Introduction to Modern Information Retrieval

Second edition

G. G. Chowdhury

facet publishing

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.

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

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