

"Chu writes concisely and clearly ... providing the essential information in a concise style. ... excellent."

—Péter Jacsó

Second Edition

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

Heting Chu

asis&t
American Society for
Information Science
and Technology

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

vi Information Representation and Retrieval in the Digital Age

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
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	185
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 207

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

viii Information Representation and Retrieval in the Digital Age

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