

# Contents

---

<b>Preface</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Biographies</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
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
<b>2 Modeling</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 A Taxonomy of Information Retrieval Models . . . . .	20
2.3 Retrieval: Ad hoc and Filtering . . . . .	21

x CONTENTS

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
<b>3</b>	<b>Retrieval Evaluation</b> . . . . .	<b>73</b>
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
<b>4</b>	<b>Query Languages</b> . . . . .	<b>99</b>
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
<b>5</b>	<b>Query Operations</b> . . . . .	<b>117</b>
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
<b>6</b>	<b>Text and Multimedia Languages and Properties</b> . . . . .	<b>141</b>
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

xii CONTENTS

6.5.4	HyTime . . . . .	159
6.6	Trends and Research Issues . . . . .	160
6.7	Bibliographic Discussion . . . . .	162
<b>7</b>	<b>Text Operations</b> . . . . .	<b>163</b>
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
<b>8</b>	<b>Indexing and Searching</b> . . . . .	<b>191</b>
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
<b>9</b>	<b>Parallel and Distributed IR</b>	<b>229</b>
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
<b>10</b>	<b>User Interfaces and Visualization</b>	<b>257</b>
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

xiv CONTENTS

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
<b>11</b>	<b>Multimedia IR: Models and Languages</b>	<b>325</b>
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
<b>12</b>	<b>Multimedia IR: Indexing and Searching</b>	<b>345</b>
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
<b>13</b>	<b>Searching the Web</b>	<b>367</b>
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
<b>14</b>	<b>Libraries and Bibliographical Systems</b>	<b>397</b>
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

xvi CONTENTS

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
<b>15</b>	<b>Digital Libraries</b>	<b>415</b>
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
	<b>Appendix: Porter's Algorithm</b>	<b>433</b>
	<b>Glossary</b>	<b>437</b>
	<b>References</b>	<b>455</b>
	<b>Index</b>	<b>501</b>



# 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 SBB'D'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

## 2 INTRODUCTION

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