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

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

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

CHAPTER 10

CHAPTER 11

Figures and Tables

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

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.

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