

Health Informatics

This series is directed to healthcare professionals leading the transformation of healthcare by using information and knowledge. For over 20 years, Health Informatics has offered a broad range of titles: some address specific professions such as nursing, medicine, and health administration; others cover special areas of practice such as trauma and radiology; still other books in the series focus on interdisciplinary issues, such as the computer based patient record, electronic health records, and networked healthcare systems. Editors and authors, eminent experts in their fields, offer their accounts of innovations in health informatics. Increasingly, these accounts go beyond hardware and software to address the role of information in influencing the transformation of healthcare delivery systems around the world. The series also increasingly focuses on the users of the information and systems: the organizational, behavioral, and societal changes that accompany the diffusion of information technology in health services environments.

Developments in healthcare delivery are constant; in recent years, bioinformatics has emerged as a new field in health informatics to support emerging and ongoing developments in molecular biology. At the same time, further evolution of the field of health informatics is reflected in the introduction of concepts at the macro or health systems delivery level with major national initiatives related to electronic health records (EHR), data standards, and public health informatics.

These changes will continue to shape health services in the twenty-first century. By making full and creative use of the technology to tame data and to transform information, Health Informatics will foster the development and use of new knowledge in healthcare.

More information about this series at <http://www.springer.com/series/1114>

William Hersh

Information Retrieval: A Biomedical and Health Perspective

Fourth Edition

 Springer

William Hersh
Oregon Health & Science University
Portland, OR
USA

ISSN 1431-1917
Health Informatics

ISSN 2197-3741 (electronic)

ISBN 978-3-030-47685-4

ISBN 978-3-030-47686-1 (eBook)

<https://doi.org/10.1007/978-3-030-47686-1>

© Springer Nature Switzerland AG 2020, corrected publication 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Sally, Becca, Alyssa, and AJ

Preface

The main goal of this book is to provide an understanding of the theory, implementation, and evaluation of information retrieval (IR) systems in biomedicine and health. There is already a great deal of “how-to” information on searching for biomedical and health information (some listed in Chap. 1). Similarly, there are also a number of high-quality basic IR textbooks (also listed in Chap. 1). This volume is different from all of the above in that it covers basic IR as do the latter books, but with a distinct focus on the biomedical and health domain.

The first three editions of this book were published in 1996, 2003, and 2009. Although subsequent editions of books in many fields represent incremental updates, this edition is profoundly rewritten and is essentially a new book. The IR world has changed substantially since I wrote the first three editions of the book. At the time of the first edition, IR systems were available and not too difficult to access if you had the means and expertise. Also, in that edition, the Internet was a “special topic” in the very last chapter of the book. By the second edition, the World Wide Web had become a widespread platform for the use of information access and delivery, but had not achieved the nearly ubiquitous and saturated use it has now. At present, however, not only must health care professionals and biomedical researchers understand how to use IR systems to be effective in their work, but patients and consumers must also as well to attain optimal health care.

Similar to previous editions will be the maintenance of a Web site for errata and updates. The Website <http://www.irbook.info/> will identify all errors in the book text as well as provide updates on important new findings in the field as they become available.

As in the first three editions, the approach is still to introduce all the necessary theory to allow coverage of the implementation and evaluation of IR systems in biomedicine and health. Any book on theoretical aspects must necessarily use technical jargon, and this book is no exception. Although jargon is minimized, it cannot be eliminated without retreating to a more superficial level of coverage. The reader’s understanding of the jargon will vary based on their background, but anyone with some background in computers, libraries, health, and/or biomedicine should be

able to understand most of the terms used. In any case, an attempt to define all jargon terms is made.

Another approach is to attempt wherever possible to classify topics, whether discussing types of information or models of evaluation. I have always found classification useful in providing an overview of complex topics. One problem, of course, is that everything does not fit into the neat and simple categories of the classification. This occurs repeatedly with IR, and the reader is forewarned.

This book had its origins in a tutorial taught at the former Symposium on Computer Applications in Medicine (SCAMC) meeting. The content continues to grow each year through my course taught to biomedical informatics students in the on-campus and disease-learning programs at OHSU. (Students often do not realize that next year's course content is based in part on the new and interesting things they teach me!) The book can be used in either a basic information science course or a biomedical and health informatics course. It should also provide a strong background for others interested in this topic, including those who design, implement, use, and evaluate IR systems.

Interest continues to grow in biomedical and health IR systems. I entered a fellowship in medical informatics at Harvard University in the late 1980s, during the initial era of medical artificial intelligence. I had assumed I would take up the banner of some aspect of that area, such as knowledge representation. But along the way I came across a reference from the field of "information retrieval." It looked interesting, so I looked at the references of that reference. It did not take long to figure out that this was where my real interests lay, and I spent many an afternoon in my fellowship tracing references in the Harvard University and Massachusetts Institute of Technology libraries. Even though I had not yet heard of the field of bibliometrics, I was personally validating all its principles. Like many in the field, I have been amazed to see IR become so "mainstream" with its routine use by almost everyone on the planet.

The book is divided into eight chapters. Chapter 1 provides basic definitions and models that will be used throughout the book. It also points to resources for the field and introduces evaluation of systems. Chapter 2 provides an overview of biomedical and health information, describing some of the issues in its production, dissemination, and use. Chapter 3 gives an overview of the great deal of content that is currently available. Chapters 4 and 5 cover the two fundamental intellectual tasks of IR, indexing and retrieval, with the predominant paradigms of each discussed in detail. Chapter 6 discusses the methods and challenges of larger information access. Chapter 7 focuses on evaluation research that has been done on state-of-the-art systems in the biomedical and health domain. Finally, Chapter 8 explores research about IR systems and their users, with an emphasis on applications in the biomedical and health domain. Within each chapter, the goal is to provide a comprehensive overview of the topic, with thorough citations of pertinent references. There is a preference for discussing biomedical and health implementations of principles, but where this is not possible, the original domain of implementation is discussed.

This book would not have been possible without the influence of various mentors, dating back to high school, who nurtured my interests in science generally

and/or biomedical and health informatics specifically, and/or helped me achieve my academic and career goals. The most prominent include Mr. Robert Koonz (then of New Trier West High School, Northfield, IL), Dr. Darryl Sweeney (then of University of Illinois at Champaign-Urbana), Dr. Robert Greenes (then of Harvard Medical School), Dr. David Evans (then of Carnegie Mellon University), Dr. Mark Frisse (then of Washington University), Dr. J. Robert Beck (then of OHSU), Dr. David Hickam (then of OHSU), Dr. Brian Haynes (McMaster University), Dr. Lesley Hallick (then of OHSU), and Dr. Jerris Hedges (then of OHSU). I must also acknowledge the late Dr. Gerard Salton (Cornell University), whose writings initiated and sustained my interest in this field.

I would also like to note the contributions of institutions and people in the federal government who aided the development of my career and this book. While many Americans increasingly question the abilities of their government to do anything successfully, the NLM, under the former directorship of the late Dr. Donald A. B. Lindberg and the current directorship of Dr. Patricia Flatley Brennan, has led the growth and advancement of the field of biomedical and health informatics. The NLM's fellowship and research funding have given me the skills and experience to succeed in this field. I would also like to acknowledge the late Oregon Senator Mark O. Hatfield through his dedication to biomedical research funding that aided myself and many others.

Finally, this book also would not have been possible without the love and support of my family. All of my parents, Mom and Jon, Dad and Gloria, as well as my brother Jeff and sister-in-law Myra, supported the various interests I developed in life and the somewhat different career path I chose. I think that now as they have become Web users and searchers, they appreciate my interest in this area. And last, but most importantly, has been the contribution of my wife, Sally, and two children, Becca and Alyssa, whose unlimited love and support made this undertaking so enjoyable and rewarding.

March 2020
Portland, OR

William Hersh

Contents

1	Foundations	1
1.1	Basic Definitions	3
1.2	Scientific Disciplines Concerned with IR	5
1.3	Models of IR	7
1.3.1	The Information World	7
1.3.2	Users	8
1.3.3	Health Decision-Making	9
1.3.4	Knowledge Acquisition and Use	9
1.4	IR Resources	11
1.4.1	Organizations	11
1.4.2	Journals	12
1.4.3	Texts	13
1.4.4	Tools	13
1.5	The Internet and World Wide Web	14
1.5.1	Users	15
1.5.2	Usage	16
1.5.3	Hypertext and Linking	18
1.6	Evaluation	19
1.6.1	Classification of Evaluation	21
1.6.2	Relevance-Based Evaluation	24
1.6.3	Challenge Evaluations	30
	References	34
2	Information	41
2.1	What Is Information?	41
2.2	Theories of Information	42
2.3	Properties of Scientific Information	45
2.3.1	Growth	45
2.3.2	Obsolescence	46
2.3.3	Fragmentation	48

2.3.4	Linkage and Citations	48
2.3.5	Propagation	60
2.4	Classification of Health Information	61
2.5	Production of Biomedical and Health Information	64
2.5.1	Generation of Scientific Information	64
2.5.2	Peer Review	69
2.5.3	Primary Literature	73
2.5.4	Systematic Reviews and Meta-Analysis	94
2.5.5	Secondary Literature	99
2.6	Electronic Publishing	101
2.6.1	Electronic Scholarly Publication	102
2.6.2	Consumer Health Information	103
2.7	Use of Knowledge-Based Health Information	108
2.7.1	Models of Physician Thinking	109
2.7.2	Physician Information Needs	110
2.7.3	Information Needs of Other Healthcare Professionals	115
2.7.4	Information Needs of Biomedical and Health Researchers	115
2.7.5	Information Needs of Consumers	116
2.8	Summary	116
	References	116
3	Content	141
3.1	Classification of Health and Biomedical Information Content	141
3.2	Bibliographic Content	143
3.2.1	Literature Reference Databases	144
3.2.2	Web Catalogs and Feeds	153
3.2.3	Specialized Registries	154
3.3	Full-Text Content	155
3.3.1	Periodicals	156
3.3.2	Books and Reports	157
3.3.3	Web Collections	159
3.4	Annotated Content	161
3.4.1	Images and Videos	162
3.4.2	Citations	163
3.4.3	Evidence-Based Medicine Resources	163
3.4.4	Molecular Biology and -Omics	166
3.4.5	Educational Resources	169
3.4.6	Linked Data	170
3.4.7	Other Annotated Content	170
3.5	Aggregations	172
3.5.1	Consumer Health	172
3.5.2	Health Professionals	173
3.5.3	Body of Knowledge	175
3.5.4	Model Organism Databases	176
3.5.5	Scientific Information	176
	References	177

4	Indexing	181
4.1	Types of Indexing	181
4.2	Factors Influencing Indexing	182
4.3	Controlled Vocabularies	183
4.3.1	General Principles of Controlled Vocabularies	184
4.3.2	The Medical Subject Headings (MeSH) Vocabulary	185
4.3.3	Other Indexing Vocabularies	191
4.3.4	The Unified Medical Language System	194
4.4	Manual Indexing	197
4.4.1	Bibliographic Manual Indexing	198
4.4.2	Full-Text Manual Indexing	200
4.4.3	Web Manual Indexing	200
4.4.4	Limitations of Manual Indexing	206
4.5	Automated Indexing	207
4.5.1	Word Indexing	207
4.5.2	Limitations of Word Indexing	207
4.5.3	Word Weighting	208
4.5.4	Link-Based Indexing	212
4.5.5	Web Crawling	213
4.6	Indexing Annotated Content	214
4.6.1	Index Imaging	214
4.6.2	Indexing Learning Objects	215
4.6.3	Indexing Biomedical and Health Data	218
4.7	Data Structures for Efficient Retrieval	218
	References	220
5	Retrieval	225
5.1	Search Process	226
5.2	General Principles of Searching	226
5.2.1	Exact-Match Searching	227
5.2.2	Partial-Match Searching	229
5.2.3	Term Selection	233
5.2.4	Other Attribute Selection	237
5.3	Searching Interfaces	237
5.3.1	Bibliographic	237
5.3.2	Full Text	249
5.3.3	Annotated	252
5.3.4	Aggregations	257
5.4	Document Delivery	257
5.5	Notification or Information Filtering	258
	References	259
6	Access	261
6.1	Libraries	261
6.1.1	Definitions and Functions of DLs	263
6.2	Access to Content	265

6.2.1	Access to Individual Items	265
6.2.2	Access to Collections	267
6.2.3	Access to Metadata	268
6.2.4	Integration with Other Applications	269
6.3	Copyright and Intellectual Property	272
6.3.1	Copyright and Fair Use	273
6.3.2	Digital Rights Management	275
6.4	Open Access and Open Science	276
6.4.1	Open-Access Publishing	277
6.4.2	NIH Public Access Policy	278
6.4.3	Predatory Journals	280
6.5	Preservation	281
6.6	Librarians, Informationists, and Other Professionals	282
6.7	Future Directions	283
	References	284
7	Evaluation	289
7.1	Usage Frequency	290
7.2	Types of Usage	292
7.3	User Satisfaction	294
7.4	Searching Quality	294
7.4.1	System-Oriented Performance Evaluations	295
7.4.2	User-Oriented Performance Evaluations	299
7.5	Factors Associated with Success or Failure	310
7.5.1	Predictors of Success	310
7.5.2	Analysis of Failure	314
7.6	Assessment of Impact	316
7.7	Research on Relevance	319
7.7.1	Topical Relevance	319
7.7.2	Situational Relevance	320
7.7.3	Research About Relevance Judgments	321
7.7.4	Limitations of Relevance-Based Measures	323
7.7.5	Automating Relevance Judgments	325
7.7.6	Measures of Agreement	326
7.8	What Has Been Learned About IR Systems?	327
	References	329
8	Research	337
8.1	Frameworks and Challenge Evaluations	337
8.2	Biomedical and Health IR Research	343
8.2.1	Early Studies	344
8.2.2	Challenge Evaluations in Biomedicine and Health	346
8.2.3	Ad Hoc Retrieval	353
8.2.4	Consumer-Oriented	355
8.2.5	Image Retrieval	356
8.2.6	High-Recall Retrieval	357
8.2.7	EHR Retrieval	358

- 8.3 General IR Research 360
 - 8.3.1 Overview of Early Research 360
 - 8.3.2 Machine Learning: Uncovering Latent Meaning 363
 - 8.3.3 Natural Language Processing 366
 - 8.3.4 Question-Answering 368
 - 8.3.5 Text Categorization 377
- 8.4 Research Systems and the User 381
 - 8.4.1 Early Research 382
 - 8.4.2 User Evaluation of Research Systems 383
 - 8.4.3 TREC Interactive Track 384
- 8.5 Looking Forward 389
- References 389
- Correction to: Retrieval C1**
- Index 407**