

Introduction to Modern Information Retrieval

Second edition

G. G. Chowdhury



facet publishing

© G. G. Chowdhury 1999, 2004

Published by
Facet Publishing
7 Ridgmount Street
London WC1E 7AE

Facet Publishing (formerly Library Association Publishing) is wholly owned by CILIP: the Chartered Institute of Library and Information Professionals.

G. G. Chowdhury has asserted his right under the Copyright, Designs and Patents Act 1988 to be identified as the author of this work.

Except as otherwise permitted under the Copyright, Designs and Patents Act 1988 this publication may only be reproduced, stored or transmitted in any form or by any means, with the prior permission of the publisher, or, in the case of reprographic reproduction, in accordance with the terms of a licence issued by The Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to Facet Publishing, 7 Ridgmount Street, London WC1E 7AE.

First published by Library Association Publishing 1999
Reprinted 2001
Reprinted by Facet Publishing 2002
This second edition 2004

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 1-85604-480-7

Typeset in 10/12pt Times and Arial by Facet Publishing.
Printed and made in Great Britain by MPG Books Ltd, Bodmin, Cornwall.

To Sudatta, Avirup and Anubhav

Contents

Preface	xi
Coverage	xi
Acknowledgements	xiii
1 Basic concepts of information retrieval systems	1
Introduction	1
Purpose	2
Functions	3
Components	3
Kinds of information retrieval systems	4
Design issues	5
Design phases	8
References	11
2 Database technology	13
Introduction	13
Data	13
The database	13
Records and fields	14
Properties of databases	14
Kinds of databases	15
Database technology	16
The development of databases in an information retrieval environment	18
Discussion	22
References	23
3 Bibliographic formats	24
Introduction	24
Bibliographic records	25
Integrated database approach	27
ISO 2709: Format for Bibliographic Information Interchange	28
MARC format	31
UNIMARC format	36
The Common Communication Format	38
Discussion	40
References	40

4	Cataloguing and metadata	42
	Introduction	42
	Cataloguing	42
	Metadata	50
	Summary	54
	References	55
5	Subject analysis and representation	57
	Introduction	57
	Classification	58
	Bibliographic Classification	65
	Classification of internet resources	66
	Subject analysis	70
	Subject indexing	72
	Exhaustivity and specificity	72
	Manual indexing	74
	Pre-coordinate indexing systems	76
	Post-coordinate indexing systems	84
	Problems of manual indexing	86
	Theory of indexing	86
	Discussion	87
	References	87
6	Automatic indexing and file organization	91
	Introduction	91
	The process of indexing	91
	Automatic classification	94
	Index file organization	96
	Inverted file	97
	Sequential access	102
	Alternative text retrieval structures	110
	Discussion	119
	References	120
7	Vocabulary control	123
	Introduction	123
	Controlled vs natural indexing	124
	Vocabulary control tools	126
	Guidelines for developing a thesaurus	142
	Criteria for evaluating a thesaurus	143
	Use of thesauri in online information retrieval	143
	References	148

8	Abstracts and abstracting	153
	Abstracts	153
	Types of abstract	153
	Qualities of abstracts	155
	Uses of abstracts	156
	The art of abstracting	157
	Automatic abstracting	160
	Recent works on text summarization	165
	Discussion	166
	References	166
9	Searching and retrieval	169
	Introduction	169
	The search strategy and its prerequisites	169
	The pre-search interview	170
	The searching process	170
	Retrieval models	171
	Alternative information retrieval models	182
	Search facilities offered by most text retrieval systems	183
	Discussion	189
	References	190
10	Users of information retrieval	192
	Introduction	192
	Users and their nature	192
	Types of information needs	193
	Information needs in different areas of activity	195
	Information seeking behaviour of users	200
	What we need to know about users	201
	User studies	205
	Possible sources of information about users	210
	References	212
11	User-centred models of information retrieval	214
	Introduction	214
	Information seeking	214
	Human information behaviour models	216
	User-centred information search models	219
	Discussion	223
	References	224
12	User interfaces	227
	Introduction	227
	The four-phase framework for interface design	227

Information seeking and user interfaces	230
User interfaces and visualization	231
User interfaces of some information retrieval systems	232
References	240
13 Evaluation of information retrieval systems	243
Introduction	243
The purpose of evaluation	244
Evaluation criteria	244
The steps of evaluation	252
New retrieval parameters	253
References	253
14 Evaluation experiments	255
Introduction	255
The Cranfield tests	256
MEDLARS	261
The SMART retrieval experiment	262
The STAIRS project	266
Limitations of early evaluation studies	267
TREC	269
References	278
15 Online and CD-ROM information retrieval	280
Introduction	280
Online searching	280
CD-ROM databases	292
Summary	296
References	298
16 Multimedia information retrieval	299
Introduction	299
Multimedia information retrieval	299
Standards	311
Summary	312
References	312
17 Hypertext and markup languages	315
Introduction	315
Hypertext	316
Markup languages	323
Discussion	327
References	328

18	Web information retrieval	330
	Introduction	330
	Traditional vs web information retrieval	330
	Web information: volume and growth	332
	Access to information on the web: the tools	335
	Web information retrieval: evaluation studies	347
	References	349
19	Intelligent information retrieval	352
	Introduction	352
	Intelligent retrieval systems	353
	Artificial intelligence	353
	Expert systems	354
	Kinds of expert systems	355
	Components of expert systems	355
	Historical development of expert systems	357
	Development methodology and approaches	357
	Knowledge elicitation and representation methods	359
	Inference strategies	360
	End-user modelling and interfaces	360
	Development tools	362
	Expert systems for library and information services	362
	Discussion	367
	References	368
20	Natural language processing and information retrieval	372
	Introduction	372
	Natural language understanding	372
	Syntactic analysis	373
	Semantic analysis	381
	Pragmatic knowledge	390
	References	394
21	Natural language processing systems	396
	Introduction	396
	Literature on natural language processing systems	397
	Natural language text processing systems	398
	Natural language user interfaces	408
	Internet, web and digital library applications of natural language processing systems	412
	Machine translation and cross-language information retrieval	413

Summary	415
References	416
22 Information retrieval in digital libraries	425
Introduction	425
Information resources in digital libraries	426
The basic design of a digital library	426
Interoperability	428
Information retrieval features of selected digital libraries	428
Common features of information retrieval in digital libraries	436
Special IR features in DLs	437
Problems and prospects	437
Summary	441
References	442
23 Trends in information retrieval	445
Introduction	445
Evaluation of information retrieval systems	447
Developments related to the input subsystem	448
Searching and retrieval	450
User studies and user modelling	452
User interfaces	454
Information retrieval standards and protocols	454
Information retrieval in the context of web and digital libraries	455
Intelligent information retrieval	457
Evaluation of natural language processing systems	459
Machine translation	459
Conclusions	461
References	462
Index	467

Preface

The rapid growth and development of the internet, the world wide web and digital libraries since the first edition of this book have brought about many significant changes in the world of information retrieval. While the original aim of this book to provide a blend of traditional and new approaches to information retrieval remains the same, this second edition has been revised to incorporate some of the wider perspectives – and the latest developments – of information retrieval.

The book aims to cover the whole spectrum of information storage and retrieval in a way that is relevant to an international readership. The primary audience I have in mind comprises students of library and information science programmes, both at undergraduate and at postgraduate levels. Written from a relatively non-technical perspective, this book is expected to meet the requirements of students undertaking courses in information retrieval, information organization, information use, digital libraries, etc. It will also help practising library and information professionals to brush up their knowledge in different areas of information retrieval.

Coverage

While the content and coverage of Chapters 1 and 2 have remained the same, they have been updated to incorporate wider perspectives of information retrieval ranging from the shelf to the web. Chapter 3 has been revised considerably by dropping discussions on the less well-known bibliographic formats in favour of new sections on the MARC21 format.

In the first edition Chapter 4 covered the areas of cataloguing and classification, and consequently the discussions on both these areas were rather brief. In this edition a chapter has been devoted to each of them. Beginning with the basics of cataloguing and AACR2, Chapter 4 discusses the implications of using AACR2 in automated cataloguing. There follow sections on the cataloguing of internet resources, and on metadata and the various metadata standards. Chapter 5 discusses the basic concepts of classification and subject indexing, which are considered to be the traditional approaches to information organization and retrieval. However, this chapter also describes the various approaches to the use of these traditional tools for organizing web information resources.

Chapter 6 discusses various automatic indexing and file organization techniques. Chapter 7 discusses the issues of vocabulary control in information retrieval. This chapter has been revised by adding new sections on the use of vocabulary control tools in online information retrieval. While the basic concepts of abstracts and abstracting remain the same, Chapter 8 updates the sections on automatic abstract-

ing to include some recent studies in this area. Chapter 9 discusses the information search process and the various information retrieval models. This chapter now also discusses various alternative information retrieval models, and provides some illustrations for the process of query expansion as part of an online information search process.

The first edition had only one chapter on information users. In this edition, Chapter 10 covers the basic issues of information users and the various approaches to user studies, while Chapter 11 discusses the various user-centred information retrieval models. Chapter 12 covers the topic of user interfaces, an essential component of an information retrieval system.

Chapters 13 and 14 cover discussions on the evaluation of information retrieval. These chapters have been revised significantly to include discussion of the TREC (Text Retrieval Conferences) series of evaluation experiments. Discussions on online and CD-ROM information retrieval have been merged to form only one chapter, Chapter 15. While some theoretical discussions on online and CD-ROM retrieval technology have been dropped, more examples have been provided to illustrate the information retrieval process.

Chapter 16 covers discussions on multimedia information retrieval, which appeared in Chapter 15 in the first edition. This chapter has been significantly revised to include discussions on image, audio and video information retrieval. Chapter 17 covers discussions on hypertext information retrieval. New sections on hypertext markup languages, including SGML, HTML, XML and XHTML, have been added.

Chapter 22 of the first edition has been completely revised to form the new Chapter 18, which covers various aspects of web information retrieval. A number of screenshots have been added to illustrate the features of various web information retrieval tools, and recent studies analysing various web information retrieval studies. Chapter 17 of the first edition has been updated to form the new Chapter 19, covering discussions on intelligent information retrieval, with new sections describing some recent studies in this area.

Four chapters (18 to 21) in the first edition have been merged and updated to form the two new Chapters 20 and 21 on natural language processing. Chapter 20 discusses the theoretical issues of natural language processing, while Chapter 21 discusses various studies on the applications of NLP techniques to the different areas of information retrieval. Chapter 22 is a new chapter in this edition that covers various aspects of information retrieval in digital libraries. Information retrieval features of selected digital libraries have been discussed here, along with discussions on some digital library research in the areas of information access and retrieval. The book ends with Chapter 23, which analyses the trends in information retrieval.

Acknowledgements

In order to illustrate the content of the various chapters in this book, I have included a number of screenshots of various information sources and services, acknowledged here:

Figures 5.1, 5.2, 5.3 and 12.3: Reproduced with permission of CDLR, Centre for Digital Library Research, <http://cdlr.strath.ac.uk/>.

Figures 7.1 and 7.2: Reproduced with permission of the Library of Congress.

Figure 7.4: Reproduced with permission of OECD Publishing.

Figure 12.1: Used with permission of California Digital Library. The content, both textual and graphical, of systems provided by the California Digital Library is copyrighted by the Regents of the University of California, unless otherwise noted.

Figures 12.2, 12.7, 12.8 and 12.9: Reproduced with permission of ProQuest Information and Learning Company. Further reproduction is prohibited without permission.

Figures 12.4, 12.5 and 12.6: NCBI Entrez and Taxonomy are freely accessible resources from the National Center for Biotechnology Information, US National Library of Medicine.

Figure 18.1: Reproduced with permission of Google, <http://www.google.com>.

Figure 18.3: Reproduced with permission of Lycos Network; HotBot® is a registered service mark of Wired Ventures.

Figure 18.4: Reproduced with permission of Kartoo, <http://www.kartoo.com>.

Figure 18.5: Reproduced with permission of Vivisimo Inc, <http://www.vivisimo.com>.

Many individuals and institutions have directly or indirectly helped me in preparing this edition. Those whose resources have been used as illustrations and discussions in this book are acknowledged with thanks. I am also indebted to various information retrieval specialists who have given their comments, in the form of both formal and informal review of the first edition.

I am especially indebted to my wife, Sudatta, who has provided constant direct and indirect support, while I have been working on this edition. I am also deeply indebted to my two wonderful sons, Avirup and Anubhav, who have been my source of inspiration throughout this period. Finally, I must express my gratitude to the staff of Facet Publishing, without whose constant help and support this book would not have seen the light of day.

G. G. Chowdhury

