

INFORMATION

**Representation**

and

**Retrieval**

in the Digital Age

Heting Chu

*asis&t*

ASIST Monograph Series

# Contents

Figures and Tables .....	xi
Preface .....	xiii

## CHAPTER 1

<b>Information Representation and Retrieval: An Overview .....</b>	<b>1</b>
1.1 History and Development of Information Representation and Retrieval .....	1
1.1.1 Major Stages .....	1
1.1.1.1 Increased Demand (1940s–early 1950s) .....	1
1.1.1.2 Rapid Growth (1950s–1980s) .....	2
1.1.1.3 Demystified Phase (1980s–1990s) .....	3
1.1.1.4 The Networked Era (1990s–present) .....	4
1.1.2 Pioneers of the Field .....	5
1.1.2.1 Mortimer Taube (1910–1965) .....	5
1.1.2.2 Hans Peter Luhn (1896–1964) .....	7
1.1.2.3 Calvin N. Mooers (1919–1994) .....	10
1.1.2.4 Gerard Salton (1927–1995) .....	11
1.2 Elaboration on Key Concepts .....	12
1.2.1 Information .....	12
1.2.2 Information Representation .....	13
1.2.3 Information Retrieval .....	13
1.2.4 Digital Age .....	14
1.3 Major Components .....	14
1.3.1 The Database .....	15
1.3.2 The Search Mechanism .....	15
1.3.3 The Language .....	16
1.3.4 The Interface .....	17
1.4 The Essential Problem in Information Representation and Retrieval .....	17
1.4.1 The Process of Information Representation and Retrieval .....	18
1.4.2 The Limits of Information Representation and Retrieval .....	19
References .....	20

## CHAPTER 2

<b>Information Representation I: Basic Approaches .....</b>	<b>25</b>
2.1 Indexing .....	25
2.1.1 Types of Indexing .....	26
2.1.2 Automated and Automatic Indexing .....	26

## iv Information Representation and Retrieval in the Digital Age

2.1.3 Indexing in the Hyperstructure Environment .....	27
2.2 Categorization .....	28
2.2.1 Types of Categorization .....	28
2.2.2 Principles of Categorization .....	28
2.2.3 The Convergence of the Two Categorization Approaches .....	29
2.3 Summarization .....	29
2.3.1 Types of Summarization .....	30
2.3.1.1 Abstracts .....	30
2.3.1.2 Summaries .....	30
2.3.1.3 Extracts .....	31
2.3.2 The Issue of Representativeness .....	31
2.4 Other Methods of Information Representation .....	31
2.4.1 Citations .....	32
2.4.2 Strings .....	33
2.5 A Review of Basic Approaches to Information Representation .....	33
References .....	34

### CHAPTER 3

<b>Information Representation II: Other Related Topics .....</b>	<b>37</b>
3.1 Metadata .....	37
3.1.1 What Is Metadata? .....	37
3.1.2 Characteristics of Digital Information on the Net .....	38
3.1.3 Examples of Metadata Standards .....	38
3.1.3.1 Dublin Core (DC) .....	38
3.1.3.2 Resource Description Framework (RDF) .....	39
3.1.4 Some Questions and Concerns about Metadata .....	40
3.2 Full Text .....	41
3.2.1 Representation of Full-Text Information .....	41
3.2.2 Difficulties in Representing Full Text .....	41
3.3 Representation of Multimedia Information .....	42
3.3.1 Types of Multimedia Information .....	42
3.3.2 Two Major Representation Approaches .....	42
3.3.3 Challenges in Representing Multimedia .....	44
3.4 Further Elaboration on Information Representation .....	45
References .....	46

### CHAPTER 4

<b>Language in Information Representation and Retrieval .....</b>	<b>47</b>
4.1 Natural Language .....	47
4.2 Controlled Vocabulary .....	48
4.2.1. Thesauri .....	48
4.2.2 Subject Heading Lists .....	49
4.2.3 Classification Schemes .....	50

4.2.4 A Comparison of Thesauri, Subject Heading Lists, and  
Classification Schemes ..... 50

4.3 Natural Language vs. Controlled Vocabulary ..... 51

4.3.1 Different Eras of IRR Languages ..... 52

4.3.2 Why Natural Language or Why Controlled Vocabulary? ..... 52

4.3.2.1 The Synonym Issue ..... 53

4.3.2.2 The Homograph Issue ..... 53

4.3.2.3 The Syntax Issue ..... 53

4.3.2.4 The Accuracy Issue ..... 54

4.3.2.5 The Updating Issue ..... 54

4.3.2.6 The Cost Issue ..... 54

4.3.2.7 The Compatibility Issue ..... 55

4.4. Language for IRR in the Digital Age ..... 55

References ..... 57

**CHAPTER 5**

**Retrieval Techniques and Query Representation ..... 59**

5.1 Retrieval Techniques ..... 59

5.1.1 Basic Retrieval Techniques ..... 59

5.1.1.1 Boolean Searching ..... 59

5.1.1.2 Case Sensitive Searching ..... 61

5.1.1.3 Truncation ..... 61

5.1.1.4 Proximity Searching ..... 62

5.1.1.5 Field Searching ..... 63

5.1.2 Advanced Retrieval Techniques ..... 64

5.1.2.1 Fuzzy Searching ..... 64

5.1.2.2 Weighted Searching ..... 65

5.1.2.3 Query Expansion ..... 67

5.1.2.4 Multiple Database Searching ..... 68

5.2 Selection of Retrieval Techniques ..... 69

5.2.1 Functions of Retrieval Techniques ..... 69

5.2.2 Retrieval Performance ..... 70

5.2.2.1 Retrieval Techniques for Improving Precision ..... 70

5.2.2.2 Retrieval Techniques for Improving Recall ..... 72

5.3 Query Representation ..... 73

5.3.1 General Steps ..... 73

5.3.1.1 Concept Analysis ..... 74

5.3.1.2 Term Variations ..... 74

5.3.1.3 Term Conversion ..... 75

5.3.1.4 Application of Boolean Operators ..... 76

5.3.1.5 Use of Other Retrieval Techniques ..... 77

5.3.2 Difficulties with Query Representation ..... 78

5.3.3 The Automatic Approach ..... 79

References ..... 80

## CHAPTER 6

<b>Retrieval Approaches</b> .....	<b>81</b>
6.1 Retrieval by Searching .....	82
6.1.1 Characteristics of Searching .....	82
6.1.2 Types of Searching .....	82
6.1.3 Search Strategies .....	84
6.1.3.1 The Building Block Approach .....	84
6.1.3.2 The Snowballing Approach .....	84
6.1.3.3 The Successive Fraction Approach .....	85
6.1.3.4 The Most Specific Facet First Approach .....	86
6.1.3.5 Toward a "Quick/Convenient" Approach .....	86
6.2 Retrieval by Browsing .....	87
6.2.1 What Is Browsing? .....	87
6.2.2 Types of Browsing .....	88
6.2.3 Browsing Strategies .....	90
6.3 Searching and Browsing Integrated in Retrieval .....	91
6.3.1 Comparison of the Two Retrieval Approaches .....	91
6.3.1.1 Information Need .....	91
6.3.1.2 Efficiency and Potential for Improvement .....	92
6.3.1.3 Cognitive Load .....	92
6.3.1.4 Serendipity .....	92
6.3.1.5 Efforts .....	92
6.3.2 The Integrated Approach .....	93
Endnote .....	94
References .....	94

## CHAPTER 7

<b>Information Retrieval Models</b> .....	<b>97</b>
7.1 Foundation of All IR Models: Matching .....	97
7.1.1 Term Matching .....	97
7.1.2 Similarity Measurement Matching .....	98
7.2 The Boolean Logic Model .....	99
7.2.1 Strengths of the Boolean Logic Model .....	99
7.2.2 Limitations of the Boolean Logic Model .....	100
7.3 Vector Space Model .....	102
7.3.1 Strengths of the Vector Space Model .....	103
7.3.2 Limitations of the Vector Space Model .....	104
7.4 Probability Model .....	106
7.4.1 Strengths of the Probability Model .....	107
7.4.2 Limitations of the Probability Model .....	108
7.5 Extensions of Major IR Models .....	108
7.5.1 Extended Boolean Logic Model .....	109
7.5.2 Fuzzy Set Model .....	109
7.6 IR Models: A Further Look .....	111

7.6.1 A Review of the Major IR Models ..... 111  
 7.6.2 IR Models vs. Retrieval Techniques ..... 112  
 7.6.3 Toward Multimodel IR Systems ..... 113  
 References ..... 113

**CHAPTER 8**

**Information Retrieval Systems ..... 117**  
 8.1 Online Systems—Pioneer IR Systems ..... 117  
     8.1.1 Features of Online IR Systems ..... 117  
     8.1.2 Online Systems and Information Retrieval ..... 118  
 8.2 CD-ROM Systems—A New Medium for IR Systems ..... 119  
     8.2.1 Features of CD-ROM Systems ..... 119  
     8.2.2 CD-ROM Systems and Information Retrieval ..... 120  
 8.3 OPACs—Computerized Library Catalogs as IR Systems ..... 121  
     8.3.1 Features of OPACs ..... 122  
     8.3.2 OPACs and Information Retrieval ..... 123  
 8.4 Internet Retrieval Systems—  
     The Newest Member in the Family of IR Systems ..... 123  
     8.4.1 Taxonomy of Internet Retrieval Systems ..... 124  
         8.4.1.1 By Retrieval Approach ..... 124  
         8.4.1.2 By Application ..... 125  
         8.4.1.3 By Content ..... 127  
     8.4.2 Features of Internet Retrieval Systems ..... 128  
         8.4.2.1 Coverage and Source Information ..... 128  
         8.4.2.2 Indexing Mechanism ..... 129  
         8.4.2.3 Searching Facilities ..... 130  
         8.4.2.4 Ranking Techniques ..... 133  
         8.4.2.5 Search Modification ..... 134  
         8.4.2.6 Interface ..... 135  
     8.4.3 Generations of Internet Retrieval Systems ..... 136  
     8.4.4 Internet Retrieval Systems and Information Retrieval ..... 138  
 8.5 Convergence of Various IR Systems ..... 139  
 References ..... 140

**CHAPTER 9**

**Retrieval of Information Unique in Content or Format ..... 145**  
 9.1 Multilingual Information ..... 145  
     9.1.1 Multilingual Information Retrieval in the Past ..... 145  
     9.1.2 Multilingual Information Retrieval on the Internet ..... 146  
     9.1.3 Research on Multilingual Information Retrieval ..... 147  
 9.2 Multimedia Information ..... 148  
     9.2.1 Still Image Retrieval ..... 150  
         9.2.1.1 Description-Based Retrieval of Still Images ..... 151

viii **Information Representation and Retrieval in the Digital Age**

9.2.1.2 Content-Based Retrieval of Still Images .....	153
9.2.1.3 Integration of the Two Image Retrieval Approaches ....	154
9.2.2 Sound Retrieval .....	155
9.2.2.1 Description-Based Retrieval of Sound Information ....	156
9.2.2.2 Content-Based Retrieval of Sound Information .....	156
9.2.3 Moving Image Retrieval .....	158
9.2.4 Multimedia Retrieval on the Internet .....	159
9.3 Hypertext and Hypermedia Information .....	161
References .....	162

**CHAPTER 10**

**The User Dimension in Information**

**Representation and Retrieval .....** 167

10.1 Users and Their Information Needs .....	167
10.2 The Cognitive Model .....	169
10.2.1 Strengths of the Cognitive Model .....	170
10.2.2 Limitations of the Cognitive Model .....	170
10.3 User and System Interaction .....	171
10.3.1 Modes of User-System Interaction .....	171
10.3.1.1 Command Language .....	172
10.3.1.2 Menu Selection .....	172
10.3.1.3 Graphical Mode of Interaction .....	173
10.3.1.4 Other Modes of User-System Interaction .....	174
10.3.1.5 The Hybrid Mode of Interaction .....	175
10.3.2 Other Dimensions of User-System Interaction .....	175
10.3.2.1 Display Features .....	175
10.3.2.2 Output Options .....	177
10.3.2.3 Help Facilities .....	177
10.3.3 Evaluation of User-System Interaction .....	178
10.3.3.1 Time Needed for the User to Learn Specific IR Functions .....	178
10.3.3.2 Speed of Interaction .....	179
10.3.3.3 Rate of Errors by the User .....	179
10.3.3.4 Retention Over Time .....	179
10.3.3.5 The User's Satisfaction .....	179
10.4 The User and Information Retrieval in the Digital Age .....	180
References .....	181

**CHAPTER 11**

**Evaluation of Information Representation and Retrieval .....** 185

11.1 Evaluation Measures for Information Representation and Retrieval .....	185
11.1.1 Evaluation Measures for Information Representation .....	185

11.1.1.1 Accuracy .....	186
11.1.1.2 Brevity .....	186
11.1.1.3 Consistency .....	186
11.1.1.4 Objectivity .....	187
11.1.1.5 Clarity, Readability, and Usability .....	187
11.1.2 Evaluation Measures for Information Retrieval .....	188
11.1.2.1 Recall and Precision .....	188
11.1.2.1.1 The Notion of Relevance .....	189
11.1.2.1.2 Determination of All the Relevant Documents in a System .....	190
11.1.2.1.3 Other Criticisms of Recall and Precision .....	191
11.1.2.1.4 Variations of Recall and Precision Measures .....	192
11.1.2.2 Fallout .....	194
11.1.2.3 Generality .....	194
11.1.2.4 Single Measures for IR Evaluation .....	194
11.1.2.5 Other Evaluation Measures for Information Retrieval .....	195
11.2 Evaluation Criteria for IR Systems .....	195
11.2.1 Evaluation Criteria for Online Systems .....	196
11.2.2 Evaluation Criteria for CD-ROM Systems .....	198
11.2.3 Evaluation Criteria for OPACs .....	200
11.2.4 Evaluation Criteria for Internet Retrieval Systems .....	202
11.3 Major Evaluation Projects for Information Representation and Retrieval .....	205
11.3.1 The Cranfield Tests .....	206
11.3.1.1 Cranfield I .....	206
11.3.1.1.1 Test Design .....	206
11.3.1.1.2 Test Findings .....	207
11.3.1.2 Cranfield II .....	209
11.3.1.2.1 Test Design .....	209
11.3.1.2.2 Test Findings .....	210
11.3.1.3 Problems with the Cranfield Tests .....	211
11.3.1.4 Significance of the Cranfield Tests .....	212
11.3.2 The TREC Series .....	212
11.3.2.1 The Design of the TREC Series .....	213
11.3.2.1.1 Participant Teams .....	213
11.3.2.1.2 Test Documents .....	214
11.3.2.1.3 Topics and Queries .....	214
11.3.2.1.4 Retrieval Tasks .....	215
11.3.2.1.5 Evaluation and Relevance Judgments .....	219
11.3.2.2 Findings of TREC .....	219
11.3.2.3 Problems with TREC .....	220
11.3.2.4 Significance of TREC .....	220



**x Information Representation and Retrieval in the Digital Age**

11.4 A Final Word on IRR Evaluation ..... 222  
References ..... 222

**CHAPTER 12**

**Artificial Intelligence in Information Representation**

**and Retrieval ..... 229**  
12.1 Overview of AI Research ..... 229  
12.2 Natural Language Processing ..... 230  
    12.2.1 The Role of Natural Language Processing in IRR ..... 230  
    12.2.2 The Natural Language Model ..... 232  
12.3 Intelligent Agents ..... 233  
12.4 AI and Information Representation and Retrieval ..... 234  
References ..... 235

**About the Author ..... 237**

**Index ..... 239**