Bolivia, September 1998.
- [577] E. S. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Fast searching on compressed text allowing errors. In *Proc. of the ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 298–306, Melbourne, Australia, August 1998.
- [578] S. Mukherjea and J. D. Foley. Visualizing the World Wide Web with the Navigational View Builder. *Computer Networks and ISDN Systems*, 27:1075–1087, 1995.
- [579] S. Mukherjea, K. Hirata, and Y. Hara. Towards a multimedia World-Wide Web information retrieval engine. In *6th Int. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [580] Kevin Mullet and Darell Sano. *Designing Visual Interfaces: Communication-Oriented Techniques*. SunSoft Press, 1995.
- [581] Brad Myers. *Creating User Interfaces by Demonstration*. Academic Press, New York, 1988.

- [582] Brad Myers. A taxonomy of window manager user interfaces. *IEEE Computer Graphics and Applications*, pages 65–84, Sept 1988.
- [583] G. Myers. A fast bit-vector algorithm for approximate pattern matching based on dynamic programming. In M. Farach-Colton, editor, *Proc. of Combinatorial Pattern Matching*, number 1448 in LNCS, pages 1–13, Rutgers, USA, 1998. Springer-Verlag.
- [584] Bonnie A. Nardi. *A Small Matter of Programming: Perspectives on End User Computing*. MIT Press, 1993.
- [585] A. Nation. Visualizing Websites using a hierarchical table of contents browser: WebTOC. In *Proc. of the 3rd Conference on Human Factors and the Web*, Denver, CO, 1997.
- [586] National Humanities Alliance, Basic Principles for Managing Intellectual Property in the Digital Environment, March 1997. http://www-ninch.cni.org/ISSUES/COPYRIGHT/-PRINCIPLES/NHA_Complete.html.
- [587] *Nation's Information Infrastructure Steering Committee, Computer Science, More than screen deep: toward every-citizen interfaces to the nation's information infrastructure*. National Academy Press, Washington, DC, USA, 1997.
- [588] G. Navarro. *Approximate Text Searching*. PhD thesis, Dept. of Computer Science, Univ. of Chile, December 1998.
- [589] G. Navarro and R. Baeza-Yates. A language for queries on structure and contents of textual databases. In *Proc. of the 18th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 93–101, Seattle, USA, July 1995.
- [590] G. Navarro and R. Baeza-Yates. Proximal nodes: A model to query document databases by content and structure. *ACM Transactions on Office and Information Systems*, 15(4):401–435, 1997.
- [591] G. Navarro, J. Kitajima, B. Ribeiro-Neto, and N. Ziviani. Distributed generation of suffix arrays. In A. Apostolico and J. Hein, editors, *Proc. of Combinatorial Pattern Matching*, number 1264 in LNCS, pages 102–115, Aarhus, Denmark, 1997. Springer-Verlag.
- [592] G. Navarro and M. Raffinot. A bit-parallel approach to suffix automata: fast extended string matching. In M. Farach-Colton, editor, *Proc. of Combinatorial Pattern Matching*, number 1448 in LNCS, pages 14–33, Rutgers, USA, 1998. Springer-Verlag.
- [593] G. Navarro and M. Raffinot. A general practical approach to pattern matching over Ziv-Lempel compressed text. Technical Report TR/DCC-98-12, Dept. of Computer Science, Univ. of Chile, 1998. <ftp://ftp.dcc.uchile.cl/pub/users/gnavarro/lzsrch.ps.gz>.
- [594] NDLTD. Networked Digital Library of Theses and Dissertations, June 1998. <http://www.ndltd.org/>.
- [595] J. Nesbit. The accuracy of approximate string matching algorithms. *Journal of Computer-Based Instruction*, 13(3):80–83, 1986.
- [596] Netcraft Web Server Survey. <http://www.netcraft.com/Survey/>, 1998.
- [597] NetSizer: Main Page. <http://www.netsizer.com>, 1998.
- [598] Network Development and MARC Standards Office, Dublin Core/MARC/GILS Crosswalk, July 1997. <http://www.loc.gov/marc/dccross.html>.
- [599] Network Wizards: Internet Domain Survey. <http://www.nw.com>, 1998.
- [600] M. Lynne Neufeld and Martha Cornog. Database history: from dinosaurs to compact discs. *Journal of the American Society for Information Science*, 37(4):183–190, 1986.
- [601] New Zealand DL Group, The New Zealand Digital Library Project. <http://www.nzdl.org/>.

- [602] D. Ngu and X. Wu. SiteHelper: a localized agent that helps incremental exploration of the World Wide Web. In *6th. Int. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [603] Wayne Niblack, Ron Barber, Will Equitz, Myron Flickner, Eduardo Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. The QBIC project: Querying images by content using color, texture and shape. In *Int. Symposium on Electronic Imaging: Science and Technology, Conf. 1908, Storage and Retrieval for Image and Video Databases*, February 1993.
- [604] Jakob Nielsen. *Hypertext and Hypermedia*. Academic Press, 1990.
- [605] Jakob Nielsen. *Usability Engineering*. Academic Press, 1993.
- [606] NISO Press, Bethesda, MD. *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*, 1995. Accessible at <http://lcweb.loc.gov/z3950/agency/>.
- [607] D. A. Norman. *The Psychology of Everyday Things*. Basic Books, New York, 1988.
- [608] Northern Light: Main page. <http://www.northernlight.com>, 1997.
- [609] G. Notess. Search Engines Showdown. <http://imt.net/~notess/search/>, 1998.
- [610] L. Nowell. *Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data*. PhD Thesis, Virginia Polytechnic and State University, Department of Computer Science, 1997.
- [611] P. J. Nuernberg, R. Furuta, J. J. Leggett, C. C. Marshall, and F. M. Shipman III. Digital Libraries: Issues and Architectures. In *Proc. of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*, Austin, Texas, 1995.
- [612] D. W. Oard. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. *D-Lib Magazine*, December 1997. <http://www.dlib.org/dlib/december97/oard/12oard.html>.
- [613] D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, 1996.
- [614] Vicki L. O'Day and Robin Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proc. of the INTERCHI '93*, Amsterdam, Netherlands, April 1993. IOS Press.
- [615] R. N. Oddy. Information retrieval through man-machine dialogue. *Journal of Documentation*, 33:1–14, 1977.
- [616] Y. Ogawa, T. Morita, and K. Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems*, 39:163–179, 1991.
- [617] Kenton O'Hara and Abigail Sellen. A comparison of reading paper and on-line documents. In *Proc. of ACM Conference on Human Factors in Computing Systems*, Atlanta, GA, USA, March 1997. <http://www.acm.org/sigchi/chi97/proceedings/paper/koh.htm>.
- [618] K. Olsen, R. Korfhage, K. Sochats, M. Spring, and J. Williams. Visualization of a document collection with implicit and explicit links: The Vibe system. *Scandinavian Journal of Information Systems*, 5, 1993.
- [619] Ed O'Neill, Brian Lavoie, and Pat McClain. OCLC Web characterization project (position paper). In *Web Characterization Workshop*, Boston, USA, Nov 1998. <http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm>.
- [620] Online Computer Library Center, Inc. (OCLC). Metadata Workshop II, April 1996. <http://www.oclc.org:5046/oclc/research/conferences/metadata2/>.
- [621] E. Oomoto and K. Tanaka. OVID: Design and implementation of a video-object database system. *IEEE Trans. on Knowledge and Data Engineering.*, 5:629–641, 1993.

- [622] Alan Victor Oppenheim and Ronald W. Schafer. *Digital Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1975.
- [623] J. A. Orenstein. A Comparison of Spatial Query Processing Techniques for Native and Parameter Spaces. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 343–352, Atlantic City, New Jersey, 1988.
- [624] A. Paepcke, C.-C. K. Chang, H. García-Molina, and T. Winograd. Interoperability for Digital Libraries Worldwide. *Communications of the ACM*, 41(4):33–43, April 1998.
- [625] A. Paepcke, S. B. Cousins, H. García-Molina, S. W. Hassan, and S. P. Ketchpel. Using distributed objects for digital library interoperability. *IEEE Computer*, 29(5):61–68, May 1996.
- [626] C. D. Paice and P. A. Jones. The identification of important concepts in highly structured technical papers. In *Proc. of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78, Pittsburgh, USA, 1993.
- [627] George Panagopoulos and Christos Faloutsos. Bit-sliced signature files for very large text databases on a parallel machine architecture. In *Proc. 4th Int. Conf. on Extending Database Technology (EDBT)*, number 779 in LNCS, pages 379–392, London, 1994. Springer-Verlag.
- [628] Apostolos Papadopoulos and Yannis Manolopoulos. Performance of nearest neighbor queries in R-trees. In Foto N. Afrati and Phokion Kolaitis, editors, *Proc. of 6th Int. Conf. on Database Theory*, number 1186 in LNCS, pages 394–408, Delphi, Greece, Jan 1997.
- [629] Michael Pazzani, Daniel Billsus, and Jack Muramatsu. Syskill & Webert: Identifying interesting Web sites. In *Proc. of the Thirteenth Annual National Conference on Artificial Intelligence*, pages 54–61, Portland, OR, USA, August 1996.
- [630] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.
- [631] Gert Schmeltz Pedersen. A browser for bibliographic information retrieval, based on an application of lattice theory. In *Proc. of the 16th Annual Int. ACM SIGIR Conference*, pages 270–279, Pittsburgh, PA, 1993.
- [632] Theresa Pepin, Jeff Barry, and W. David Penniman. Automated system marketplace: the competitive edge: expanded access drives vendors. *Library Journal*, 122(6):47–56, 1997.
- [633] C. Peters and E. Picchi. across languages, across cultures: Issues in multilinguality and digital libraries. *D-Lib Magazine*, May 1997. <http://www.dlib.org/dlib/may97/peters/05peters.html>.
- [634] P. E. Peters. Digital libraries are much more than digitized collections. *EDUCOM Review*, 30(4), 1995.
- [635] Euripides G. M. Petrakis and Christos Faloutsos. Similarity searching in medical image databases. *IEEE Trans. on Knowledge and Data Engineering*, 9(3):435–447, May 1997.
- [636] Lori Phelps. Active documentation: Wizards as a medium for meeting user needs. In *Proc. of the ACM 15th International Conference on Systems Documentation*, pages 207–210, 1997.
- [637] D. Pimienta. Languages, culture, and Internet (in French). <http://funredes.org/>, March 1998.
- [638] Peter Pirolli and Stuart Card. Information foraging models of browsers for very large document spaces. In *Advanced Visual Interfaces*, L'Aquila, Italy, May 1998.

- [639] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the Web. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 118–125, Zurich, Switzerland, May 1996. ACM Press.
- [640] Peter Pirolli, Patricia Schank, Marti A. Hearst, and Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 213–220, Zurich, Switzerland, May 1996.
- [641] J. E. Pitkow. Summary of WWW characterizations. In *7th WWW Conference*, Brisbane, Australia, 1998.
- [642] J. E. Pitkow and K. A. Bharat. WebViz: A tools for World Wide Web access log analysis. In *Proc. of the 1st Int. WWW Conference*, Geneva, Switzerland, May 1994. <http://www1.cern.ch/PapersWWW94/pitkow-webvis.ps>.
- [643] Catherine Plaisant, Tom Bruns, Ben Shneiderman, and Khoa Doan. Query previews in networked information systems: the case of EOSDIS. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 2 of *Formal Video Program*, pages 202–203, Atlanta, GA, USA, March 1997.
- [644] Catherine Plaisant, David Carr, and Ben Shneiderman. Image-browser taxonomy and guidelines for designers. *IEEE Software*, 12(2):21–32, 1995.
- [645] Christine A. Pogue and Peter Willet. Use of text signatures for document retrieval in a highly parallel environment. *Parallel Computing*, 4:259–268, 1987.
- [646] Steven Pollitt. Interactive information retrieval based on faceted classification using views. In *Proc. of the 6th Int. Study Conference on Classification (FID/CR)*, University College, London, June 1997.
- [647] A. Pollock and A. Hockley. What's wrong with Internet searching. *D-Lib Magazine*, March 1997.
- [648] M. F. Porter. An algorithm for suffix stripping. In K. Sparck Jones and P. Willet, editors, *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers, Inc., 1997.
- [649] J. Powell and E. Fox. Multilingual federated searching across heterogeneous collections. *D-Lib Magazine*, September 1998.
- [650] Wanda Pratt. Dynamic organization of search results using the UMLS. In *American Medical Informatics Association Fall Symposium*, Nashville, TN, USA, October 1997.
- [651] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.
- [652] G. Pringle, L. Allison, and D. L. Dowe. What is a tall poppy among Web pages? In *7th WWW Conference*, Brisbane, Australia, April 1998.
- [653] Teresa Pritchard-Schoch. Comparing natural language retrieval: WIN & Freestyle. *Online*, 19(4):83–87, 1995.
- [654] The PURL Team, Persistent Uniform Resource Locator (PURL). <http://purl.oclc.org/>.
- [655] Yonggang Qiu and H. P. Frei. Concept based query expansion. In *Proc. of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA, USA, 1993.
- [656] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. Querying semistructured heterogeneous information. In *Deductive and Object-Oriented Databases, Proc. of the DOOD '95 Conference*, pages 319–344, Singapore, December 1995. Springer-Verlag.
- [657] F. Rabitti and P. Savino. Retrieval of multimedia document by imprecise query specification. In F. Bancilhon, C. Thanos, and D. Tsichritzis, editors, *Proc. of the Int. Conference on Extended Database Technologies*, number 416 in LNCS, pages 203–218, Venice, Italy, 1990.

- [658] T. Radecki. Mathematical model of information retrieval system based on the concept of fuzzy thesaurus. *Information Processing & Management*, 12:313–318, 1976.
- [659] T. Radecki. Mathematical model of time-effective information retrieval system based on the theory of fuzzy sets. *Information Processing & Management*, 13:109–116, 1977.
- [660] T. Radecki. Fuzzy set theoretical approach to document retrieval. *Information Processing & Management*, 15:247–259, 1979.
- [661] T. Radecki. On the inclusiveness of information retrieval systems with documents indexed by weighted descriptors. *Fuzzy Sets and Systems*, 5:159–176, 1981.
- [662] Tadeusz Radecki. Trends in research on information retrieval—the potential for improvements in conventional Boolean retrieval systems. *Information Processing & Management*, 24:219–227, 1988.
- [663] V. V. Raghavan, P. Bollmann, and G. S. Jung. Retrieval system evaluation using recall and precision: Problems and answers. In *Proc. of the 12th ACM SIGIR Conference*, pages 59–68, Cambridge, USA, June 1989.
- [664] V. V. Raghavan, G. S. Jung, and P. Bollmann. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Office and Information Systems*, 7(3):205–229, 1989.
- [665] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Sciences*, 37(5):279–287, 1986.
- [666] R. Rao and S. K. Card. The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 318–322, Boston, MA, USA, April 1994.
- [667] R. Rao, S. K. Card, H. D. Jelinek, J. D. Mackinlay, and G. G. Robertson. The Information Grid: A framework for building information retrieval and retrieval-centered applications. In *Proc. of the ACM Symposium on User Interface Software and Technology*, Monterey, CA, USA, Nov 1992.
- [668] E. Rasmussen. Clustering algorithms. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 419–442. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [669] David F. Redmiles. Reducing the variability of programmers’ performance through explained examples. In *Proc. of the ACM Conference on Human Factors in Computing Systems*, pages 67–73, 1993.
- [670] A. Renear. The digital library research agenda: What’s missing — and how humanities textbase projects can help. *D-Lib Magazine*, July/August 1997. <http://www.dlib.org/dlib/july97/07renear.html>.
- [671] Earl Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proc. of UIST’94, ACM Symposium on User Interface Software and Technology*, pages 3–12, New York, 1994.
- [672] Paul Resnick and Hal Varian. Introduction: Special issue on collaborative filtering. *Communications of the ACM*, 40(3):56–58, March 1997.
- [673] B. Ribeiro-Neto and R. Barbosa. Query performance in tightly coupled distributed digital libraries. In *Proc. of the 3rd Annual Int. ACM Conference on Digital Libraries*, Pittsburgh, PA, USA, 1998.
- [674] Berthier A. Ribeiro-Neto and Richard Muntz. A belief network model for IR. In *Proc. of the 19th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, Zurich, Switzerland, 1996.

- [675] J. Rissanen and G. G. Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23:149–162, 1979.
- [676] George C. Robertson, Stuart K. Card, and Jock D. Mackinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4):56–71, 1993.
- [677] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, 27(3):129–146, 1976.
- [678] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System — Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.
- [679] Peter Roget. *Roget's II The New Thesaurus*. Houghton Mifflin Company, Boston, USA, 1988.
- [680] M. Roscheisen, C. Mogensen, and T. Winograd. Interaction design for shared World-Wide Web annotations. Stanford Digital Library Project Working Paper, February 1995. <http://walrus.stanford.edu/diglib/pub/reports/brio-chi95.html> .
- [681] David E. Rose, Richard Mander, Tim Oren, Dulce B. Poncelón, Gitta Salomon, and Yin Yin Wong. Content awareness in a file system interface: Implementing the 'pile' metaphor for organizing information. In *Proc. of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 260–269, 1993.
- [682] Louis Rosenfeld and Peter Morville. *Information Architecture for the World Wide Web*. O'Reilly & Associates, 1998.
- [683] A. Ross. Library functions, scholarly communication, and the foundation of the digital library: Laying claim to the control zone. *The Library Quarterly*, 66:239–265, July 1996.
- [684] Mary Beth Rosson, John M. Carroll, and Rachel K. E. Bellamy. Smalltalk Scaffolding: A case study of minimalist instruction. In *Proc. of the ACM Conference on Human Factors in Computing Systems*, pages 423–429, 1990.
- [685] Steven F. Roth, Mei C. Chuah, Stephan Kerpedjiev, John A. Kolojejchick, and Peter Lucas. Towards an information visualization workspace: Combining multiple means of expression. *Human-Computer Interaction*, 12(1-2):131–185, 1997.
- [686] Nick Roussopoulos, Steve Kelley, and F. Vincent. Nearest neighbor queries. In *Proc. of ACM-SIGMOD*, pages 71–79, San Jose, CA, May 1995.
- [687] N. Roussopoulos, C. Faloutsos, and T. Sellis. An efficient pictorial database system for PSQL. *IEEE Transactions on Software Engineering*, 14(5):639–650, May 1988.
- [688] Jennifer Rowley. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2):108–119, 1994.
- [689] Mary Beth Ruskai, Gregory Beylkin, Ronald Coifman, Ingrid Daubechies, Stephane Mallat, Yves Meyer, and Louise Raphael. *Wavelets and their Applications*. Jones and Bartlett Publishers, Boston, MA, 1992.
- [690] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The cost structure of sensemaking. In *Proc. of ACM Conference on Human Factors in Computing Systems*, pages 269–276, 1993.
- [691] W. M. Sachs. An approach to associative retrieval through the theory of fuzzy sets. *Journal of the American Society for Information Sciences*, pages 85–87, 1976.
- [692] M. Sahami, S. Yusufali, and M. Q. W. Baldonado. SONIA: A service for organizing networked information autonomously. In *Proc. of the 3rd Annual Conference on Digital Libraries*, pages 200–209, New York, USA, 1998.

- [693] A. Salminen and F. Tompa. PAT expressions: an algebra for text search. Technical Report OED-92-02, UW Centre for the New Oxford English Dictionary, July 1992.
- [694] G. Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, October 1963.
- [695] G. Salton. *The SMART Retrieval System — Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.
- [696] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
- [697] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, January 1968.
- [698] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.
- [699] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Sciences*, 26(1):33–44, 1975.
- [700] Gerard Salton. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.
- [701] Gerard Salton and Chris Buckley. Parallel text search methods. *Communications of the ACM*, 31(2):202–215, February 1988.
- [702] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *J. of the American Society for Information Science*, 41(4):288–297, 1990.
- [703] Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, November 1983.
- [704] Gerard Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372, 1973.
- [705] P. Samuelson. Copyright and digital libraries. In *Proc. of the 2nd ACM International Conference on Digital Libraries*, pages 113–114, Philadelphia, PA, USA, 1997.
- [706] David Sankoff and Joseph B. Kruskal. *Time warps, string edits and macromolecules: The theory and practice of sequence comparisons*. Addison-Wesley, Reading, MA, 1983.
- [707] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3):161–176, 1988.
- [708] S. Sato and T. Kanade. Name-it: Association of Face and Name in Video. Technical Report CMU-CS-96-205, CMU School of Computer Science, 1996.
- [709] L. Schamber. What is a document? Rethinking the concept in uneasy times. *Journal of the American Society for Information Science*, 47:669–671, September 1996.
- [710] B. Schatz and H. Chen. Building large-scale digital libraries: Guest editors' introduction to theme issue on the US Digital Library Initiative. *IEEE Computer*, 29(5):22–26, May 1996. <http://computer.org/computer/dli/>.
- [711] B. R. Schatz. Information retrieval in digital libraries: Bringing search to the Net. *Science*, 275:327–335, January 1997.
- [712] Manfred Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W. H. Freeman and Company, New York, 1991.
- [713] E. S. Schwartz and B. Kallick. Generating a canonical prefix encoding. *Communications of the ACM*, 7:166–169, 1964.
- [714] S. Sclaroff, L. Taycher, and M. La Cascia. ImageRover: A content-based image browser for the World Wide Web. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, San Juan, Puerto Rico, June 1997.

- [715] E. Selberg and O. Etzioni. Multi-service search and comparison using the MetaCrawler. *4th Int. WWW Conference*, December 1995.
- [716] P. Sellers. The theory and computation of evolutionary distances: pattern recognition. *Journal of Algorithms*, 1:359–373, 1980.
- [717] T. Sellis, N. Roussopoulos, and C. Faloutsos. The R^+ -tree: A dynamic index for multi-dimensional objects. In *Proc. of the 13th Int. Conference on Very Large Data Bases*, pages 507–518, Brighton, England, 1987.
- [718] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:398–403, 1948.
- [719] Jacob Shapiro, Vladimir G. Voiskunskii, and Valery J. Frants. *Automated Information Retrieval : Theory and Text-Only Methods*. Academic Press, 1997.
- [720] Upendra Shardanand and Pattie Maes. Social information filtering: Algorithms for automating Word of Mouth. In *Proc. of ACM Conference on Human Factors in Computing Systems*, pages 210–217, Denver, CO, USA, 1995.
- [721] W. M. Shaw, J. B. Wood, R. E. Wood, and H. R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.
- [722] Frank M. Shipman, III, Catherine C. Marshall, and Thomas P. Moran. Finding and using implicit structure in human-organized spatial layouts of information. In *Proc. of ACM Conference on Human Factors in Computing Systems*, volume 1, pages 346–353, Denver, CO, USA, 1995.
- [723] N. Shivakumar and H. García-Molina. Finding near-replicas of documents on the Web. In *Workshop on Web Databases*, Valencia, Spain, March 1998.
- [724] Ben Shneiderman. The eyes have it: A task by data type taxonomy. In *Proc. of IEEE Symp. Visual Languages 96*, pages 336–343, Boulder, CO, USA, 1996.
- [725] Ben Shneiderman. *Designing the User Interface: Strategies for Effective Human-computer Interaction*. Addison-Wesley, Reading, MA, 1997.
- [726] Ben Shneiderman, Donald Byrd, and W. Bruce Croft. Sorting out searching: A user-interface framework for text searches. *Communications of the ACM*, 41(4):95–98, 1998.
- [727] Ben Shneiderman and Greg Kearsley. *Hypertext Hands-On! An Introduction to a New Way of Organizing and Accessing Information*. Addison-Wesley Publishing Co., Reading, MA, 1989.
- [728] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large alta vista query log. Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, CA, USA, 1998.
- [729] Singapore advanced research and education network (singaren). <http://www.singaren.net.sg/>.
- [730] H. Small. Co-citation in the scientific literature: A new measure of relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.
- [731] H. Small. The relationship of information science to the social sciences: A co-citation analysis. *Information Processing & Management*, 17(1):39–50, 1981.
- [732] H. Small and M. E. D. Koenig. Journal clustering using a bibliographic coupling method. *Information Processing & Management*, 13(5):277–288, 1977.
- [733] D. C. Smith and S.-F. Chang. Searching for images and videos on the World-Wide Web. *IEEE Multimedia Magazine*, 4(3):12–20, 1997.
- [734] T. R. Smith. The Meta-information environment of digital libraries. *D-Lib Magazine*, July/August 1996. <http://www.dlib.org/dlib/july96/new/07smith.html>.

- [735] Dagobert Soergel. *Indexing Languages and Thesauri: Construction and Maintenance*. Melville Publishing Co., Los Angeles, CA, 1974.
- [736] Ellen Spertus. ParaSite: Mining structural information on the Web. In *6th Int. WWW Conference*, Santa Clara, CA, USA, April 1997.
- [737] M. Spiliopoulou and L. Faulstich. WUM — A tool for WWW utilization analysis. In *Workshop on Web Databases*, pages 109–115, Valencia, Spain, March 1998.
- [738] Anselm Spoerri. InfoCrystal: A visual tool for information retrieval & management. In *Proc. of Information Knowledge and Management '93*, pages 11–20, Washington, DC, Nov 1993.
- [739] P. Srinivasdan. Thesaurus construction. In W. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 161–218. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [740] Craig Stanfill. Partitioned posting files: A parallel inverted file structure for information retrieval. In *Proc. 13th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 413–428, Brussels, Belgium, 1990.
- [741] Craig Stanfill. Parallel information retrieval algorithms. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval Data Structures & Algorithms*, pages 459–497. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [742] Craig Stanfill and Brewster Kahle. Parallel free-text search on the Connection Machine system. *Communications of the ACM*, 29(12):1229–1239, December 1986.
- [743] Craig Stanfill and Brewster Kahle. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.
- [744] Craig Stanfill, Robert Thau, and David Waltz. A parallel indexed algorithm for information retrieval. In *Proc. 12th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 88–97, Cambridge, USA, June 1989.
- [745] Stanford DLI Team, Stanford University Digital Libraries Project. <http://www.diglib.stanford.edu/diglib/>.
- [746] M. Stefik. Trusted systems. *Scientific American*, March 1997. <http://www.sciam.com/-0397issue/0397stefik.html>.
- [747] J. A. Storer. *Data Compression: Methods and Theory*. Computer Science Press, 1988.
- [748] Tomek Strzalkowski, editor. *Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999. To appear.
- [749] D. Sullivan. Search Engine Watch: Main page. <http://www.searchenginewatch.com>, 1997.
- [750] K. Summers. Toward a taxonomy of logical document structures. In *DAGS95: Electronic Publishing and the Information Superhighway, May 30–June 2, 1995*, 1995. <http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html>.
- [751] D. Sunday. A very fast substring search algorithm. *Communications of the ACM*, 33(8):132–142, August 1990.
- [752] Elaine Svenonius. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5):331–340, 1986.
- [753] R. Swick. Resource Description Framework (RDF), June 1998. <http://www.w3.org/RDF>.
- [754] J. Tague-Sutcliffe. Measuring the informativeness of a retrieval process. In *Proc. of the 15th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 23–36, Copenhagen, Denmark, 1992.
- [755] V. A. Tahani. A fuzzy model of document retrieval systems. *Information Processing & Management*, 12:177–187, 1976.

- [756] TEI. A gentle introduction to SGML. Technical report, Text Encoding Initiative, 1996. <http://www.sil.org/sgml/gentle.html>.
- [757] Carol Tenopir and Jeff Barry. Database marketplace: data dealers face stormy weather. *Library Journal*, 123(9):38–46, 1998.
- [758] Carol Tenopir and Pamela Cahn. TARGET & FREESTYLE: DIALOG and Mead join the relevance ranks. *Online*, 18(3):31–47, 1994.
- [759] C. Thanos, editor. *Multimedia Office Filing: The MULTOS Approach*. Elsevier Science Publishers B.V. (North Holland), 1990.
- [760] P. Thompson, H. Turtle, B. Yang, and J. Flood. TREC-3 ad-hoc retrieval and routing experiments using the WIN system. In *Proc. of the Text REtrieval Conference*, pages 211–218, Gaithersburg, MD, USA, 1995.
- [761] R. H. Thompson and B. W. Croft. Support for browsing in an intelligent text retrieval system. *International Journal of Man-Machine Studies*, 30(6):639–668, 1989.
- [762] Anthony Tomasic and Hector García-Molina. Caching and database scaling in distributed shared-nothing information retrieval systems. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 129–138, Washington, DC, USA, May 1993.
- [763] Anthony Tomasic and Hector García-Molina. Performance of inverted indices in shared-nothing distributed text document information retrieval systems. In *Proc. of the 2nd Int. Conf. on Parallel and Distributed Information Systems*, pages 8–17, San Diego, USA, 1993.
- [764] Anthony Tomasic and Hector García-Molina. Performance issues in distributed shared-nothing information retrieval systems. *Information Processing & Management*, 32(6):647–665, 1996.
- [765] Anthony Tomasic, Luis Gravano, Calvin Lue, Peter Schwarz, and Laura Haas. Data structures for efficient broker implementation. *ACM Transactions on Information Systems*, 15(3):223–253, 1997.
- [766] A. Tombros and M. Sanderson. Advantages of query biased summaries in Information Retrieval. In B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10, Melbourne, Australia, 1998.
- [767] Geoffrey Towell, Ellen M. Voorhees, Narendra K. Gupta, and Ben Johnson-Laird. Learning collection fusion strategies for information retrieval. In *Proc. of the 12th Int. Conference on Machine Learning*, pages 540–548. Morgan Kaufmann, 1995.
- [768] The TREC NIST site, 1998. <http://trec.nist.gov>.
- [769] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Chelshire, CT, 1983.
- [770] Edward Tufte. *Envisioning Information*. Graphics Press, Chelshire, CT, 1990.
- [771] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.
- [772] Howard Turtle and W. Bruce Croft. Inference networks for document retrieval. In *Proc. of the 13th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–24, Brussels, Belgium, 1990.
- [773] L. A. Tweedie, R. Spence, D. Williams, and R. Bhogal. The attribute explorer. In *Proc. of the Video Track of the ACM Conference on Human Factors in Computing Systems*, pages 435–436, Boston, MA, USA, April 1994.
- [774] UC Berkeley Digital Library Project, About MVD version 1.0alpha3. <http://elib.cs.berkeley.edu/java/help/About.html>.
- [775] UC Berkeley DLI Team, UC Berkeley Digital Library Project. <http://elib.cs.berkeley.edu/>.

- [776] UC Santa Barbara DLI Team, Alexandria Digital Library. <http://alexandria.sdc.ucsb.edu/>.
- [777] UIUC DLI Team, University of Illinois at Urbana-Champaign Digital Libraries. <http://dli.grainger.uiuc.edu/default.htm>.
- [778] The UK Office for Library and Information Networking, Electronic Libraries Programme, eLib, March 1998. <http://www.ukoln.ac.uk/services/elib/>.
- [779] E. Ukkonen. Approximate string matching over suffix trees. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors, *Proc. of Combinatorial Pattern Matching*, number 684 in LNCS, pages 228–242, Padova, Italy, 1993. Springer-Verlag.
- [780] E. Ukkonen. Constructing suffix trees on-line in linear time. *Algorithmica*, 14(3):249–260, Sep 1995.
- [781] Esko Ukkonen. Finding approximate patterns in strings. *Journal of Algorithms*, 6:132–137, 1985.
- [782] G. M. Underwood, P. P. Maglio, and R. Barrett. User-centered push for timely information delivery. In *7th WWW Conference*, Brisbane, Australia, 1998.
- [783] Unicode Consortium, Unicode. <http://www.unicode.org/>.
- [784] University of Michigan DLI Team, Univ. of Michigan Digital Library Project. <http://www.si.umich.edu/UMDL/>.
- [785] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.
- [786] J. Verhoeff, W. Goffmann, and Jack Belzer. Inefficiency of the use of Boolean functions for information retrieval systems. *Communications of the ACM*, 4(12):557–558, 594, December 1961.
- [787] Charles L. Viles and James C. French. Dissemination of collection wide information in a distributed information retrieval system. In *Proc. 18th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 12–20, Seattle, WA, USA, July 1995.
- [788] S. Vinoski. CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications Magazine*, 14(2), February 1997.
- [789] J. S. Vitter. Algorithm 673: Dynamic Huffman coding. *ACM Transactions on Mathematical Software*, 15:158–167, 1989.
- [790] A. N. Vo and A. Moffat. Compressed inverted files with reduced decoding overhead. In *Proc. of the ACM-SIGIR Int. Conference on Research and Development in Information Retrieval*, pages 290–297, Melbourne, Australia, 1998.
- [791] E. M. Voorhees and R. M. Tong. Multiple search engines in database merging. In *Proc. of the 2nd ACM International Conference on Digital Libraries*, pages 93–102, Philadelphia, PA, USA, 1997.
- [792] Ellen M. Voorhees, Narendra K. Gupta, and Ben Johnson-Laird. The collection fusion problem. In Donna K. Harman, editor, *The Third Text REtrieval Conference (TREC-3)*, pages 95–104, Gaithersburg, MD, USA, 1995. Dept. of Commerce, National Institute of Standards and Technology. Special Publication 500-226.
- [793] E. M. Voorhees. *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval*. PhD thesis, Cornell University, 1986.
- [794] E. M. Voorhees and D. K. Harman. Overview of the 6th text retrieval conference (TREC-6). In E. M. Voorhees and D. K. Harman, editors, *Proc. of the Sixth Text REtrieval Conference*, pages 1–24. NIST Special Publication, Gaithersburg, MD, USA, 1997.
- [795] W3C. Extensible markup language (XML) 1.0. Technical report, WWW Consortium (W3C), 1998. <http://www.w3.org/TR/1998/REC-xml-19980210>.

- [796] W3C. HTML 4.0 specification. Technical report, WWW Consortium (W3C), 1998. <http://www.w3.org/TR/1998/REC-html40-19980424/>.
- [797] W3C. Web characterization activity, 1998. <http://www.w3.org/WCA/>.
- [798] W3C. XML linking language (XLink). Technical report, WWW Consortium (W3C), 1998. <http://www.w3.org/TR/1998/WD-xlink-19980303>.
- [799] W3C. XSL requirements summary. Technical report, WWW Consortium (W3C), 1998. <http://www.w3.org/TR/1998/WD-XSLReq-19980511>.
- [800] Howard D. Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. Intelligent access to digital video: Informedia project. *IEEE Computer*, 29(5):46–52, May 1996.
- [801] Marilyn A. Walker, Jeanne Fromer, Giuseppe Di Fabbrizio, Craig Mestel, and Don Hindle. What can I say: Evaluating a spoken language interface to email. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 582–589, Los Angeles, CA, March 1998.
- [802] Gregory K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):31–44, April 1991.
- [803] S. Wartick. Boolean operations. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures & Algorithms*, pages 264–292. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.
- [804] John A. Waterworth and Mark H. Chignell. A model of information exploration. *Hypermedia*, 3(1):35–58, 1991.
- [805] WebTaxi: Main Page. <http://www.webtaxi.com>, 1998.
- [806] S. Weibel, J. Godby, E. Miller, and R. Daniel. OCLC/NCSA Metadata Workshop Report: The Essential Elements of Network Object Description, March 1995. <http://purl.oclc.org/oclc/rsch/metadataI>.
- [807] S. Weibel and E. Miller. Dublin Core Metadata, 1997. URL: http://purl.org/metadata/dublin_core.
- [808] Andreas S. Weigend and Neil A. Gerschenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison Wesley, 1994.
- [809] Bella H. Weinberg. Bibliographic coupling: A review. *Information Storage and Retrieval*, 10(5-6):189–196, 1974.
- [810] R. Weiss, B. Vélez, M. Sheldon, C. Nemprenpre, P. Szilagyi, and D. K. Gifford. HyPursuit: A hierarchical network engine that exploits content-link hypertext clustering. In *Proc. of the 7th ACM Conference on Hypertext and Hypermedia*, pages 180–193, Washington, DC, USA, 1996.
- [811] H. G. Wells. World Brain: The idea of a permanent world encyclopedia. Contribution to the New Encyclopedie Francaise, 1937. http://sherlock.berkeley.edu/wells/world_brain.html.
- [812] H. D. White and K. W. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186, 1989.
- [813] Howard S. White, editor. *Library Technology Reports*, volume 33(5). American Library Association, Chicago, IL, September/October 1997.
- [814] Robert Wilensky, Yigal Arens, and David N. Chin. Talking to UNIX in English: An overview of UC. *Communications of the ACM*, 27(6), 1984.
- [815] R. Wilkinson and P. Hingston. Using the cosine measure in a neural network for document retrieval. In *Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–210, Chicago, USA, Oct 1991.
- [816] Peter Willet. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.

- [817] J. Williams. *Bots and other Internet Beasts*. Prentice Hall, Englewood Cliffs, NJ, 1996.
- [818] Michael David Williams. What makes rabbit run? *International Journal of Man-Machine Studies*, 21(4):333–352, 1984.
- [819] W. Willinger and V. Paxson. Where mathematics meets the Internet. *Notices of the AMS*, 45(8):961–970, 1998.
- [820] T. Winograd. Conceptual models for comparison of digital library systems and approaches. Stanford Digital Library Project Working Paper, July 1995. <http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC13.html>.
- [821] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, and Anne Schur. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proc. of the Information Visualization Symposium 95*, pages 51–58. IEEE Computer Society Press, 1995.
- [822] N. Wiseman. Implementing a national access management system for electronic services: Technology alone is not enough. *D-Lib Magazine*, March 1998. <http://www.dlib.org/dlib/march98/wiseman/03wiseman.html>.
- [823] I. H. Witten, R. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–541, 1987.
- [824] Ian H. Witten, Craig Nevill-Manning, Roger McNab, and Sally Jo Cunningham. A public library based on full-text retrieval. *Communications of the ACM*, 41(4):71–75, 1998.
- [825] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York, 1994.
- [826] Kent Wittenburg and Eric Sigman. Integration of browsing, searching, and filtering in an applet for Web information access. In *Proc. of the ACM Conference on Human Factors in Computing Systems, Late Breaking Track*, Atlanta, GA, USA, 1997. <http://www1.acm.org:82/sigs/sigchi/chi97/proceedings/short-talk/kw.htm>.
- [827] D. Woelk, W. Kim, and W. Luther. An object-oriented approach to multimedia databases. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 312–325, Washington, DC, USA, June 1986.
- [828] L. Wong. BioKleisli. <http://corona.iss.nus.sg:8080/biokleisli.html>.
- [829] L. Wong. BioKleisli Architecture. <http://sdmc.krdl.org.sg/kleisli/kleisli/Architecture.html>.
- [830] S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):39–68, 1995.
- [831] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong. On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems*, 12(2):299–321, 1987.
- [832] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector space model in information retrieval. In *Proc. 8th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25, New York, USA, 1985.
- [833] A. Wood. DC-4: NLA/DSTC/OCLC Dublin Core Down Under / The 4th Dublin Core Metadata Workshop, March 1997. <http://www.dstc.edu.au/DC4/>.
- [834] A. Woodruff, P. Aoki, E. Brewer, P. Gauthier, and L. Rowe. An investigation of documents from the World Wide Web. In *5th WWW Conference*, Paris, France, 1996.
- [835] H. Wu and G. Salton. The estimation of term relevance weights using relevance feedback. *Journal of Documentation*, 37(4):194–214, 1981.
- [836] S. Wu and U. Manber. Agrep — a fast approximate pattern-matching tool. In *Proc. of USENIX Technical Conference*, pages 153–162, San Francisco, CA, USA, 1992.

- [837] S. Wu and U. Manber. Fast text searching allowing errors. *Communications of the ACM*, 35(10):83–91, October 1992.
- [838] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, 1996.
- [839] Yahoo!: Main page. <http://www.yahoo.com>, 1995.
- [840] Byoung-Kee Yi, H. V. Jagadish, and Christos Faloutsos. Efficient retrieval of similar time sequences under time warping. In *ICDE'98*, pages 201–208, Orlando, Florida, Feb 1998.
- [841] Degi Young and Ben Shneiderman. A graphical filter/flow model for Boolean queries: An implementation and experiment. *Journal of the American Society for Information Science*, 44(6):327–339, July 1993. With permission from University of Maryland, USA, Shneiderman & Young, 1993.
- [842] C. T. Yu and G. Salton. Precision weighting — an effective automatic indexing method. *Journal of the ACM*, 23(1):76–88, January 1976.
- [843] C. T. Yu, C. Buckley, K. Lam, and G. Salton. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2(4):129–154, October 1983.
- [844] B. Yuwono and D. L. Lee. Search and ranking algorithms for locating resources on World Wide Web. In *Proc. of the Int. Conference on Data Engineering (ICDE)*, pages 164–171, New Orleans, USA, 1996.
- [845] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the Internet. In *Proc. of the 5th Int. Conf. on Databases Systems for Advanced Applications (DASFAA)*, pages 41–50, Melbourne, Australia, 1997.
- [846] L. A. Zadeh. Fuzzy sets. In D. Dubois, H. Prade, and R. R. Yager, editors, *Readings in Fuzzy Sets for Intelligent Systems*. Morgan Kaufmann, 1993.
- [847] G. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.
- [848] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.
- [849] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.
- [850] J. Zobel and A. Moffat. Adding compression to a full text retrieval system. *Software Practice and Experience*, 25(8):891–903, August 1995.
- [851] Justin Zobel, Alistair Moffat, and Kotagiri Ramamohanarao. Inverted files versus signature files for text indexing. Technical Report CITRI/TR-95-5, Collaborative Information Technology Research Institute, Department of Computer Science, Royal Melbourne Institute of Technology, Australia, July 1995.
- [852] Albert Y. Zomaya, editor. *Parallel and Distributed Computing Handbook*. McGraw-Hill, New York, 1996.

Note: some of the bibliography is only available on the Web, in particular conferences and electronic journals related to the Web itself or digital libraries. Most of these references are available on the Web page of the book (see Chapter 1).