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 ISSN 2197-3741 (electronic)
Health Informatics
ISBN 978-3-030-47685-4 ISBN 978-3-030-47686-1 (eBook)
https://doi.org/10.1007/978-3-030-47686-1

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 William Hersh
Portland, OR

Contents