DOI: 10.1111/j.1471-1842.2008.00815.x

Future perspectives

Ubiquitous but unfinished: grand challenges for information retrieval

William Hersh, *Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon, USA*

Over the past 25 years information retrieval (IR) in biomedicine has seen a growth from a task performed by an elite cadre of search experts to something done by almost all literate people. A large segment of the population in the developed world, and a growing fraction in the developing world, use search engines. Of those who search, a substantial majority use them for searching on health-related topics. The biomedical literature and a growing array of other sources are available for free or accessible through institutional subscriptions. Searching is an indispensable skill for the modern health-care professional or researcher. Yet there are still challenges for users and developers of these nearly ubiquitous biomedical IR systems and, as such, the task for those of us who develop systems, provide expertise in their use, and point research in future directions is still unfinished.

Search is essentially ubiquitous. An estimated 80% of consumers who have used a search engine have searched for information related to personal health.[1,2] Likewise, most clinicians are connected. About 98% of US physicians use the Internet and half use PDAs,[3] with higher use among those who see more patients.[4]

What are the grand challenges in biomedical IR as we near close to the end of the first decade of the 21st century? I see four of them

- Content—the right information for the right task;
- Indexing—metadata for Web content;
- Linkage—across multiple resources; and
- Access—open access but protective of intellectual property.

Correspondence: William Hersh, Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University. E-mail: hersh@ohsu.edu

Some may be surprised that searching itself is not on this list. Clearly searching is important, and no doubt we could use improvement in search systems. Each of the four challenges above needs good searching capabilities. But in my view, search proper is not the major challenge.

Content

In seeking content, it is essential that we provide people with the right information for their task. Part of that challenge is to understand the role(s) that individuals play and their information needs. We also need to understand the type of information they need, its extent, and the form in which it is to be delivered. Table 1 lists the dimensions that we must take into account when developing systems for users.

We also need to recognize some of the challenges of producing usable content. Clearly we must understand what users need, which requires that we understand the questions they bring to systems.[5–7] We must also recognize that mass production of information is challenging for a variety of reasons. Not only are the academic rewards for tasks other than knowledge production not clear, but pure volunteerism has not proven scalable. For example, the Cochrane Collaboration has had difficulty reaching its original goals that were based on a model of volunteerism. Likewise, the challenges of the business model will be described below in the Linkage section.

Indexing

Although we produce the quantity and type of content that is necessary to meet the needs of users, we still need to index it properly so that users can find it accurately and efficiently. Although the

Table 1 The dimensions of information content in information retrieval systems

Role	Type	Extent	Form
Personal health—consumers, patients	Patient-specific information—generated in the care of patients	Studies—reports of original research	Bibliographic—pointers (links) to information, rich in metadata
Health care provider—physicians, nurses, others	Knowledge-based information—the scientific literature of health care	Syntheses—systematic synthesis of studies	Full-text—journals, textbooks, Websites, etc.
Population health—public health officials		Synopses—summaries of synthesized knowledge	Annotated—non-text or highly structured text
Information enablers—librarians, information technology personnel		Actionable knowledge—rules and other forms that systems can act upon	Aggregations—bringing all the above together
Researchers—basic, clinical, and translational Policy makers—officials and leaders			

vastness of content makes its complete manual indexing impossible, we must still investigate the methods—human or automated—to provide better indexing.

There is probably still a role for manual indexing of resources such as MEDLINE. In a variety of places, Medical Subject Headings (MeSH) indexing has shown value. A number of research studies have demonstrated it improves the performance of different aspects of searching.[9–11] Other systems of metadata also provide value, such as the annotation of genes and their functions using the Gene Ontology (GO).[12] We still do not know how to best annotate non-textual objects, such as images, as demonstrated in results from the ImageCLEF medical retrieval task.[13] If (a big if!) the semantic Web ever becomes reality, some and perhaps a great deal of manual or semi-automated indexing will be required.[14]

Linkage

Consider this scenario: A primary care clinician observing an elderly patient who has hypertension, congestive heart failure, sleep apnea and obesity. She has charted pertinent information in the electronic health record and now wants recommendations from a guideline with overview of supporting evidence. Later, she would like to explore these recommendations in more detail, including reading a systematic review and some original clinical trials it has included. Or she may want a basic review of topics seen infrequently in practice.

There are many impediments for this clinician.

- She cannot link directly from guideline to supporting or background information.
- She may want to access pertinent sections of a favourite textbook directly.
- She does not want to go to each Website, log on, and use that site's search engine.
- She would also like to navigate across levels of evidence from compendium to systematic review to original clinical trial or other study.
- She may want to create a personal digital library of preferred content.

There are impediments for others as well. Publishers might desire to allow access to pieces of content but need assurances of revenue and intellectual property protection. Content aggregators may want to 'mix and match' content that is 'best of breed' but difficult to do across content of different publishers.

A recurrent problem in these scenarios is that most information is in silos. The way to overcome this impediment is through inter-operability, which is defined by the IEEE[15] as the 'ability of two or more systems ... to exchange information or use the information that has been exchanged.' The term is used in the digital library community to describe seamless access and integration to

content across publishers, Websites, etc. Required to facilitate IR inter-operability are a minimum set of metadata and inter-application interfaces and cooperation among publishers, vendors, and others to agree upon standards.

How can we achieve a move from silos to inter-operability? A possible starting point is the Open Archives Initiative (OAI, www.openarchives.org). OAI promotes the 'exposure' of archives' metadata such that systems can know what content is available and how it can be harvested. Each record in an OAI collection contains metadata. Are there any good examples of integrated resources? Yes, from the bioinformatics community are the databases of National Center for Biotechnology Information (NCBI), which are linked in the Entrez system and to other databases external to it.[16]

Access

A final problem to solve is access to information. The impediments to its wider dissemination are economic and political, not technical. Journals have monopolies due to promotion and tenure concerns. There is a growing concern over the cost of journals in an era of constrained library budgets and the shift from paper to electronic access, where users no longer get to keep their back issues. This has led to calls for 'open access' to scientific research results.[17]

The rationale for these calls is that most research is publicly funded, yet reports of results are copyrighted by publishers. If such information may be life-saving (in the case of biomedical research), should it be freely available? Furthermore, freely available articles are more likely to be cited.[18,19] But the production of information is not free. Furthermore, although authors are sympathetic, they have higher concerns than free access, that is, publication in prestigious journals.[20]

Conclusions

IR systems, especially in biomedicine, have become 'mainstream'. Searching is an essential skill for knowledge workers and perhaps the rest of the world as well. Basic searching is simple and easy to do. Challenges remain in creating and providing access to the right information for the right task while preserving the incentive to produce it.

Conflicts of interest

WH has declared no conflicts.

References

1 Fox, S. *Online Health Search 2006*. Washington, DC: Pew Internet & American Life Project, 2006. Available at: http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf. Accessed 18 June 2008.

2 Anonymous. *Number of 'Cyberchondriacs'—Adults Who Have Ever Gone Online for Health Information—Increases to an Estimated 160 Million Nationwide*. Rochester, NY: Harris Interactive, 2007. Available at: http://www.harrisinteractive.com/harris_poll/index.asp?PID=792. Accessed 18 June 2008.

3 Anonymous. *Physician Internet Use Statistics*. 2005. Available at: http://www.max.md/pdf/PhysicianInternetUseStatistics.pdf. Accessed 18 June 2008.

4 Taylor, H. & Leitman, R. *The Increasing Impact of eHealth on Physician Behavior*, November 13, 2001 Available at: http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2001Vol1_iss31.pdf. Accessed 18 June 2008.

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

6 Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L. & Evans, E. R. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal* 1999, **319**, 358–361.

7 Roberts, P. M. & Hayes, W. S. *Information Needs and the Role of Text Mining in Drug Development. Pacific Symposium on Biocomputing*. Big Island, Hawaii: World Scientific Press, 2008: 592–603. Available at: http://psb.stanford.edu/psb-online/proceedings/psb08/roberts.pdf. Accessed 18 June 2008.

8 Grimshaw, J. So what has the Cochrane Collaboration ever done for us? A report card on the first 10 years. *Canadian Medical Association Journal* 2004, **171**, 747–749.

9 Hersh, W. R., *et al*. OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland: Springer-Verlag, 1994: 192–201.

10 Srinivasan, P. Query expansion and MEDLINE. *Information Processing and Management* 1996, **32**, 431–444.

11 Hersh, W. R., *et al*. Enhancing access to the bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration* 2006, **1**, 3. Available at: http://www.j-biomed-discovery.com/content/1/1/3. Accessed 18 June 2008.

12 Anonymous. *The Gene Ontology Project in 2008. Nucleic Acids Research* 2008, **36**, D440–D444.
13 Hersh, W. R., *et al*. Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association* 2006, **13**, 488–496.
14 Robu, I., Robu, V. & Thirion, B. An introduction to the Semantic Web for health sciences librarians. *Journal of the Medical Library Association* 2006, **94**, 198–205.
15 Anonymous. *IEEE Standard Computer Dictionary: a Compilation of IEEE Standard Computer Glossaries*. Piscataway, NJ: IEEE Press, 1990.
16 Wheeler, D. L., *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 2008, **36**, D13–D21.
17 Albert, K. M. Open access: implications for scholarly publishing and medical libraries. *Journal of the Medical Library Association* 2006, **94**, 253–262.
18 Eysenbach, G. Citation advantage of open access articles. *PLoS Biology* 2006, **4**, e157.
19 Moed, H. F. The effect of open access on citation impact: an analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science & Technology* 2007, **58**, 2047–2054.
20 Schroter, S. & Tite, L. Open access publishing and author-pays business models: a survey of authors' knowledge and perceptions. *Journal of the Royal Society of Medicine* 2006, **99**, 141–148.

