Inf Retrieval (2010) 13:96–100
DOI 10.1007/s10791-009-9100-5

BOOK REVIEW

William Hersh: Information retrieval: a health and biomedical perspective, 3rd ed

Health and Informatics Series, Springer, 2009, 504 pp, $79.95, ISBN: 978-0-387-78702-2)

Nicola Stokes

Published online: 30 June 2009

As a graduate student I received an invaluable piece of advice—*know your data*. At the time, like many a wet-eared student, I had fallen into the trap of blindly changing parameters in an attempt to get my results to match my preconceived conclusions. Examining my data collection and user information needs (queries) rather than my evaluation metric scores, quickly illuminated the problem. From that day forth, I never tackled an experiment without first familiarising myself with the data; an approach which has serves me well when the domain of interest is written in my native tongue and on a topic that requires little additional expertise to interpret. No such luck during my recent adventures in the world of TREC Genomics which I'm reluctant to admit, required some significant undergraduate biology revision.

Another information source which would have greatly aided me in this endeavour, if it had been published in time, is William Hersh's new edition of his health informatics textbook—*Information Retrieval*: *A health and biomedical perspective*. Hersh is one of those rare individuals who has a true understanding of what medical professionals want, and what computer scientists can do for them. He is both a qualified MD and a Prof. and Chair at the Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, where he teaches among other things a biomedical informatics course from which the aforementioned book evolved.

Naturally then this publication has a textbook feel to it. It is split into three parts. Part I: Basic Concepts, covers just that—the IR concepts that are necessary for novices to understand more of the technical content introduced in subsequent chapters. Part II: State of the Art, looks at four different aspects of biomedical IR—available content (from literature reference databases to consumer information); how this data is indexed for optimal retrieval; the retrieval process (from query formulation and term weighting to interface design); and the role of IR systems in Digital Libraries. Part III: Research Directions examines future and current research topics in biomedical IR: relevance metrics versus user-oriented evaluation; IR system- and user-oriented research; Natural Language Processing (NLP) applications such as summarisation and question answering.

N. Stokes (✉)
School of Computer Science & Informatics, University College Dublin, Dublin, Ireland
e-mail: nicola.stokes@ucd.ie

Like any good textbook, Hersh has evolved its contents to reflect changing trends in healthcare and IR since its first edition was published in 1996. Hersh states that the third and latest incarnation of the book has been “profoundly rewritten and is essentially a new book compared to the 2003 version”. These changes reflect the profound effect that the growth of the World Wide Web has had on information dissemination in this scientific domain. For example, by the 2nd edition electronically available publications were on the increase, by 2009 they are nearly the norm. In addition, Hersh points out that we now live in a world where the consumer demographic of healthcare information includes healthcare professionals as well as patients and even *cyberchondriacs*. Hersh believes that in order to achieve optimal healthcare both professional and lay communities need to understand how biomedical IR engines tick.

As an undergraduate/graduate textbook for Computer Science students, this book lacks sufficient IR implementation details (data structures and algorithms) and all the nitty-gritty mathematically content, and end-of-chapter exercises that go with that. There are, of course, plenty of decent books that already cover IR theory at that level of detail (Salton and McGill 1986; Witten et al. 1999; Manning et al. 2008; to name but a few). Instead Hersh’s book offers engaging content on a domain specific application of search from an Information Science perspective. This book will no doubt help Computer Science students studying IR realise the importance of capturing user information needs in a field that primarily focuses on system engineering considerations.

For more experienced IR readers, in particular those who find themselves as I did branching out in a Genomic or similar biomedical domain, this book is essential reading. In this case, the expectations of the reader will typically focus on three areas: what types of information repositories constitute biomedical data; how do information needs vary across the different biomedical professions; what unique IR research challenges does biomedical IR present, and what has been achieved so far?

Chapter 2 concerns itself with explaining aspects of scientific publishing; its motivations; its measures; its characteristics; its flaws. For seasoned researchers, there is nothing new here, until the second half of the chapter where we reach a series of subsections on health informatics topics such as the electronic publishing of consumer health information, the use of knowledge-based health information and Evidence-based Medicine.

Obviously, what we are looking for here are pointers to studies that report on typical information seeking behaviours of health and biomedical professionals. While there is no reason to suggest that Hersh’s background research is anything but thorough, it is still surprising to see that so many of the studies discussed are from the 1990s; a time when, for example, only a 26% of family physicians had computers in their offices, and 45% of them a computer at home (Ely et al. 1999). Hersh alludes to the difficulties faced by Information Scientists seeking these answers given that many health professionals aren’t directly aware of gaps in their knowledge. Also early studies showed that health professionals have in the past had very modest interaction with IR systems “where the average user seeks answers to clinical questions with online resources only a few times per month”. Even when clinicians were found to have two unanswered questions for every three patients sessions, they only pursued answers one-third of the time (Gorman 1995). In any case, it is clear that new studies are now required that reflect the fact that digitised content is the norm, and user communities extend beyond the general practitioner, e.g., biomedical researchers, nurses, patients, etc.

Hersh’s final section in this chapter, prompts the reader to wonder how IR systems can be evolved to support the Evidence-based Medicine paradigm where reports on “best practice” are used to inform clinical decision making. Finding suitable studies and other meta-analysis; distinguishing evidence from propaganda; synthesising and condensing this knowledge into guidelines that can inform healthcare intervention—are all examples of challenging "next generation" information processing tasks that will be of interest to IR and NLP researchers alike.

A categorisation of *health and biomedical information content* is provided in Chapter 3. IR applications reported in the literature have primarily focussed on the retrieval of abstracts from bibliographic database such as MEDLINE. I must admit during my own categorisation attempts I was shocked to discover the sheer volume of electronically available information that didn't fit into the MEDLINE subcategory. Chapter 3 provides an overview of these alternative data sources: *web catalogs and feeds*; *full texts* such as books and reports; *annotated databases* (image databases, model organism databases etc., molecular biology databases (e.g., Entrez Gene)), and *aggregated data collections* such as MedlinePlus (for consumer health) and NLM Gateway (for molecular biology).

The availability of these repositories should in theory encourage the adoption of new research tasks; however, as Chapter 6 (*Digital Libraries*) points out much of the data in this domain is proprietary. In recent times, there has been a move by prominent funding agencies (e.g., NIH, the UK Medical Research Council, the Welcome Trust etc.) to ensure that all publicly funded research papers are submitted to open-access repositories within 12 months of publication. It is therefore hoped that in the near future Open-Access Publishing will become the norm, and IR researchers will have access to larger collections of full-text scientific publications—which at the moment is a significant barrier for entry into this domain.

In Chapter 5 on *Retrieval*, IR systems are described with respect to the health and biomedical information sources they index. Like Chapter 3, the aim of this section of the book is to be representative rather than exhaustive in coverage. A large portion of the chapter is therefore dedicated to the PubMed retrieval engine (from ranking strategy to user interface design), which is one of the most widely used engines for accessing MEDLINE's 16 million scientific references. An integral part of the PubMed retrieval framework is MeSH, a controlled vocabulary of around 23,000 medical subject headings. MeSH and the proliferation of other indexing vocabularies used in the medical domain are described in detail in Chapter 4, which also briefly covers automatic IR indexing strategies.

Complementing Chapters 4 and 5 is Chapter 8, which provides an overview of the state-of-the-art in *System- and User-oriented IR Research*, as well as an historical account of how we got there. Section 8.1 on *System-oriented Research* discusses IR models and their various term weighting schemes (without a complicated formula in sight!). Subsections on relevance feedback, query expansion and passage retrieval culminate the discussion of what Hersh refers to as *lexical-statistical IR systems*. *Linguistic IR systems*, on the other hand, attempt to take advantage of linguistic mark-up such as syntactic classes (parts-of-speech) and semantic annotations (disambiguated text). Applications of these systems are discussed in the biomedical domain (TREC Genomics, Medical Image Retrieval at CLEF) and elsewhere (TREC Web, Blog Tracks; Cross-lingual IR and CLEF).

Section 8.2 discusses advances in content delivery and other aspects of User-oriented IR research. A reoccurring observation made through out the book is that—"the build it and they will come" mentality is overly simplified in the health and biomedical domain. For example, studies published so far indicated that the impact of IR engines is modest in a clinical setting and is only used to answer a small amount of information needs, with people preferring to use textbooks or seek advice from colleagues. Hence, it seems the focus for researchers should be on building complete solutions rather than enabling technologies. This of course involves embedding IR functionality into systems that health professionals actively already use. Hersh provides a number of examples such as integrating clinical information in an *Electronic Health Record* system where the IR system is working in the background to formulate a series of useful queries and relevant responses. *Infobuttons* embedded in the health record provide users with the option of, for example, clicking to access more information on a particular drug treatment. The importance of handheld alternatives to desktop search is also stressed since health care professionals tend to spend many more hours away from their desks than at them.

Hersh's emphasis on user-oriented IR research is also mirrored in his overview of IR *Evaluation* in Chapter 7. In contrast to many other IR textbooks, the focus is less on evaluation metrics in an in vitro evaluation environment (e.g., TREC), and more on measuring performance in real user-system interaction scenarios.

Chapter 9, the book's final chapter, examines text processing technologies beyond IR, specifically those that fall under the NLP banner such as information extraction, text mining, text categorisation, question answering and text summarisation. There is a symbiotic relationship between many of these NLP technologies and IR, so it makes sense that Hersh has added this chapter to his latest edition. For example, in the case of NLP applications, text retrieval is often used as an initial filtering step to reduce the size of a large document collection that would have otherwise made computationally expensive, linguistic analysis infeasible. While for search technologies, NLP offers alternative results presentation strategies (summaries, exact answers, automatic database population etc.) that can help users wade through vast amounts of information.

To avoid the need for additional background content, this chapter outlines NLP tasks rather than techniques. Pointers to more in depth coverage of the BioNLP community's activities are provided (Ananiadou and McNaught 2006). What this chapter shows is that biomedical text processing has also been enthusiastically embraced by the NLP community. There are a number of reasons for this including the fact that basic processing tools such as tokenisers, part-of-speech and named entity taggers have all had to be re-engineered for this domain. In addition, results from the TREC Genomics track show that this is one of those rare domains where ontology-based query expansion can provide *significant* and *consistent* gains in performance over baseline runs (Zhou et al. 2007; Stokes et al. 2009); a result which has eluded researchers looking at WordNet-based query expansion (Voorhees 1994).

The challenges faced by the BioNLP community are familiar as they surround the gaining of access to confidential data (e.g., clinical narratives), proprietary data, and sufficient amounts of annotated data for training statistical based techniques. Hersh points out that the robustness of linguistic techniques when faced with noisy data (e.g., misspellings and ungrammatical sentences) is also an area where BioNLP techniques need to prove their mettle. So there is still plenty of work to be done, and many motivating reasons why NLP and IR researchers should be working on these tasks together.

In conclusion then, Hersh's book offers the reader a fascinating insight into the additional complexities associated with the building of systems for users as opposed to evaluation competitions. Hence, I would recommend this book to IR researchers, including those who aren't working on biomedical IR. I also found Hersh to be an engaging writer; even when I was familiar with the content, his clear and concise writing style encouraged me to read on. Hersh's homepage (http://www.billhersh.info/) contains a wealth of supplementary information on the emerging field of Health Informatics (see his *Informatics Professor blog*). Likewise readers will find the book's website (http://medir.ohsu.edu/~hersh/irbook/) a valuable source of dynamic content, containing errata and links to relevant papers not published in time to make it into this edition of the book.

References

Ananiadou, S., & McNaught, J. (Eds.). (2006). *Text mining for biology and biomedicine*. Boston, MA: Artech House.

Ely, J., Osheroff, J., et al. (1999). Analysis of questions asked by family doctors regarding patients care. *British Medical Journal, 319*, 358–361.

Gorman, P. (1995). Information needs of physicians. *Journal of the American Society for Information Science, 46*, 729–736.

Manning, C. D., Raghavan, P., & Schtze, H. (2008). *Introduction to information retrieval*. New York, USA: Cambridge University Press.

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, USA: McGraw-Hill, Inc.

Stokes, N., Li, Y., Cavedon, L., & Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval, 12*, 17–50.

Voorhees, E. M. (1994). Query expansion using lexical-semantic relations. In Proceedings of the 17th ACM-SIGIR conference, pp. 61–69.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: compressing and indexing documents and images*. San Francisco, California: Morgan Kaufmann.

Zhou, W., Yu, C., Smalheiser, N., Torvik, V., & Hong, J. (2007). Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: Proceedings of the 30th ACM-SIGIR conference, pp. 655–662.