

INFORMATION **Representation** and **Retrieval** in the Digital Age

Second Edition

Heting Chu

asis&t
ASIST Monograph Series

Published for the
American Society for Information Science and Technology by

 **Information Today, Inc.**
Medford, New Jersey

First printing, 2010

Information Representation and Retrieval in the Digital Age, Second Edition

Copyright © 2010 by American Society for Information Science and Technology

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without permission in writing from the publisher, except by a reviewer, who may quote brief passages in a review. Published by Information Today, Inc., 143 Old Marlton Pike, Medford, New Jersey 08055.

Publisher's Note: The author and publisher have taken care in preparation of this book but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book and Information Today, Inc. was aware of a trademark claim, the designations have been printed with initial capital letters.

Library of Congress Cataloging-in-Publication Data

Chu, Heting, 1957-

President and CEO: Thomas H. Hogan, Sr.
Editor-in-Chief and Publisher: John B. Bryans
Managing Editor: Amy R. Reeve
ASIST Monograph Series Editor: Samantha Hastings
VP Graphics and Production: M. Heide Dengler
Book Designer: Kara Mia Jalkowski
Cover Designer: Dana J. Stevenson
Copy Editor: Bonnie Freeman
Proofreader: Penelope Mathiesen
Indexer:

Contents

Figures and Tables	ix
Preface to the Second Edition	xi
Preface to the First Edition	xiii

CHAPTER 1

Information Representation and Retrieval: An Overview	1
1.1 History and Development of Information Representation and Retrieval	1
1.1.1 Major Stages	1
1.1.2 Pioneers of the Field	4
1.2 Elaboration on Key Concepts	13
1.2.1 Information	13
1.2.2 Information Representation	14
1.2.3 Information Retrieval	15
1.2.4 Digital Age	15
1.3 Major Components	16
1.3.1 The Database	16
1.3.2 The Search Mechanism	17
1.3.3 The Language	17
1.3.4 The Interface	18
1.4 The Essential Problem in Information Representation and Retrieval	18
1.4.1 The Process of Information Representation and Retrieval . . .	19
1.4.2 The Limits of Information Representation and Retrieval . . .	20
References	21

CHAPTER 2

Information Representation I: Basic Approaches	27
2.1 Indexing	27
2.1.1 Types of Indexing	28
2.1.2 Automated and Automatic Indexing	28
2.1.3 Indexing in the Hyperstructure Environment	29
2.1.4 Social Tagging	30
2.2 Categorization	31
2.2.1 Types of Categorization	31
2.2.2 Principles of Categorization	31
2.2.3 The Convergence of the Two Categorization Approaches . . .	32

iv Information Representation and Retrieval in the Digital Age

2.3 Summarization	32
2.3.1 Types of Summarization	33
2.3.2 The Issue of Representativeness	35
2.4 Other Methods of Information Representation	35
2.4.1 Citations	35
2.4.2 Strings	36
2.5 A Review of Basic Approaches to Information Representation	37
References	38

CHAPTER 3

Information Representation II: Related Topics 41

3.1 Metadata	41
3.1.1 What Is Metadata?	41
3.1.2 Characteristics of Digital Information on the Internet	42
3.1.3 Examples of Metadata Standards	42
3.1.4 Some Questions and Concerns About Metadata	45
3.2 Full Text	46
3.2.1 Representation of Full-Text Information	46
3.2.2 Difficulties in Representing Full Text	47
3.3 Representation of Multimedia Information	48
3.3.1 Types of Multimedia Information	48
3.3.2 Two Major Representation Approaches	48
3.3.3 Challenges in Representing Multimedia	50
3.4 Further Elaboration on Information Representation	51
References	51

CHAPTER 4

Language in Information Representation and Retrieval 53

4.1 Natural Language	53
4.2 Controlled Vocabulary	54
4.2.1 Thesauri	54
4.2.2 Subject Heading Lists	55
4.2.3 Classification Schemes	56
4.2.4 A Comparison of Thesauri, Subject Heading Lists, and Classification Schemes	56
4.3 Natural Language Versus Controlled Vocabulary	57
4.3.1 Different Eras of Information Representation and Retrieval Languages	58
4.3.2 Why Natural Language or Why Controlled Vocabulary?	59
4.4. Language for Information Representation and Retrieval in the Digital Age	61
4.4.1 Taxonomies	63
4.4.2 Folksonomies	64

4.4.3 Ontologies 64
 References 66

CHAPTER 5

Retrieval Techniques and Query Representation 69

5.1 Retrieval Techniques 69
 5.1.1 Basic Retrieval Techniques 69
 5.1.2 Advanced Retrieval Techniques 74
 5.2 Selection of Retrieval Techniques 80
 5.2.1 Functions of Retrieval Techniques 80
 5.2.2 Retrieval Performance 81
 5.3 Query Representation 84
 5.3.1 General Steps 84
 5.3.2 Difficulties with Query Representation 89
 5.3.3 The Automatic Approach 90
 References 91

CHAPTER 6

Retrieval Approaches 93

6.1 Retrieval by Searching 93
 6.1.1 Characteristics of Searching 94
 6.1.2 Types of Searching 94
 6.1.3 Search Strategies 96
 6.2 Retrieval by Browsing 99
 6.2.1 What Is Browsing? 99
 6.2.2 Types of Browsing 100
 6.2.3 Browsing Strategies 102
 6.3 Searching and Browsing Integrated in Retrieval 103
 6.3.1 Comparison of the Two Retrieval Approaches 103
 6.3.2 The Integrated Approach 105
 References 106

CHAPTER 7

Information Retrieval Models 109

7.1 Foundation of All Information Retrieval Models: Matching 109
 7.1.1 Term Matching 109
 7.1.2 Similarity Measurement Matching 110
 7.2 The Boolean Logic Model 111
 7.2.1 Strengths of the Boolean Logic Model 112
 7.2.2 Limitations of the Boolean Logic Model 112
 7.3 Vector Space Model 115
 7.3.1 Strengths of the Vector Space Model 116
 7.3.2 Limitations of the Vector Space Model 117
 7.4 Probability Model 118

vi Information Representation and Retrieval in the Digital Age

7.4.1 Strengths of the Probability Model	119
7.4.2 Limitations of the Probability Model	120
7.5 Extensions of Major Information Retrieval Models	121
7.5.1 Extended Boolean Logic Model	121
7.5.2 Fuzzy Set Model	122
7.6 Information Retrieval Models: A Further Look	123
7.6.1 A Review of the Major Information Retrieval Models	124
7.6.2 Information Retrieval Models Versus Retrieval Techniques	125
7.6.3 Toward Multimodel Information Retrieval Systems	125
References	126

CHAPTER 8

Information Retrieval Systems 129

8.1 Online Systems: Pioneer Information Retrieval Systems	129
8.1.1 Features of Online Information Retrieval Systems	130
8.1.2 Online Systems and Information Retrieval	130
8.2 CD-ROM Systems: A Different Medium for Information Retrieval Systems	131
8.2.1 Features of CD-ROM Systems	131
8.2.2 CD-ROM Systems and Information Retrieval	133
8.3 OPACs: Computerized Library Catalogs as Information Retrieval Systems	133
8.3.1 Features of OPACs	134
8.3.2 OPACs and Information Retrieval	136
8.4 Internet Retrieval Systems: The Newest Member in the Family of Information Retrieval Systems	137
8.4.1 Taxonomy of Internet Retrieval Systems	138
8.4.2 Features of Internet Retrieval Systems	141
8.4.3 Generations of Internet Retrieval Systems	149
8.4.4 Internet Retrieval Systems and Information Retrieval	152
8.5 Information Retrieval Systems: Some Trends	153
8.5.1 Convergence of Information Retrieval Systems	153
8.5.2 Web 2.0 and Information Retrieval Systems	154
References	155

CHAPTER 9

Retrieval of Information Unique in Content or Format 161

9.1 Multilingual Information	161
9.1.1 Multilingual Information Retrieval in the Past	162
9.1.2 Multilingual Information Retrieval on the Internet	163
9.1.3 Research on Multilingual Information Retrieval	164
9.2 Multimedia Information	165

9.2.1 Still-Image Retrieval	167
9.2.2 Sound Retrieval	172
9.2.3 Moving-Image Retrieval	176
9.2.4 Multimedia Retrieval on the Internet	178
9.3 Hypertext and Hypermedia Information	179
References	180

CHAPTER 10

The User Dimension in Information

Representation and Retrieval	185
10.1 Users and Their Information Needs	188
10.2 The Cognitive Model and Other User-Centered Models	188
10.2.1 The Cognitive Model	188
10.2.2 Other User-Centered Models of Information Retrieval	190
10.3 User and System Interaction	191
10.3.1 Modes of User-System Interaction	191
10.3.2 Other Dimensions of User-System Interaction	196
10.3.3 Evaluation of User-System Interaction	199
10.4 The User and Information Retrieval in the Digital Age	201
References	201

CHAPTER 11

Evaluation of Information Representation and Retrieval

11.1 Evaluation Measures for Information	
Representation and Retrieval	207
11.1.1 Evaluation Measures for Information Representation	207
11.1.2 Evaluation Measures for Information Retrieval	210
11.2 Evaluation Criteria for Information Retrieval Systems	218
11.2.1 Evaluation Criteria for Online Systems	218
11.2.2 Evaluation Criteria for OPACS	220
11.2.3 Evaluation Criteria for Internet Retrieval Systems	223
11.2.4 Evaluation Criteria for Multimedia Retrieval Systems	226
11.2.5 Usability as Evaluation Criteria	227
11.3 Major Evaluation Projects for Information	
Representation and Retrieval	228
11.3.1 The Cranfield Tests	229
11.3.2 The TREC Series	236
11.4 A Final Word on Evaluation of Information Representation	
and Retrieval	249
References	250

CHAPTER 12

Artificial Intelligence in Information Representation

and Retrieval 259

- 12.1 Overview of Artificial Intelligence Research 259
- 12.2 Natural Language Processing 261
 - 12.2.1 The Role of Natural Language Processing in
Information Representation and Retrieval 261
 - 12.2.2 Automatic Summarization 264
 - 12.2.3 Question Answering 267
 - 12.2.4 Natural Language Searching 272
- 12.3 The Semantic Web 274
 - 12.3.1 Semantic Web Architecture 274
 - 12.3.2 The Semantic Web and Information Representation
and Retrieval 277
 - 12.3.3 Challenges to Semantic Web as an Artificial Intelligence
Application 280
- 12.4 Artificial Intelligence and Information Representation
and Retrieval 281
- References 282

About the Author 287

Index 289

Figures and Tables

Figure 1.1	Process of Information Representation and Retrieval	19
Table 2.1	Basic Approaches to Information Representation	37
Figure 3.1	Taxonomy of Multimedia Information	49
Table 4.1	Comparisons of Controlled Vocabularies	57
Figure 5.1	Process of Query Expansion	77
Table 5.1	Concept Analysis of a Search Request	85
Table 5.2	Concept Groups for Query Representation	86
Figure 5.2	Application of Boolean Operators	87
Table 6.1	Comparison of Searching and Browsing	105
Table 7.1	Features of Three Information Retrieval Models	124
Table 8.1	Major Internet Applications and Corresponding Retrieval Systems	139
Table 8.2	Three Generations of Internet Retrieval Systems	150
Table 11.1	Possible Retrieval Outcomes	210
Table 11.2	Recall Ratio of Four Indexing Systems	230
Table 11.3	Indexing Time Versus Recall Ratio	231
Table 11.4	Number of Participating Teams in TREC	237
Table 11.5	An Overview of TREC Tracks: TREC-4 through TREC-10 . . .	239
Table 11.6	An Overview of TREC Tracks: TREC 2002 through TREC 2009	240
Figure 12.1	Semantic Web Architecture	275
Figure 12.2	Example of RDF Graph	276

Preface to the Second Edition

Around eight years have passed since I completed the manuscript for the first edition of this book in July 2001. The time span itself spells out the need for a new edition in the rapidly developing field of information representation and retrieval (IRR). In addition, at least two-dozen reviews were written for the previous edition of the book. Those reviews contain many valuable suggestions and comments, which I not only appreciate highly but also would like to incorporate into this book. These two factors combined subsequently become the catalyst for writing the second edition.

While the entire book is updated and revised, the following list details the major changes I made to individual chapters in the present edition.

- Chapter 1: New section §1.1.2.5 Karen Spärck Jones (1935–2007)
- Chapter 2: New sections §2.1.4 Social Tagging and §2.3.1.4 RSS
- Chapter 3: New section §3.1.3.3 Digital Object Identifiers (DOIs)
- Chapter 4: New sections §4.4.1 Taxonomies, §4.4.2 Folksonomies, and §4.4.3 Ontologies
- Chapter 5: Added Search/Retrieve Web Service (SRW) and Search/Retrieve via URL (SRU) into §5.1.2.4 Multiple Database Searching
- Chapter 8: Added next generation of OPACs into §8.3 OPACs: Computerized Library Catalogs as Information Retrieval Systems; Major updates of §8.4.2.4 Ranking Techniques; New section §8.5.2 Web 2.0 and Information Retrieval Systems
- Chapter 9: Major updates of §9.1 Multilingual Information, §9.2.2 Sound Retrieval, and §9.2.3 Moving Image Retrieval
- Chapter 10: New section §10.2.2 Other User-Centered Models of Information Retrieval; Added Organic User Interface (OUI) into §10.3.1.4 Other Models of User-System Interaction
- Chapter 11: Deleted §11.2.2 Evaluation Criteria for CD-ROM Systems; New sections §11.2.4 Evaluation Criteria for Multimedia Retrieval Systems, §11.2.5 Usability as

xii Information Representation and Retrieval in the Digital Age

Evaluation Criteria; Major updates of §11.3.2.1.4 Retrieval Tasks

- Chapter 12: Expansion and major revision, including deletion of §12.2.2 Natural Language Model; addition of §12.2.2 Automatic Summarization, §12.2.3 Question Answering, §12.2.4 Natural Language Searching, and §12.3 The Semantic Web

In addition, I would like to clarify the focus and orientation of this book based on the feedback I received about the previous edition. First, IRR in this book is examined and discussed from the perspective of library and information science rather than from the viewpoint of computer science. Thus, systems design and implementation specifics such as algorithms are not included in this book. Second, information representation is presented as one necessary component in the process of IR and not treated as a domain parallel to information retrieval. Therefore, coverage of information representation in this book is much less than that of IR. Third, certain topics (e.g., information seeking behavior), are only described briefly in this book for two reasons: 1) This book is not intended to treat them extensively, and 2) excellent coverage of them can be found in other publications that are typically listed as references at the end of related chapters.

Finally, I wish to thank Long Island University for granting me sabbatical leave, without which I would not have been able to write a second edition of this book. The efforts of my graduate assistant, Fenfei Ouyang, in gathering materials for me are also gratefully acknowledged. Furthermore, I truly appreciate the advice and guidance of Samantha Hastings, ASIS&T Monograph Series Editor, Amy Reeve, Managing Editor of Books, and John B. Bryans, Editor-in-Chief and Publisher, Information Today, Inc., during the course of preparing the current edition. It is a pleasure to work with them all as always.

Heting Chu
Long Island, New York

Preface to the First Edition

Another book on information retrieval (IR)? Yes, because there are new topics and developments that need to be discussed in this field as we enter the digital age. In addition, two chapters of this book are devoted exclusively to information representation, a step that must be taken before information becomes retrievable.

We begin with an overview of information representation and retrieval (IRR) in Chapter 1, which reviews key concepts, key people, key events, and major developmental stages of the field. Chapters 2 and 3 examine basic approaches to information representation and other related topics. Given the significance of language in information representation and retrieval, Chapter 4 discusses natural language and controlled vocabulary, the two types of languages used in the field, along with their relationship and characteristics. Chapters 5, 6, and 7 focus on various aspects of retrieval: retrieval techniques, retrieval approaches, and retrieval models. Major types of information retrieval systems are then considered in Chapter 8 with special coverage of Internet retrieval systems, the rising star in the family of IR systems. Chapter 9 explores the retrieval of multilingual information, multimedia information, and hyper-structured information. The user dimension, a fundamental aspect in information representation and retrieval, is covered in Chapter 10. Chapter 11 surveys the complex and multifaceted evaluation issue in the field, including evaluation measures, evaluation methodology, and major evaluation projects. The last chapter of the book, Chapter 12, analyzes the role and potential of artificial intelligence (AI) in information representation and retrieval.

I have attempted to present a systematic, thorough yet nontechnical view of the field by using plain language, wherever possible, to explain complex topics. The emphasis of the book is placed upon the principles and fundamentals of information representation and retrieval rather than on descriptions of specific procedures, systems, or corresponding practices in the field. Once the reader understands these IRR principles and fundamentals, he or she should be able to apply them in different situations and environments. Attention is also paid specifically to topics and developments regarding information representation and retrieval in the digital age.

While *Information Representation and Retrieval* has an orientation toward the user, IRR system designers should find it helpful for understanding the field from the perspective of users. My intent is a book that will be useful to anyone who is interested in learning about the field, particularly those who are new to it. I strongly recommend that beginners read the book in the order in which it is organized, as later chapters are built upon the earlier ones.

xiv Information Representation and Retrieval in the Digital Age

Looking back, I might not have written this book at this point in my career if my colleague, Richard Smiraglia, had not initiated the contact for me with the publisher. My gratitude also goes to my students, who have shared my interest and enthusiasm in the field. I appreciate as well the sabbatical leave granted by Long Island University, without which I would not have had the luxury of time for undertaking this project. In addition, I would like to thank Michael Koenig, dean of Palmer School with which I am affiliated, and John Bryans, editor-in-chief of the book publishing division of Information Today, Inc., for making the publication of my manuscript a reality.

This book to a large extent was an integral part of my family life during the time of writing as my husband and daughter were also deeply involved with my work. For example, the book became a regular topic at our dinner table, and my family often had to spend weekend and evening hours without me at home. My daughter, Fangfei, is perhaps exposed more to *information retrieval* than any other kids her age. She continually urged me to work on my book even though, in her heart, she would have preferred I spend the time with her. The love, understanding, and support from my family have continuously been a source of energy and inspiration.

Heting Chu
Long Island, New York
hchu@liu.edu