

# Modern Information Retrieval

(the concepts and technology behind search)

Second edition



Ricardo Baeza-Yates

Berthier Ribeiro-Neto

# Modern Information Retrieval

**PEARSON**

We work with leading authors to develop the  
strongest educational materials in information retrieval,  
bringing cutting-edge thinking and best  
learning practice to a global market.

Under a range of well-known imprints, including  
Addison Wesley, we craft high quality print and  
electronic publications which help readers to understand  
and apply their content, whether studying or at work.

To find out more about the complete range of our  
publishing, please visit us on the World Wide Web at:  
[www.pearsoned.co.uk](http://www.pearsoned.co.uk)

# Modern Information Retrieval

the concepts and technology behind search  
Second edition

Ricardo Baeza-Yates  
Berthier Ribeiro-Neto

Addison Wesley  
is an Imprint of

**PEARSON**

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

Pearson Education Limited

Edinburgh Gate

Hatton

Essex CM9 5JE

England

and Associated Companies throughout the world

Visit us on the World Wide Web at:

[www.pearsoned.co.uk](http://www.pearsoned.co.uk)

First published 1999

Second edition published 2011

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

This edition copyright © Pearson Education Limited 2011

The rights of Ricardo Basco-Yates and Berthier Ribeiro-Neto to be identified as 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, Saffron House, 6-10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

Pearson Education is not responsible for the content of third party internet sites.

Microsoft product screenshots reprinted with permission from Microsoft Corporation.

ISBN: 978-0-321-41691-9

#### **British Library Cataloguing-in-Publication Data**

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

#### **Library of Congress Cataloguing-in-Publication Data**

A catalog record for this book is available from the Library of Congress

10 9 8 7 6 5 4 3 2

13 12 11

Typeset in CMR10 10/12 by 23

Printed and bound in Great Britain by Henry Ling Ltd at the Dorset Press, Dorchester,  
Dorset

About the cover: the word cloud was produced using the frequency of keywords in the text  
content of the book, using wordcloud.net

# Contents

---

Preface to the Second Edition	xix
Preface to the First Edition	xxi
Authors' Acknowledgements to the Second Edition	xxiii
Authors' Acknowledgements to the First Edition	xxv
Publishers' Acknowledgements	xxvii
<b>1 Introduction</b>	
1.1 Information Retrieval . . . . .	1
1.1.1 Early Developments . . . . .	1
1.1.2 Information Retrieval in Libraries and Digital Libraries . . . . .	1
1.1.3 IR at the Center of the Stage . . . . .	3
1.2 The IR Problem . . . . .	3
1.2.1 The User's Task . . . . .	4
1.2.2 Information versus Data Retrieval . . . . .	5
1.3 The IR System . . . . .	5
1.3.1 Software Architecture of the IR System . . . . .	5
1.3.2 The Retrieval and Ranking Processes . . . . .	7
1.4 The Web . . . . .	8
1.4.1 A Brief History . . . . .	8
1.4.2 The e-Publishing Era . . . . .	9
1.4.3 How the Web Changed Search . . . . .	10
1.4.4 Practical Issues on the Web . . . . .	11
1.5 Organization of the Book . . . . .	12
1.5.1 Errata of the Book . . . . .	12
1.5.2 Book Contents . . . . .	13
1.6 The Book Web Site: A Teaching Resource . . . . .	16
1.7 Bibliographic Discussion . . . . .	17
<b>2 User Interfaces for Search by Mark Weibel</b>	21
2.1 Introduction . . . . .	21
2.2 How People Search . . . . .	21

2.2.1	Information Lookup versus Exploratory Search . . . . .	22
2.2.2	Classic versus Dynamic Model of Information Seeking . . . . .	23
2.2.3	Navigation versus Search . . . . .	24
2.2.4	Observations of the Search Process . . . . .	24
2.3	Search Interfaces Today . . . . .	25
2.3.1	Getting Started . . . . .	25
2.3.2	Query Specification . . . . .	26
2.3.3	Query Specification Interfaces . . . . .	27
2.3.4	Retrieval Results Display . . . . .	29
2.3.5	Query Reformulation . . . . .	32
2.3.6	Organizing Search Results . . . . .	35
2.4	Visualization in Search Interfaces . . . . .	40
2.4.1	Visualizing Boolean Syntax . . . . .	42
2.4.2	Visualizing Query Terms within Retrieval Results . . . . .	43
2.4.3	Visualizing Relationships Among Words and Documents . . . . .	47
2.4.4	Visualization for Text Mining . . . . .	49
2.5	Design and Evaluation of Search Interfaces . . . . .	50
2.6	Trends and Research Issues . . . . .	54
2.7	Bibliographic Discussion . . . . .	54
<b>3</b>	<b>Modeling</b> . . . . .	<b>57</b>
3.1	IR Models . . . . .	57
3.1.1	Modeling and Ranking . . . . .	57
3.1.2	Characterization of an IR Model . . . . .	58
3.1.3	A Taxonomy of IR Models . . . . .	59
3.2	Classic Information Retrieval . . . . .	61
3.2.1	Basic Concepts . . . . .	61
3.2.2	The Boolean Model . . . . .	64
3.2.3	Term Weighting . . . . .	66
3.2.4	TF-IDF Weights . . . . .	68
3.2.5	Document Length Normalization . . . . .	75
3.2.6	The Vector Model . . . . .	77
3.2.7	The Probabilistic Model . . . . .	79
3.2.8	Brief Comparison of Classic Models . . . . .	86
3.3	Alternative Set-Theoretic Models . . . . .	87
3.3.1	Set-Based Model . . . . .	87
3.3.2	Extended Boolean Model . . . . .	92
3.3.3	Fuzzy Set Model . . . . .	95
3.4	Alternative Algebraic Models . . . . .	98
3.4.1	Generalized Vector Space Model . . . . .	98
3.4.2	Latent Semantic Indexing Model . . . . .	101
3.4.3	Neural Network Model . . . . .	103
3.5	Alternative Probabilistic Models . . . . .	104
3.5.1	BM25 . . . . .	104
3.5.2	Language Models . . . . .	107
3.5.3	Divergence from Randomness . . . . .	113
3.5.4	Bayesian Network Models . . . . .	116
3.6	Other Models . . . . .	124

3.6.1	The Hypertext Model . . . . .	124
3.6.2	Web-based Models . . . . .	125
3.6.3	Structured Text Retrieval . . . . .	126
3.6.4	Multimedia Retrieval . . . . .	126
3.6.5	Enterprise and Vertical Search . . . . .	126
3.7	Trends and Research Issues . . . . .	127
3.8	Bibliographic Discussion . . . . .	128
<b>4</b>	<b>Retrieval Evaluation</b> . . . . .	<b>131</b>
4.1	Introduction . . . . .	131
4.2	The Crossfield Paradigm . . . . .	132
4.2.1	A Brief History . . . . .	132
4.2.2	Reference Collections . . . . .	134
4.3	Retrieval Metrics . . . . .	134
4.3.1	Precision and Recall . . . . .	135
4.3.2	Single Value Summaries: P@n, MAP, MRR, F . . . . .	139
4.3.3	User-Oriented Measures . . . . .	144
4.3.4	DCG: Discounted Cumulative Gain . . . . .	145
4.3.5	BPREF: Binary Preferences . . . . .	150
4.3.6	Rank Correlation Metrics . . . . .	153
4.4	Reference Collections . . . . .	158
4.4.1	The TREC Collections . . . . .	159
4.4.2	Other Reference Collections . . . . .	166
4.4.3	Other Small Test Collections . . . . .	167
4.5	User-Based Evaluation . . . . .	168
4.5.1	Human Experimentation in the Lab . . . . .	168
4.5.2	Side-by-Side Panels . . . . .	169
4.5.3	A/B Testing . . . . .	169
4.5.4	Crowdsourcing . . . . .	170
4.5.5	Evaluation using Clickthrough Data . . . . .	171
4.6	Practical Considerations . . . . .	173
4.7	Trends and Research Issues . . . . .	174
4.8	Bibliographic Discussion . . . . .	174
<b>5</b>	<b>Relevance Feedback and Query Expansion</b> . . . . .	<b>177</b>
5.1	Introduction . . . . .	177
5.2	A Framework for Feedback Methods . . . . .	178
5.3	Explicit Relevance Feedback . . . . .	180
5.3.1	Relevance Feedback for the Vector Model: Rocchio Method . . . . .	181
5.3.2	Relevance Feedback for the Probabilistic Model . . . . .	183
5.3.3	Evaluation of Relevance Feedback . . . . .	184
5.4	Explicit Feedback Through Clicks . . . . .	185
5.4.1	Eye Tracking and Relevance Judgements . . . . .	185
5.4.2	User Behavior . . . . .	186
5.4.3	Clicks as a Metric of User Preferences . . . . .	187
5.5	Implicit Feedback Through Local Analysis . . . . .	190
5.5.1	Implicit Feedback Through Local Clustering . . . . .	190
5.5.2	Implicit Feedback through Local Context Analysis . . . . .	193

5.6	Implicit Feedback Through Global Analysis . . . . .	195
5.6.1	Query Expansion based on a Similarity Thesaurus . . . . .	195
5.6.2	Query Expansion based on a Statistical Thesaurus . . . . .	198
5.7	Trends and Research Issues . . . . .	200
5.8	Bibliographic Discussion . . . . .	200
6	<b>Documents: Languages &amp; Properties</b> with Gonzalo Navarro and Niraj Zutani . . . . .	203
6.1	Introduction . . . . .	203
6.2	Metadata . . . . .	205
6.3	Document Formats . . . . .	206
6.3.1	Text . . . . .	206
6.3.2	Multimedia . . . . .	207
6.3.3	Graphics and Virtual Reality . . . . .	208
6.4	Markup Languages . . . . .	208
6.4.1	SGML . . . . .	209
6.4.2	HTML . . . . .	211
6.4.3	XML . . . . .	214
6.4.4	RDF: Resource Description Framework . . . . .	216
6.4.5	ByTime . . . . .	217
6.5	Text Properties . . . . .	218
6.5.1	Information Theory . . . . .	218
6.5.2	Modeling Natural Language . . . . .	219
6.5.3	Text Similarity . . . . .	222
6.6	Document Preprocessing . . . . .	223
6.6.1	Lexical Analysis of the Text . . . . .	224
6.6.2	Elimination of Stopwords . . . . .	226
6.6.3	Stemming . . . . .	226
6.6.4	Keyword Selection . . . . .	227
6.6.5	Thesauri . . . . .	228
6.7	Organizing Documents . . . . .	231
6.7.1	Taxonomies . . . . .	231
6.7.2	Folksonomies . . . . .	232
6.8	Text Compression . . . . .	233
6.8.1	Basic Concepts . . . . .	234
6.8.2	Statistical Methods . . . . .	234
6.8.3	Statistical Methods: Modeling . . . . .	235
6.8.4	Statistical Methods: Coding . . . . .	238
6.8.5	Dictionary Methods . . . . .	243
6.8.6	Preprocessing for Compression . . . . .	246
6.8.7	Comparing Text Compression Techniques . . . . .	248
6.8.8	Structured Text Compression . . . . .	249
6.9	Trends and Research Issues . . . . .	250
6.10	Bibliographical Discussion . . . . .	253
7	<b>Queries: Languages &amp; Properties</b> with Gonzalo Navarro . . . . .	255
7.1	Query Languages . . . . .	256

7.1.1	Keyword-based Querying . . . . .	256
7.1.2	Beyond Keywords . . . . .	259
7.1.3	Structural Queries . . . . .	262
7.1.4	Query Protocols . . . . .	265
7.2	Query Properties . . . . .	267
7.2.1	Characterizing Web Queries . . . . .	267
7.2.2	User Search Behavior . . . . .	269
7.2.3	Query Intent . . . . .	270
7.2.4	Query Topic . . . . .	272
7.2.5	Query Sessions and Messions . . . . .	273
7.2.6	Query Difficulty . . . . .	274
7.3	Trends and Research Issues . . . . .	278
7.4	Bibliographical Discussion . . . . .	279
<b>8</b>	<b>Text Classification</b>	<b>281</b>
	with Alvaro Gómez-Pérez	
8.1	Introduction . . . . .	281
8.2	A Characterization of Text Classification . . . . .	282
8.2.1	Machine Learning . . . . .	282
8.2.2	The Text Classification Problem . . . . .	283
8.2.3	Text Classification Algorithms . . . . .	284
8.3	Unsupervised Algorithms . . . . .	286
8.3.1	Clustering . . . . .	286
8.3.2	Naïve Text Classification . . . . .	290
8.4	Supervised Algorithms . . . . .	291
8.4.1	Decision Trees . . . . .	291
8.4.2	The k-NN Classifier . . . . .	299
8.4.3	The Rocchio Classifier . . . . .	300
8.4.4	Probabilistic Naïve Bayes Document Classification . . . . .	303
8.4.5	The SVM Classifier . . . . .	306
8.4.6	Ensemble Classifiers . . . . .	316
8.4.7	Final Remarks on Supervised Algorithms . . . . .	319
8.5	Feature Selection or Dimensionality Reduction . . . . .	320
8.5.1	Term-Class Incidence Table . . . . .	321
8.5.2	Term Document Frequency . . . . .	322
8.5.3	TF-IDF Weights . . . . .	322
8.5.4	Mutual Information . . . . .	323
8.5.5	Information Gain . . . . .	323
8.5.6	Chi Square . . . . .	324
8.5.7	Impact of Feature Selection . . . . .	325
8.6	Evaluation Metrics . . . . .	325
8.6.1	Contingency Table . . . . .	325
8.6.2	Accuracy and Error . . . . .	326
8.6.3	Precision and Recall . . . . .	327
8.6.4	F-measure and $F_1$ . . . . .	327
8.6.5	Cross-Validation . . . . .	329
8.6.6	Standard Collections . . . . .	329
8.7	Organizing the Classes – Taxonomies . . . . .	330

8.8 Trends and Research Issues . . . . .	333
8.9 Bibliographic Discussion . . . . .	334
<b>9 Indexing and Searching with Gonzalo Navarro</b>	<b>337</b>
9.1 Introduction . . . . .	337
9.2 Inverted Indexes . . . . .	340
9.2.1 Basic Concepts . . . . .	340
9.2.2 Full Inverted Indexes . . . . .	341
9.2.3 Searching . . . . .	345
9.2.4 Ranking . . . . .	348
9.2.5 Construction . . . . .	351
9.2.6 Compressed Inverted Indexes . . . . .	354
9.2.7 Structural Queries . . . . .	357
9.3 Signature Files . . . . .	357
9.4 Suffix Trees and Suffix Arrays . . . . .	360
9.4.1 Structure: Tries and Suffix Trees . . . . .	361
9.4.2 Searching for Simple Strings . . . . .	362
9.4.3 Searching for Complex Patterns . . . . .	363
9.4.4 Construction . . . . .	365
9.4.5 Compressed Suffix Arrays . . . . .	367
9.5 Sequential Searching . . . . .	372
9.5.1 Simple Strings: Horspool . . . . .	373
9.5.2 Complex Patterns: Automata and Bit-Parallelism . . . . .	375
9.5.3 Faster Bit-Parallel Algorithms . . . . .	379
9.5.4 Regular Expressions . . . . .	382
9.5.5 Multiple Patterns . . . . .	384
9.5.6 Approximate Searching . . . . .	385
9.5.7 Searching Compressed Text . . . . .	389
9.6 Multi-dimensional Indexing . . . . .	391
9.7 Trends and Research Issues . . . . .	393
9.8 Bibliographic Discussion . . . . .	394
<b>10 Parallel and Distributed IR with Eric Broos</b>	<b>399</b>
10.1 Introduction . . . . .	399
10.2 A Taxonomy of Distributed IR Systems . . . . .	402
10.3 Data Partitioning . . . . .	404
10.3.1 Collection Partitioning . . . . .	405
10.3.2 Collection Selection . . . . .	407
10.3.3 Inverted Index Partitioning . . . . .	409
10.3.4 Partitioning other Indexes . . . . .	412
10.4 Parallel IR . . . . .	414
10.4.1 Introduction . . . . .	414
10.4.2 Parallel IR on MIMD Architectures . . . . .	416
10.4.3 Parallel IR on SIMD Architectures . . . . .	418
10.5 Cluster-based IR . . . . .	421
10.6 Distributed IR . . . . .	424

10.6.1	Introduction . . . . .	424
10.6.2	Indexing . . . . .	428
10.6.3	Query Processing . . . . .	431
10.6.4	Web Issues . . . . .	437
10.7	Federated Search . . . . .	438
10.8	Retrieval in Peer-to-Peer Networks . . . . .	440
10.9	Trends and Research Issues . . . . .	444
10.10	Bibliographic Discussion . . . . .	445
<b>11</b>	<b>Web Retrieval</b>	<b>447</b>
	<i>with Yossi Matza</i>	
11.1	Introduction . . . . .	447
11.2	A Challenging Problem . . . . .	449
11.3	The Web . . . . .	451
11.3.1	Characteristics . . . . .	451
11.3.2	Structure of the Web Graph . . . . .	452
11.3.3	Modeling the Web . . . . .	454
11.3.4	Link Analysis . . . . .	456
11.4	Search Engine Architectures . . . . .	458
11.4.1	Basic Architectures . . . . .	458
11.4.2	Cluster-based Architecture . . . . .	459
11.4.3	Caching . . . . .	462
11.4.4	Multiple Indexes . . . . .	464
11.4.5	Distributed Architectures . . . . .	466
11.5	Search Engine Ranking . . . . .	468
11.5.1	Ranking Signals . . . . .	469
11.5.2	Link-based Ranking . . . . .	470
11.5.3	Simple Ranking Functions . . . . .	473
11.5.4	Learning to Rank . . . . .	473
11.5.5	Learning the Ranking Function . . . . .	474
11.5.6	Quality Evaluation . . . . .	475
11.5.7	Web Spam . . . . .	476
11.6	Managing Web Data . . . . .	477
11.6.1	Assigning Identifiers to Documents . . . . .	477
11.6.2	Metadata . . . . .	478
11.6.3	Compressing the Web Graph . . . . .	478
11.6.4	Handling Duplicated Data . . . . .	479
11.7	Search Engine User Interaction . . . . .	480
11.7.1	The Search Rectangle Paradigm . . . . .	481
11.7.2	The Search Engine Result Page . . . . .	488
11.7.3	Educating the User . . . . .	497
11.8	Browsing . . . . .	498
11.8.1	Flat Browsing . . . . .	499
11.8.2	Structure Guided Browsing and Web Directories . . . . .	499
11.9	Beyond Browsing . . . . .	501
11.9.1	Hypertext and the Web . . . . .	501
11.9.2	Combining Searching with Browsing . . . . .	501
11.9.3	Web Query Languages . . . . .	503

13.3.2 Model Based on Proximal Nodes . . . . .	550
13.3.3 Ranking Structured Text Results . . . . .	551
<b>13.4 XML Retrieval . . . . .</b>	<b>551</b>
13.4.1 Challenges in XML Retrieval . . . . .	551
13.4.2 Indexing Strategies . . . . .	553
13.4.3 Ranking Strategies . . . . .	554
13.4.4 Resolving Overlaps . . . . .	555
<b>13.5 XML Retrieval Evaluation . . . . .</b>	<b>556</b>
13.5.1 Document Collections . . . . .	556
13.5.2 Topics . . . . .	557
13.5.3 Retrieval Tasks . . . . .	558
13.5.4 Performance . . . . .	559
13.5.5 Measures . . . . .	571
<b>13.6 Query Languages . . . . .</b>	<b>573</b>
13.6.1 Characteristics . . . . .	574
13.6.2 Classification of XML Query Languages . . . . .	575
13.6.3 Examples of XML Query Languages . . . . .	577
<b>13.7 Trends and Research Issues . . . . .</b>	<b>582</b>
<b>13.8 Bibliographic Discussion . . . . .</b>	<b>585</b>
<b>14 Multimedia Information Retrieval . . . . .</b>	<b>587</b>
<i>by Duke Porteleta and Maureen Slaney</i>	
<b>14.1 Introduction . . . . .</b>	<b>587</b>
14.1.1 What is Multimedia? . . . . .	587
14.1.2 Multimedia IR . . . . .	588
14.1.3 Text IR versus Multimedia IR . . . . .	589
<b>14.2 The Challenges . . . . .</b>	<b>589</b>
14.2.1 The Semantic Gap . . . . .	590
14.2.2 Feature Ambiguity . . . . .	591
14.2.3 Machine-generated Data . . . . .	591
<b>14.3 Content-based Image Retrieval . . . . .</b>	<b>592</b>
14.3.1 Color-Based Retrieval . . . . .	593
14.3.2 Texture . . . . .	593
14.3.3 Saliency Points . . . . .	596
<b>14.4 Audio and Music Retrieval . . . . .</b>	<b>597</b>
14.4.1 Fingerprinting . . . . .	598
14.4.2 Speech Recognition . . . . .	599
14.4.3 Speaker Identification . . . . .	601
14.4.4 Spoken Document Retrieval . . . . .	602
14.4.5 Audio Basics . . . . .	603
<b>14.5 Retrieving and Browsing Video . . . . .</b>	<b>606</b>
14.5.1 Video Abstracts . . . . .	606
14.5.2 Static Summaries . . . . .	607
14.5.3 Mosaics and Saliency Stills . . . . .	608
14.5.4 Dynamic Summaries . . . . .	609
14.5.5 Interactive Summaries . . . . .	611
14.5.6 Visual vs. Audio Browsing . . . . .	612
14.5.7 Evaluating Summaries . . . . .	613

14.6	Fusion Models: Combining It All . . . . .	614
14.6.1	Naming Faces . . . . .	614
14.6.2	Naming Images . . . . .	615
14.6.3	Naming Audio . . . . .	616
14.6.4	Combining Audio and Video for ASR . . . . .	617
14.6.5	Combining Audio and Video for Multimedia . . . . .	620
14.7	Segmentation . . . . .	620
14.7.1	A Video Segmentation Example . . . . .	620
14.7.2	Segmentation Schemes for Video . . . . .	622
14.7.3	Video Segmentation with Edges . . . . .	623
14.7.4	Speech Segmentation . . . . .	624
14.7.5	Segmentation Evaluation . . . . .	625
14.8	Compression and MPEG Standards . . . . .	626
14.8.1	Intensity and Sampling . . . . .	626
14.8.2	Color . . . . .	626
14.8.3	Lossy Compression . . . . .	628
14.8.4	Lossless Compression . . . . .	628
14.8.5	Temporal Redundancy . . . . .	630
14.8.6	Motion Prediction . . . . .	631
14.8.7	MPEG Standards . . . . .	633
14.9	Trends and Research Issues . . . . .	636
14.10	Bibliographic Discussion . . . . .	637
15	<b>Enterprise Search</b> <i>by David Balogun</i> . . . . .	641
15.1	Introduction . . . . .	641
15.1.1	Characteristics and Applications of Enterprise Search . . . . .	642
15.1.2	Enterprise Search Software . . . . .	643
15.1.3	Workplace Search . . . . .	644
15.2	Enterprise Search Tasks . . . . .	644
15.2.1	Examples of Search-Supported Tasks . . . . .	644
15.2.2	Search Types . . . . .	647
15.2.3	Studying Enterprise Search . . . . .	647
15.3	Architecture of Enterprise Search Systems . . . . .	648
15.3.1	Gathering . . . . .	648
15.3.2	Extracting . . . . .	651
15.3.3	Indexing . . . . .	653
15.3.4	Indexing Textual Annotations . . . . .	653
15.3.5	Query Processing . . . . .	654
15.3.6	Presentation of Search Results . . . . .	655
15.3.7	Security Models . . . . .	657
15.3.8	Federation/Metasearch . . . . .	659
15.4	Enterprise Search Evaluation . . . . .	662
15.4.1	Published Test Collections for Enterprise Search . . . . .	662
15.4.2	Internal Enterprise Search Evaluations . . . . .	663
15.4.3	Enterprise Search Tuning . . . . .	665
15.4.4	What is it Reasonable to Expect? . . . . .	666
15.5	Potential Reasons for Dissatisfaction . . . . .	667

15.6 Content and Personalization . . . . .	698
15.6.1 Controls and Levers for Contextualization . . . . .	699
15.6.2 Contextualization: Local, Enterprise or Global? . . . . .	700
15.6.3 Privacy of Profiles . . . . .	700
15.6.4 Defining, Creating and Maintaining a Profile . . . . .	707
15.6.5 User Modeling . . . . .	707
15.6.6 Implicit Measures . . . . .	709
15.6.7 Information Filtering . . . . .	709
15.6.8 Social Recommender Systems . . . . .	710
15.7 Trends and Research Issues . . . . .	711
15.8 Bibliographic Discussion . . . . .	711
<b>16 Library Systems</b> . . . . .	<b>715</b>
<i>by Edie Rasmussen</i>	
16.1 The Information Environment in the Library . . . . .	715
16.2 Online Public Access Catalogues . . . . .	717
16.2.1 OPACs and Bibliographic Records . . . . .	719
16.2.2 Information Retrieval from the ILS . . . . .	721
16.2.3 Integrating the Hybrid Library . . . . .	723
16.2.4 OPACs and End Users . . . . .	724
16.2.5 ILS Vendors and Products . . . . .	725
16.3 IR Systems and Document Databases . . . . .	727
16.3.1 Bibliographic and Full-text Databases . . . . .	728
16.3.2 Content of Database Records . . . . .	729
16.3.3 The Online Industry: Database Vendors . . . . .	731
16.3.4 Information Retrieval from Document Databases . . . . .	732
16.4 Information Retrieved in Organizations . . . . .	736
16.5 Trends and Research Issues . . . . .	738
16.6 Bibliographic Discussion . . . . .	739
<b>17 Digital Libraries</b> . . . . .	<b>741</b>
<i>by Marcos González</i>	
17.1 Introduction . . . . .	741
17.2 Defining Digital Libraries . . . . .	742
17.3 A General Architecture . . . . .	743
17.4 Fundamentals . . . . .	744
17.4.1 Digital Objects and Collections . . . . .	744
17.4.2 Metadata and Catalogs . . . . .	746
17.4.3 Repositories/Archives . . . . .	749
17.4.4 Services . . . . .	753
17.5 Social-Economical Issues . . . . .	755
17.5.1 Social Issues . . . . .	755
17.5.2 Economical Issues . . . . .	756
17.6 Software Systems . . . . .	757
17.6.1 Greenstone . . . . .	758
17.6.2 Eprints . . . . .	758
17.6.3 DSpace . . . . .	758
17.6.4 Fedora . . . . .	759

17.6.5 Open Digital Libraries . . . . .	729
17.6.6 The IS Suite . . . . .	730
17.7 DL Case Studies . . . . .	731
17.7.1 The Networked DL of Theses and Dissertations . . . . .	731
17.7.2 The National Science Digital Library . . . . .	732
17.7.3 The ETANA-DL Archaeological Digital Library . . . . .	732
17.8 Trends and Research Issues . . . . .	733
17.8.1 Evaluation . . . . .	733
17.8.2 Integration . . . . .	733
17.8.3 Other Research Challenges . . . . .	734
17.9 Bibliographic Discussion . . . . .	735
<b>A Open Source Search Engines</b> . . . . .	<b>737</b>
with Christian Middleton	
A.1 Introduction . . . . .	737
A.2 Search Engines . . . . .	738
A.2.1 Preliminary Selection of Search Engines . . . . .	738
A.2.2 Features . . . . .	741
A.2.3 Evaluation . . . . .	742
A.3 Methodology . . . . .	743
A.3.1 Document Collections . . . . .	743
A.3.2 Evaluation Tests . . . . .	744
A.3.3 Experimental Setup . . . . .	744
A.4 Experimental Results . . . . .	745
A.4.1 Test A – Indexing . . . . .	745
A.4.2 Test B – Incremental Indexing . . . . .	749
A.4.3 Test C – Search Performance . . . . .	749
A.4.4 Global Evaluation . . . . .	752
A.5 Conclusions . . . . .	753
<b>B Biographies</b> . . . . .	<b>755</b>
<b>References</b> . . . . .	<b>761</b>
<b>Index</b> . . . . .	<b>893</b>

# Preface to the Second Edition

Since the first edition of this book, much has happened in the Information Retrieval (IR) arena, particularly with regard to the Web. For one, the gargantuan volume of information on the Web has transformed the search engines into key tools for seeking and finding information of interest. Further, since the search engines are fundamentally IR systems at their very core, they have become the main proof-of-concept of the application of IR technologies to vast document collections with huge query traffic.

We closely followed this evolution by starting search engines in Brazil and Chile, just a few months after the first edition appeared. Later on, we joined the two major search engine companies, Google and Yahoo!, getting even closer to all the action. Hence, this second edition of *Modern Information Retrieval* reflects not only the changes in the IR field, but also our own experience in research, development, and execution of IR technology, particularly when applied to the Web.

The first edition of *Modern Information Retrieval* was not a book written in standard fashion, given that we asked contributors to write chapters in areas in which we felt we did not have enough expertise. So, in some sense, we promoted the Web 2.0 trend of development in collaboration with a team. We aimed at a well integrated book by carefully coordinating and supervising all the writing. To a certain extent, our efforts worked well. Indeed, the first edition of the book sold very well and became the IR best seller, having been reprinted many times. The book has been adopted by hundreds of universities and schools. It has also been translated first to Korean and then to Chinese, with an special non-expensive edition having been printed in India. Hence, just a couple of years after the first edition was printed, we started talking about a second edition. The idea did not materialize until 2004 when we submitted a proposal to the publisher, which was approved. We eventually started working on the second edition by November 2006, more than five years ago. Today, we have finally finished!

In this second edition of *Modern Information Retrieval* we have followed the same methodology, as it clearly worked with the first edition. Nonetheless, in this case we are authors or co-authors of more chapters and we have taken a stronger hand in shaping the content of the contributed chapters. As a consequence, we have had to change completely many chapters and add many new ones. As a result, 60-70% of this second edition is made of new material, which mostly differs from the first edition

in the following aspects:

1. A complete reorganization of the content of the first chapters.
2. New chapters on text classification, Web crawling, structured text retrieval, and enterprise search plus a new appendix on open source search engines.
3. Fully rewritten chapters on user interfaces, multimedia retrieval, and digital libraries.
4. Expanded chapters to include new and important developments such as language models, new evaluation measures, characteristics of queries, cluster-based and distributed IR, learning to rank, search engines interfaces, click-through data, and personalization.
5. An improved Web site, see [www.mk2ed.org](http://www.mk2ed.org), aimed at becoming a reference IR teaching resource, which includes a full set of slides for all chapters in the book and recommended lists of exercises.

The final outcome is a book that is almost twice as long and that contains more than twice the number of references of the first edition. In summary, if you liked the first edition of *Modern Information Retrieval*, we hope you will like this second edition even more. And, in case you did not like the first edition, we hope that this time you will change your mind.

Ricardo Baeza-Yates, Barcelona, Spain  
Bernier Ribeiro-Neto, Belo Horizonte, Brazil  
December, 2010

# Preface to the First Edition

## the Second Edition

---

Information retrieval (IR) has changed considerably in the last years with the expansion of the Web (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 which has led to the introduction of new IR books recently. Nevertheless, we believe that there is still great need of a book that approaches the field in a rigorous and complete way from a computer-science perspective (in opposition 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 is composed of two portions which complement and balance each other. The core portion includes 9 chapters authored or coauthored by the designers of the book. The second portion, which is fully integrated with the first, is formed by 6 state-of-the-art chapters written by leading researchers in their fields. A same notation and glossary are employed in all the chapters. Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors 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 broadness of coverage and richness of details. 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  
October, 1998

## Authors' Acknowledgements to the Second Edition

---

We would like to sincerely thank the various people who, over a period of several years, provided us with useful and helpful comments, reviews, and suggestions. The improvements in the book content and in the organization of the material are largely due to them. Without their help, this second edition would not be of the same quality. Any errors that remain, hopefully only few, are entirely our responsibility.

First, we would like to thank all the chapter contributors, for their dedication and interest. To Eric Brown, Carlos Castillo, Marcos Gonçalves, David Hawking, Marti Hearst, Mounia Lalmas, Yacine Maarek, Christian Middleton, Gonzalo Navarro, Dalek Pensecic, Edie Rasmussen, Malcolm Slaney, and Nirio Ziviani, whose contributions reflect expertise we certainly have not fully mastered ourselves.

Second, to all the people who directly or indirectly contributed or influenced the new content of this second edition. We have to thank Omar Alonso (who pointed out that we were laying out the important trend of crowdsourcing), Paolo Boldi (Web graph compression), Pavel Calado (text classification), Marco Cristo (whose comments on the text classification chapter led to an overall organization of the material), Christos Faloutsos (multi-dimensional indexing), Winston Heu (multimedia), Flávio Junqueira (distributed retrieval), Edlene Moura (retrieval evaluation), Vanessa Murkoff (query difficulty), Martin Porter (his stemming algorithm), Mark Sanderson (whose sharp comments led to great improvements on the retrieval evaluation chapter), Fabrizio Silvestri (URL ordering), and Gleb Skobchyan (P2P IR). Further, we also acknowledge the contributions of various graduate students of Marcos Gonçalves at the Federal University of Minas Gerais, Brazil, who reviewed and wrote extensive comments on the text classification chapter.

Third, we need to thank all the people who sent us errata for the first edition, comments for improvements and also comments on drafts of the second edition. In the case of errata we mention the first people who detected a mistake, as otherwise the list would be too long. They are, with the risk of omitting someone, Omar Alonso, José Hilario Caño, Bertrand Barla Cambazoglu, Ernie Davis, Anne Dickson, Bill Diurn, Joaquim Gaharro, Jamie Goldes, Eduardo Gracilis, Kyoung-Soo Han, Claudia Henaff, Shouqie He, Ben Houston, Paay-Leng Lee, Songwook Lee, Shian-Hua Lin,

Mildrid Ljostad, Chang-Tien Lu, Mari Carmen Marcos, Peter Mika, Vanessa Murdoch, Joanne Plattner, Luis Rello, Hie-Chael Seo, Ben Shneiderman, Helge Grenager Solheim, Ellen Sparreus, Markus Stocker, Kazunori Sugiyama, Satoru Takabayashi, Jutta Tikkinen, Luong Minh Thang, Yannis Tsitlikas, Fredrik Wallenborg, Theo van der Weide, John Westbrook, Judith Winter, Sai Xi, Peng Yong, Hugo Zaragoza, and Youhai Zhang.

Fourth, special thanks to David Fernandes who made the teaching slides that can be found on the book Web site and patiently pointed out many small errors and inconsistencies throughout the book. Also, we need to mention the implicit support of our employers, Yahoo! and Google, in the above difficult task of writing a book.

Fifth, we have to thank our editors at Pearson Education. We started with Kate Brewin, then Simon Plumtree, next Owen Knight, and finally Rufus Curnow, who supported us during the most important part of the publishing process. During this process we had the help of Anita Atkinson as our desk editor and Jenny Oates as proof reader.

Finally, and most important, to Helena, Rose, and our children, who care more about us than we do ourselves, and who have been the best慰藉 to us during these last four years, their recurrent question was: when will you finish the boat?

concerned. We sincerely hope that our acknowledgements will help to make our thanks more meaningful to those who have contributed to this book.

## Authors' Acknowledgements to the First Edition

---

We would like to sincerely thank the various people who, throughout 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 materialized.

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 Ferrai, Ed Fox, Marti Hearst, Gonzalo Navarro, Eddie Rasmussen, Ohm Senni, and Nivio Ziviani, whose contributions reflect expertise we certainly have not fully mastered 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 expressed interest in publishing this book, in particular, 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 Harrison, 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 nice 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 arguments of an unknown referee. A separate chapter on Retrieval Evaluation was only included after another unknown 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 effort since 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 expressed 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 Fapes project SIAM (Information Systems for Mobile Computers).

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

Downloaded from <http://www.informaworld.com> at 09:45 09 September 2009

## Publisher's Acknowledgements

We are grateful to the following for permission to reproduce copyright material:

### Figures

Figures 2.1 and 2.12 from Yelp!, <http://www.yelp.co.uk/>; Yelp! Inc.; Figure 2.3 from NextBite.com; Figures 2.5, 4.13b, 11.10c, 11.11a and 11.13 from Google screenshots, [www.google.co.uk](http://www.google.co.uk); Figure 2.6 from <http://research.berkeley.edu>, copyright M. A. Hearst; Figure 2.7 from Microsoft product screenshots reprinted with permission from Microsoft Corporation; Figure 2.13 from FindEx, copyright (c) 2010 FindEx.com, Inc., and its licensors; Figure 2.18 from Graphical query specification and dynamic result previews for a digital library, *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST'98)* pp. 143-51 (Jones, S. 1998), <http://doi.acm.org/10.1145/288392.288595>, © 1998 Association for Computing Machinery, Inc. Reprinted by permission; Figure 2.14 from Research TileBars, <http://people.csail.mit.edu/hearst/research/tilebars.html>, copyright M. A. Hearst; Figure 2.17a from *Search User Interfaces*, Cambridge University Press (Hearst, M.A. 2009) figure 10.1(a), copyright M. A. Hearst; Figure 2.17b from INSYDEB: a content-based visual-information-seeking system for the web, *International Journal on Digital Libraries*, pp. 25-41 (Reiterer, H., Tallus, G. and Mann, T.M. 2005), with kind permission from Springer Science + Business Media and OCLC; and Professor H. Reiterer; Figure 2.18 from Using thumbnails to search the Web, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pp. 199-206 (Woodruff, A., Fushing, A., Rosenblatt, R., Morrison, J. and Piroli, P. 2001), <http://doi.acm.org/10.1145/365024.365098>, © 2001 Association for Computing Machinery, Inc. Reprinted by permission; Figure 2.20b from Evaluating a system for interactive exploration of large, hierarchically structured document repositories, *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pp. 127-33 (Granitzer, M., Kimareich, W., Sabol, V., Andrews, K. and Kibler, W. 2004), © 2004 IEEE; Figure 2.20b from Search result visualisation with xFIND, *Proceedings of User Interfaces to Data Intensive Systems (UIDAS 2001)*, pp. 59-8 (Andrews, K., Gotl, C., Moser, J., Sabol, V. and Lockner, W. 2001), © 2001 IEEE; Figure 2.21 from <http://kylescholtz.com/projects/wordnet/>; Kyle Scholtz;

Figure 2.22 from The Word tree, an interactive visual concordance, *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pp. 1221–8 (Wittenberg, M. and Fernanda, B. 2008), © 2008 IEEE; Figure 2.23 from Baby names popularity graph NameVoyager, <http://www.babynamewizard.com>; Figure 2.24 from Avian flu case study with aSpace and GeoTime, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST'06)* pp. 27–34 (Proulx, P. et al. 2006), © 2006 IEEE; Figure 5.4 after Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search, *ACM Transactions on Information Systems*, 25(2) (Jozefiniak, T., Granka, L., Pan, B., Hembroski, H., Radlinski, F. and Gay, G. 2007), <http://doi.acm.org/10.1145/1229179.1229141>, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figures 7.4 and 7.5 from The impact of caching on search engines, *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)* (Baeza-Yates, R. et al. 2007), <http://doi.acm.org/10.1145/1277741.1277775>, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figure 7.6 from Query usage mining in search engines, *Web Mining Applications and Techniques* (Baeza-Yates, R., Scime, A. ed.) (2004), Idea Group, reprinted by permission of the publisher, IGI Global; Figure 10.1 adapted from Load balancing for term-distributed parallel retrieval, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 348–355 (Moffat, A., Webber, W. and Zobel, J. 2006), <http://doi.acm.org/10.1145/1148170.1148332>, © 2006 Association for Computing Machinery, Inc. Reprinted by permission; Figures 10.12 and 10.13 from Challenges on distributed web retrieval, *Proceedings of ICDB 2007*, pp. 6–20 (2007), © 2007 IEEE; Figure 10.14 from A pipelined architecture for distributed text query evaluation, *Information Retrieval*, 10(3), pp. 205–31 (Webber, W., Moffat, A., Zobel, J. and Baeza-Yates, R. 2007), with kind permission from Springer Science + Business Media; Figure 11.1 from Graph structure in the web: experiments and models, *Proceedings of the North Conference on World Wide Web*, pp. 209–20 (Inder, A., Kumar, R., Magdon, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wieser, J. 2000), copyright Elsevier (2000); Figure 11.3a from M. Crovella, 1998; Figure 11.3b from Self-similarity in World Wide Web traffic: evidence and possible causes, *SIGMETRICS'98: Proceedings of the 1998 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 24, pp. 160–9 (Crovella, M.E. and Bestavros, A. 1998), <http://doi.acm.org/10.1109/90.656143>, © 1998 Association for Computing Machinery, Inc. Reprinted by permission; Figures 11.4 and 11.5 from Generic damping functions for propagating importance in link-based ranking algorithms, *Internet Mathematics*, 3(4), pp. 443–78 (Baeza-Yates, R., Boldi, P. and Castillo, C. 2006), copyright 2006 A. K. Peters, Ltd.; Figure 11.7 after Challenges in building large-scale information retrieval systems: invited talk presentation, <http://research.google.com/people/jeff/WSDM09-keynote.pdf>, Jeffrey Dean; Figure 11.8 from Design trade-offs for search engine caching, *TWEB*, 2(4) (Baeza-Yates, R.A., Gionis, A., Junqueira, F., Murdoch, V., Plachouras, V. and Shastri, F. 2008), <http://doi.acm.org/10.1145/1409220.1409223>, © 2008 Association for Computing Machinery, Inc. Reprinted by permission; Figure 11.10a, Ask screenshot, © IAC Search & Media, Inc., 2010. All rights reserved. ASK.COM, ASK JEEVES, the ASK logo, the ASK JEEVES logo and other trade marks appearing on the Ask.com and Ask Jeeves websites are property of IAC Search & Media, Inc. and/or its licensors; Figures 11.10b and 11.15 from Bing screenshots, reprinted with permission from Microsoft.

Corporation; Figure 12.8 from Synchronizing a database to improve freshness, *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, pp. 117–28 (Cho, J. and Garcia-Molina, H. 2000), <http://doi.acm.org/10.1145/342009.335391>, © 2000 Association for Computing Machinery, Inc. Reprinted by permission; Figure 13.9 from INEX 2006 assessment interface, Professor Mónica Llorente; Figure 14.4 from IBM Almaden Research Center; Figures 14.6 and 14.8 are courtesy of Jim Halter from the QBIC system, IBM Almaden Research Center; Figure 14.9 from A bipartite graph model for associating images and text, *IJCACI-2007 Workshop on Multimedia Information Retrieval* (Srinivasan, S.H. and Stasney, M. 2007); Figure 14.10 from Image retrieval on large-scale image databases, *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR 07)*, pp. 17–24 (Horster, E., Liebhart, R. and Stasney, M. 2007), <http://doi.acm.org/10.1145/1282260.1282283>, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figures 14.13 and 14.14 from Kyoko Lee; Figure 14.16 from Video skimming for quick browsing based on audio and image characterization, *Technical Report CMU-CS-95-188* (Smith, M.A. and Kanade, T. 1995) School of Computer Science Tech Report, Carnegie Mellon University; Figure 14.17 from Video summary generating semantically meaningful video summaries, *MULTIMEDIA '99: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pp. 283–92 (Uchihashi, S. et al. 1999), <http://doi.acm.org/10.1145/319463.319484>, © 1999 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.18 from Harpreet Sawhney, Samoff Corporation; Figure 14.19 from Salient stills, *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(1), pp. 16–36 (Tessios, L. and Bender, W. 2005), <http://doi.acm.org/10.1145/1047936.1047940>, © 2005 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.20 from PanoramaExample: Extracting and parking panoramas for video browsing, *MULTIMEDIA '97: Proceedings of the Fifth ACM International Conference on Multimedia*, pp. 427–36 (Taniguchi, Y., Akatsu, A. and Tomonaga, Y. 1997), <http://doi.acm.org/10.1145/266100.266398>, © 1997 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.21 from Hierarchical brushing in a collection of video data, *Proceedings of Hawaii International Conference on System Sciences (HICSS) (2001)*, © 2001, IEEE; Figure 14.26 from Automatic recognition of audivisual specific recent progress and challenges, *Proceedings of the IEEE (Portmann, G., Neti, C., Gravier, G., Garg, A. and Senior, A.W. 2003)*, © 2003 IEEE; Figure 14.28 from Multimedia edges: Finding hierarchy in all dimensions, *Proceedings of 8th ACM International Conference on Multimedia (Stasney, M., Poncelet, D. and Kauffman, J. 2001)*, <http://doi.acm.org/10.1145/500141.500148>, © 2001 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.29 from Comparison of automatic shot boundary detection algorithms, *SPIE Image and Video Processing VII*, pp. 326–729 (Liebhart, R. 1999), SPIE; Figure 15.3 from Oxford Australia; Figure 15.5 from Evaluation by comparing result sets in context, *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* pg. 94–101 (Thomas, P. and Hawking, D. 2006), <http://doi.acm.org/10.1145/1193614.1193632>, © ACM, 2006; Figure 16.1 from Edie Rasmussen, and with permission from The Network Development and MARC Standards Office; Figure 16.2 from Find... books or journals, <http://www.library.ubc.ca/hora/research.html>; The University of British Columbia Library website (2010). Used with permission; Figures 16.4, 16.5, 16.6 and 16.7 from DIALOG, Dialog® interface and screen shots reproduced with permission of Dialog.

LLC. The Dialog product name is a registered trademark of Dialog, LLC; Figure 16.4 with permission from EBSCO Publishing, Inc.

## Tables

Table 4.2 after Overview of the sixth text retrieval conference (TREC- 6), *Proceedings of the Sixth Text Retrieval Conference (TREC-6)* (Voorhees, E. and Harman, D. 1997); Table 7.1 from From e-commerce to e-commerce: Web search changes, *Computer*, 35(3), pp. 107-9 (Spink, A., Jansen, B.J., Wolfson, D. and Saracevic, T. 2002), © 2002 IEEE.

## Text

Extract on page 339 from <http://trec.nist.gov>, NIST.

In some instances we have been unable to trace the owners of copyright material, and we would appreciate any information that would enable us to do so.