

Wolfgang G. Stock,
Mechtild Stock

Handbook of Information Science

DE GRUYTER
SAUR

Contents

A. Introduction to Information Science — 1

- A.1 What is Information Science? — 3
- A.2 Knowledge and Information — 20
- A.3 Information and Understanding — 50
- A.4 Documents — 62
- A.5 Information Literacy — 78

Information Retrieval

B. Propaedeutics of Information Retrieval — 91

- B.1 History of Information Retrieval — 93
- B.2 Basic Ideas of Information Retrieval — 105
- B.3 Relevance and Pertinence — 118
- B.4 Crawlers — 129
- B.5 Typology of Information Retrieval Systems — 141
- B.6 Architecture of Retrieval Systems — 157

C. Natural Language Processing — 167

- C.1 n -Grams — 169
- C.2 Words — 179
- C.3 Phrases – Named Entities – Compounds – Semantic Environments — 198
- C.4 Anaphora — 219
- C.5 Fault-Tolerant Retrieval — 227

D. Boolean Retrieval Systems — 239

- D.1 Boolean Retrieval — 241
- D.2 Search Strategies — 253
- D.3 Weighted Boolean Retrieval — 266

E. Classical Retrieval Models — 275

- E.1 Text Statistics — 277
- E.2 Vector Space Model — 289
- E.3 Probabilistic Model — 301
- E.4 Retrieval of Non-Textual Documents — 312

F. Web Information Retrieval — 327

- F.1 Link Topology — 329
- F.2 Ranking Factors — 345
- F.3 Personalized Retrieval — 361
- F.4 Topic Detection and Tracking — 366

G. Special Problems of Information Retrieval — 375

G.1 Social Networks and “Small Worlds” — 377

G.2 Visual Retrieval Tools — 389

G.3 Cross-Language Information Retrieval — 399

G.4 (Semi-)Automatic Query Expansion — 408

G.5 Recommender Systems — 416

G.6 Passage Retrieval and Question Answering — 423

G.7 Emotional Retrieval and Sentiment Analysis — 430

H. Empirical Investigations on Information Retrieval — 443

H.1 Informetric Analyses — 445

H.2 Analytical Tools and Methods — 453

H.3 User and Usage Research — 465

H.4 Evaluation of Retrieval Systems — 481

Knowledge Representation

I. Propaedeutics of Knowledge Representation — 501

I.1 History of Knowledge Representation — 503

I.2 Basic Ideas of Knowledge Representation — 519

I.3 Concepts — 531

I.4 Semantic Relations — 547

J. Metadata — 565

J.1 Bibliographic Metadata — 567

J.2 Metadata about Objects — 586

J.3 Non-Topical Information Filters — 598

K. Folksonomies — 609

K.1 Social Tagging — 611

K.2 Tag Gardening — 621

K.3 Folksonomies and Relevance Ranking — 629

L. Knowledge Organization Systems — 633

L.1 Nomenclature — 635

L.2 Classification — 647

L.3 Thesaurus — 675

L.4 Ontology — 697

L.5 Faceted Knowledge Organization Systems — 707

L.6 Crosswalks between Knowledge Organization Systems — 719

M. Text-Oriented Knowledge Organization Methods — 733

M.1 Text-Word Method — 735

M.2 Citation Indexing — 744

N. Indexing — 757

N.1 Intellectual Indexing — 759

N.2 Automatic Indexing — 772

O. Summarization — 781

O.1 Abstracts — 783

O.2 Extracts — 796

P. Empirical Investigations on Knowledge Representation — 807

P.1 Evaluation of Knowledge Organization Systems — 809

P.2 Evaluation of Indexing and Summarization — 817

Q. Glossary and Indexes — 827

Q.1 Glossary — 829

Q.2 List of Abbreviations — 851

Q.3 List of Tables — 854

Q.4 List of Figures — 855

Q.5 Index of Names — 860

Q.6 Subject Index — 879

Q.3 List of Tables

- Table B.6.1: Field Schema for Document Indexing and Display in the ifo Literature Database — 161
- Table C.3.1: Word-Concept Matrix — 209
- Table C.5.1: Dynamic Programming Algorithm for Computing the Edit Distance between *surgery* and *survey* — 233
- Table D.1.1: Boolean Operators — 245
- Table D.1.2: Monadic not-Operator — 246
- Table F.1.1: Link Matrix with Hub and Authority Weights — 337
- Table G.1.1: Examples for Graphs of Relevance for Information Retrieval — 378
- Table H.4.1: Functionality of a Professional Information Service — 486
- Table H.4.2: Performance Parameters of Retrieval Systems in the Cranfield Tests — 490
- Table H.4.3: Typical Query in TReC — 491
- Table I.4.1: Reflexivity, Symmetry and Transitivity of Paradigmatic Relations — 560
- Table I.4.2: Knowledge Organization Systems and the Relations They Use — 560
- Table L.2.1: Universal Classifications — 662
- Table L.2.2: Classifications in Health Care — 663
- Table L.2.3: Classifications in Intellectual Property Rights — 665
- Table L.2.4: Economic Classifications — 667
- Table L.2.5: Geographic Classifications — 669
- Table L.3.1: Abbreviations in Thesaurus Terminology — 685
- Table P.1.1: Mean Similarity Measures Comparing Different Methods of Knowledge Representation — 814
- Table P.1.2: Dimensions and Indicators of the Evaluation of KOSs — 815

Q.4 List of Figures

- Figure A.1.1: Information Science and its Sub-Disciplines — 5
- Figure A.1.2: Information Science and its Neighboring Disciplines — 14
- Figure A.2.1: Schema of Signal Transmission Following Shannon — 21
- Figure A.2.2: Knowledge Spiral in the SECI Model — 32
- Figure A.2.3: Building Blocks of Knowledge Management Following Probst et al. — 33
- Figure A.2.4: Simple Information Transmission — 36
- Figure A.2.5: Information Transmission with Human Intermediator — 37
- Figure A.2.6: Information Transmission with Mechanical Intermediation — 38
- Figure A.2.7: Intermediation—Phase I: Information Indexing — 44
- Figure A.2.8: Intermediation—Phase II: Information Retrieval — 45
- Figure A.2.9: Intermediation—Phase III: Further Processing of Retrieved Information — 45
- Figure A.3.1: Feedback Loop of Understanding Information — 52
- Figure A.3.2: Dimensions of the Analysis of Cognitive Work — 58
- Figure A.3.3: The “Cognitive Actor” in Knowledge Representation and Information Retrieval — 60
- Figure A.4.1: Digital Text Documents, Data Documents and their Surrogates in Information Services — 67
- Figure A.4.2: A Rough Classification of Document Types — 68
- Figure A.4.3: Documents and Surrogates — 69
- Figure A.5.1: Levels of Literacy — 79
- Figure A.5.2: Studying Information Literacy in Primary Schools — 84
- Figure B.2.1: Documentary Reference Unit — 109
- Figure B.2.2: Documentary Unit (Surrogate) of the Example of Figure B.2.1 in PubMed — 110
- Figure B.2.3: Information Indexing and Information Retrieval — 111
- Figure B.3.1: Aspects of Relevance — 120
- Figure B.3.2: Relevance Distributions: Power Law and Inverse-Logistic Distribution — 125
- Figure B.4.1: Basic Crawler Architecture — 131
- Figure B.5.1: Four Interpretations of “The man saw the pyramid on the hill with the telescope” — 142
- Figure B.5.2: Retrieval Systems and Terminological Control — 144
- Figure B.5.3: Interplay of Information Linguistics and Retrieval Models for Relevance Ranking in Content-Based Text Retrieval — 145
- Figure B.5.4: Working Fields of Information-Linguistic Text Processing — 147
- Figure B.5.5: Retrieval Models — 150
- Figure B.5.6: Retrieval Dialog — 152
- Figure B.6.1: Building Blocks of a Retrieval System — 158
- Figure B.6.2: Short German Sentence in the ASCII 7-Bit Code — 159
- Figure B.6.3: The Sentence from Figure B.6.2 in the ISO 8859-1 Code — 159
- Figure B.6.4: Inverted File for Texts in the Body of Websites — 163
- Figure C.2.1: Reading Directions in an Arabic Text — 180
- Figure C.2.2: The Ten Most Frequent Words in Training Documents for Four Languages — 182
- Figure C.2.3: List of Endings to Be Removed in the Lovins Stemmer — 189
- Figure C.2.4: Iterative Approach in the Porter Stemmer. Working Step 1 — 190
- Figure C.2.5: Document Frequency of the Tetragrams of the Word Form “Juggling” — 194
- Figure C.3.1: Additional Inverted Files via Word Processing — 199
- Figure C.3.2: Statistical Phrase Building — 201

- Figure C.3.3: Natural- and Technical-Language Knowledge Organization Systems as Tools for Indexing the Semantic Environment of a Concept — 209
- Figure C.3.4: Hierarchical Retrieval — 211
- Figure C.3.5: Excerpt of the WordNet Semantic Network — 212
- Figure C.3.6: Excerpt from a KOS — 213
- Figure C.3.7: Excerpt from a KOS with Weighted Relations — 214
- Figure C.4.1: Concepts—Words—Anaphora — 220
- Figure C.5.1: Personal Names Arranged by Sound — 229
- Figure D.1.1: Boolean Operators from the Perspective of Set Theory — 245
- Figure D.2.1: Command-Based Boolean Search on the Example of DialogWeb — 253
- Figure D.2.2: Menu-Based Boolean Search on the Example of Profound — 254
- Figure D.2.3: Host-Specific Database Search on the Example of Dialog File 411 — 259
- Figure D.2.4: The Building Blocks Strategy during Query Modification — 262
- Figure D.2.5: Growing “Citation Pearls” during Query Modification — 262
- Figure E.1.1: Frequency and Significance of Words in a Document — 279
- Figure E.2.1: Document-Term Matrix — 289
- Figure E.2.2: Document Space — 290
- Figure E.2.3: Three Documents and Two Queries in Vector Space — 292
- Figure E.3.1: Program Steps in Probabilistic Retrieval — 302
- Figure E.4.1: Dimensions of Facial Recognition — 315
- Figure E.4.2: Shot and Scene — 317
- Figure E.4.3: Music Representation via Audio Signals, Time-Stamped Events and Musical Notation — 321
- Figure F.1.1: Display with Direct Answer, Hit List and Further Search Options — 332
- Figure F.1.2: Fundamental Link Relationships — 334
- Figure F.1.3: Enhancement of the Initial Hit List (“Root Set”) into the “Base Set” — 335
- Figure F.1.4: Calculating Hubs and Authorities — 336
- Figure F.1.5: ModelWeb to Demonstrate the PageRank Calculation — 341
- Figure F.3.1: User Characteristics in the Search Process — 362
- Figure F.4.1: Working Steps of Topic Detection and Tracking — 368
- Figure F.4.2: The Role of “Named Entities” and “Topic Terms” in Identifying a New Topic — 370
- Figure G.1.1: Centrality Measurements in a Network — 380
- Figure G.1.2: Cutpoint in a Graph — 383
- Figure G.1.3: Bridge in a Graph — 384
- Figure G.1.4: “Small World” Network — 385
- Figure G.2.1: Term Cloud — 390
- Figure G.2.2: Statistical KOS—Low Resolution — 391
- Figure G.2.3: Statistical KOS—High Resolution — 392
- Figure G.2.4: Visualization of Search Results — 394
- Figure G.2.5: Mapping Informetric Results — 395
- Figure G.2.6: Mash-Up of Informetric Results and Maps — 396
- Figure G.2.7: Visualization of Photographer Movement in New York City — 397
- Figure G.3.1: Original, Translation and Retranslation via a Translation Program — 399
- Figure G.3.2: Working Steps in Cross-Language Information Retrieval — 401
- Figure G.3.3: Transitive Translation in Cross-Language Retrieval — 403
- Figure G.3.4: Gleaning a Translated Query via Parallel Documents and Text Passages — 405
- Figure G.4.1: Options for Query Expansion — 409
- Figure G.5.1: Social Network of CiteULike-Users Based on Bookmarks — 418
- Figure G.5.2: Social Network of Authors Based on CiteULike Tags — 419

- Figure G.5.3: Collaborative Item-to-Item Filtering on Amazon — 421
- Figure G.6.1: Working Steps in Question-Answering Systems — 426
- Figure G.7.1: Photo (from Flickr) and Emotion Tagging via Scroll Bar — 432
- Figure G.7.2: Presentation of Search Results of an Emotional Retrieval System — 435
- Figure H.1.1: Subjects and Research Areas of Informetrics — 446
- Figure H.2.1: Working Steps in Informetric Analyses — 457
- Figure H.2.2: Command-Based Informetric Analysis on the Example of Dialog — 458
- Figure H.2.3: Menu-Based Informetric Analysis on the Example of Web of Knowledge — 459
- Figure H.2.4: Informetric Time Series on the Example of STN International via the TABULATE Command — 460
- Figure H.2.5: Information Flow Analysis of Important Articles on Alexius Meinong Using Web of Science and HistCite — 462
- Figure H.3.1: Model of Information Behavior — 466
- Figure H.3.2: Variables of the Analysis of Corporate Information Needs — 471
- Figure H.3.3: Kuhlthau's Model of the Information Search Process — 472
- Figure H.4.1: A Comprehensive Evaluation Model for Retrieval Systems — 483
- Figure H.4.2: Calculation of MAP — 493
- Figure I.1.1: Example of a Figura of the Ars Magna by Llull — 506
- Figure I.1.2: Camillo's Memory Theater in the Reconstruction by Yates — 507
- Figure I.2.1: Methods of Knowledge Representation and their Actors — 526
- Figure I.2.2: Indexing and Summarizing — 528
- Figure I.3.1: The Semiotic Triangle in Information Science — 532
- Figure I.3.2: Epistemological Foundations of Concept Theory — 534
- Figure I.4.1: Semantic Relations — 548
- Figure I.4.2: Specific Meronym-Holonym Relations — 556
- Figure I.4.3: Expressiveness of KOSs Methods and the Breadth of their Knowledge Domains — 561
- Figure J.1.1: Traditional Catalog Card — 567
- Figure J.1.2: Perspectives on Formally Published Documents — 570
- Figure J.1.3: Document Relations: The Same or Another Document? — 572
- Figure J.1.4: Documents, Names and Concepts on Aboutness as Controlled Access Points to Documents and Other Information — 574
- Figure J.1.5: The Interplay of Document Relations with Names and Aboutness — 575
- Figure J.1.6: Catalog Entry in the Exchange Format (MARC) — 577
- Figure J.1.7: User Interface of the Catalog Entry from Figure J.1.6 at the Library of Congress — 578
- Figure J.1.8: User Interface of the Catalog Entry from Figure J.1.6 at the Healey Library of the University of Massachusetts Boston — 579
- Figure J.1.9: A Webpage's Metatags — 583
- Figure J.2.1: Updating a Surrogate about an Object — 588
- Figure J.2.2: Beilstein Database. Fields of the Attributes of Liquids and Gases — 589
- Figure J.2.3: Beilstein Database. Attribute "Critical Density of Gases" — 590
- Figure J.2.4: Beilstein Database. Searching on STN — 591
- Figure J.2.5: Hoppenstedt Firmendatenbank. Display of a Document — 592
- Figure J.2.6: Description of the Field Values for Identifying Watermarks According to the CDWA — 594
- Figure J.2.7: Example of Real-Time Flight Information — 595
- Figure J.3.1: Keyword List for Genres in the Hollis Catalog — 602
- Figure J.3.2: Entry of the Hollis Catalog — 603

- Figure J.3.3: Target-Group-Specific Access to Different Forms of Information — 604
- Figure K.1.1: Documents, Tags and Users in a Folksonomy — 613
- Figure K.1.2: Ideal-Typical Tag Distribution in Docsonomies — 614
- Figure K.1.3: “Dorothy’s Ruby Slippers” — 616
- Figure K.2.1: Tag Clusters Concerning *Java* on Flickr — 624
- Figure K.2.2: Power Tags in a Power Law Distribution — 625
- Figure K.2.3: Power Tags in an Inverse-Logistic Tag Distribution — 626
- Figure K.2.4: Tag Co-Occurrences with Tag *web2.0* in BibSonomy — 627
- Figure K.3.1: Criteria of Relevance Ranking when Using a Folksonomy — 630
- Figure L.1.1: Example of a Keyword Entry from the Keyword Norm File — 636
- Figure L.1.2: Keyword Entry in the CAS Registry File — 640
- Figure L.1.3: Connection Table of a Chemical Structure in the CAS Registry File — 641
- Figure L.1.4: Shorthand of a Chemical Compound — 641
- Figure L.1.5: Two Molecules (One with Isotopes) with Weak Bonds — 642
- Figure L.1.6: Markush Structure — 642
- Figure L.2.1: Multiple-Language Terms of a Notation — 651
- Figure L.2.2: Simulated Simple Example of a Classification System — 652
- Figure L.2.3: Search Sequence with Indirect Hits for Syncategoremata — 653
- Figure L.2.4: Extensional Identity of a Class and the Union of its Subclasses — 656
- Figure L.2.5: Thematic Relevance Ranking via Citation Order — 659
- Figure L.3.1: Entry, Preferred and Candidate Vocabulary of a Thesaurus — 676
- Figure L.3.2: Vocabulary and Conceptual Control — 677
- Figure L.3.3: Vocabulary Relation 1: Designations-Concept (Synonymy) — 678
- Figure L.3.4: Vocabulary Relation 2: Designation-Concepts (Homonymy) — 679
- Figure L.3.5: Vocabulary Relation 3: Intra-Concepts Relation (Splitting) — 679
- Figure L.3.6: Vocabulary Relation 4: Inter-Concepts Relation as Bundling — 680
- Figure L.3.7: Vocabulary Relation 5: Inter-Concepts Relation as Specification — 680
- Figure L.3.8: Descriptor Entry in MeSH — 687
- Figure L.3.9: Structure of a Multilingual Thesaurus — 691
- Figure L.3.10: Bottom-Up Approach of Thesaurus Construction and Maintenance — 693
- Figure L.5.1: Thesaurus Facet of Industries in Dow Jones Factiva — 712
- Figure L.5.2: Faceted Nomenclature for Searching Recipes — 714
- Figure L.5.3: Dynamic Classing as a Chart of Two Facets — 716
- Figure L.6.1: Shell Model of Documents and Preferred Models of Indexing — 720
- Figure L.6.2: Different Semantic Perspectives on Documents — 721
- Figure L.6.3: Upgrading a KOS to a More Expressive Method — 722
- Figure L.6.4: Cropping a Subset of a KOS — 724
- Figure L.6.5: Direct Concordances — 725
- Figure L.6.6: Concordances with Master — 725
- Figure L.6.7: Two Cases of One-Manyness in Concordances — 726
- Figure L.6.8: Non-Exact Intersections of Two Concepts — 726
- Figure L.6.9: Statistical Derivation of Concept Pairs from Parallel Corpora — 727
- Figure L.6.10: The Process of Unifying KOSs via Merging — 729
- Figure M.1.1: Surrogate According to the Text-Word Method — 737
- Figure M.1.2: Surrogate Following the Text-Word Method with Translation Relation — 740
- Figure M.2.1: Shepardizing on LexisNexis — 745
- Figure M.2.2: Bibliographic Coupling and Co-Citation — 750
- Figure N.1.1: Fixed Points of the Indexing Process — 760
- Figure N.1.2: Elements and Phases of Indexing — 762

- Figure N.1.3: Allocation of Concepts to the Objects of the Aboutness of a Documentary Reference Unit — 764
- Figure N.1.4: Typical Index Entry — 769
- Figure N.2.1: Fields of Application for Automatic Procedures during Indexing — 772
- Figure N.2.2: Rule-Based Automatic Indexing — 775
- Figure N.2.3: Cluster Formation via Single Linkage — 778
- Figure N.2.4: Cluster Formation via Complete Linkage — 779
- Figure O.1.1: Homomorphous and Paramorphous Information Condensation — 786
- Figure O.1.2: Indicative Abstract — 790
- Figure O.1.3: Informative Abstract — 790
- Figure O.1.4: Structured Abstract — 791
- Figure O.2.1: Working Steps in Automatic Extracting — 799
- Figure P.1.1: An Example of Semantic Inconsistency — 811
- Figure P.1.2: An Example of a Circularity Error — 812
- Figure P.1.3: An Example of a Skipping Error — 813
- Figure P.2.1: Indexing Consistency and the “Meeting” of User Interests — 821