Information Retrieval in the Ubiquitous Search Era: 
A View from the Biomedical/Health Domain

William Hersh, MD 
Professor and Chair 
Department of Medical Informatics & Clinical Epidemiology 
School of Medicine 
Oregon Health & Science University 
Email: hersh@ohsu.edu 
Web: www.billhersh.info 
Blog: http://informaticsprofessor.blogspot.com

References


Information Retrieval in the Ubiquitous Search Era: View from the Biomedical/Health Domain

William Hersh, MD
Professor and Chair
Department of Medical Informatics & Clinical Epidemiology
School of Medicine
Oregon Health & Science University
Email: hersh@ohsu.edu
Web: www.billhersh.info
Blog: http://informaticsprofessor.blogspot.com

Overview

• Role of IR in health and biomedicine
• Personal journey: IR evaluation in health and biomedicine
  – Early work
  – Task-oriented evaluation
  – Use case-driven batch evaluation
• Future directions and recommendations
The world of IR has changed

- Evolution of my book
  - In first edition (1996), last chapter devoted to “special topic” of the Internet and Web
- Most people have used a search engine
  - And have strong opinions about them
- Previous concern of access to information (e.g., Gregor Mendel) has given way to information overload, data smog, and information chaos
- 91% of US Internet users (73% of US adults) have used a search engine (Purcell, 2012)

IR and online access firmly planted in health and biomedicine

- Biology is now defined as an “information science” (Insel, 2003)
- Pharmaceutical companies compete for informatics/library talent (Davies, 2006)
- Search for health information by clinicians, researchers, and patients/consumers is ubiquitous (Purcell, 2012; Google/Manhattan Research, 2012)
  - It’s even part of “meaningful use” rule for electronic health record adoption! (Metzger, 2012)
Popular IR-related icons permeate our lives

Models show us that IR in biomedicine is more than just searching

- Medical decision-making
- Knowledge management
Medical decision-making (Mulrow, 1997)

- EVIDENCE
  - Patient data
  - Basic, clinical, and epidemiological research
  - Randomized controlled trials
  - Systematic reviews

- KNOWLEDGE
  - Patient data
  - Cultural beliefs
  - Personal values

- CLINICAL DECISION

- PATIENT/CLINICIAN PREFERENCES
  - Formal policies and laws
  - Community standards
  - Time
  - Financial

- GUIDELINES

- ETHICS

- CONSTRAINTS

IR in context of biomedical knowledge management (Hersh, 2009)

- All literature
  - Possibly relevant literature (abstracts)
  - Definitely relevant literature (full text)
  - Actionable knowledge

- Information retrieval
  - Information extraction, text mining
Personal journey in IR evaluation in health and biomedical domain

- SAPHIRE
- Toward task-oriented evaluation
- Factors association with successful searching
- Domain-specific retrieval evaluation

Concept-based IR using UMLS Metathesaurus (Hersh, 1990)
Set out to evaluate SAPHIRE and IR in biomedicine

• Concept-based approach did not impart value over word indexing and searching (Hersh, JAMIA, 1994)
• Experience of several evaluations led to concern with use of recall/precision (Hersh, JASIS, 1994)
  – How much difference is meaningful?
  – How valid is batch evaluation for understand how well user will search?

Led to “task-oriented” evaluation approaches

• Motivated by Egan (1989) and Mynatt (1992)
• Major task in medicine: answering questions
• How can we evaluate systems in interactive use for answering questions?
• Undertook parallel approaches in
  – Medicine – Using bibliographic databases and electronic textbooks
  – General news – TREC Interactive Track
Medical textbook – Boolean vs. natural language (1995)

• Searching medical textbook (*Scientific American Medicine*) with Boolean and natural language interfaces
  – Medical students answering ten short-answer questions
  – Randomized to one interface or other, asked to search on questions they rated lowest confidence before searching
  – Pre-searching correctness very low (1.7/10)
  – Correctness improved markedly with searching (4.0/5)
  – When incorrect with searching, document with correct answer retrieved two-thirds of time and viewed half of time

MEDLINE – Boolean vs. natural language (1996)

• Searching MEDLINE with Ovid (Boolean) and Knowledge Finder (natural language)
  – Medical students answering yes/no clinical questions
  – 37.5% answered correctly before searching
  – 85.4% answered correctly after searching
  – No difference across systems in time taken, relevant articles retrieved, or user satisfaction
Factors associated with successful searching (Hersh, 2002)

- Medical and nurse practitioner (NP) students success of using a retrieval system to answer clinical questions
  - Had to provide not only answer but level of evidence supporting it
    - Yes with good evidence
    - Indeterminate evidence
    - No with good evidence
- Look at factors associated with success
  - Based on model of factors associated with successful use of retrieval systems (Fidel, 1983) adapted to this setting
  - Dependent variable was correctness of answer

Major categories and some factors in the model

- Associated answering question correctly with independent variables
  - Answers – correct before searching, certainty, time
  - Demographic – age, gender, school
  - Computer experience – general, searching, specific MEDLINE features
  - Cognitive – set of factors shown in past to be associated with successful computer and/or retrieval system use
  - Search mechanics – sets retrieved, references viewed
  - User satisfaction – from Questionnaire for User Interface Satisfaction (QUIS)
  - Retrieval – recall, precision
Results

• 66 participants, 45 medical and 21 NP students
  – NP students all female, medical students evenly divided
  – NP students older, with more computer use but less searching and EBM experience
  – Medical students scored higher on cognitive tests, especially of spatial visualization
• Prior to searching, rate of correctness (32.1%) about equal to chance for both groups
  – Rating of certainly low for both groups
• With searching, medical students increased rate of correctness to 51.6% but NP students remained virtually unchanged at 34.7%
  – NP student difference was likely due to judging evidence

Results (cont.)

<table>
<thead>
<tr>
<th>Pre-Search</th>
<th>Post-Search</th>
<th>Incorrect</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>133 (41%)</td>
<td>81 (36%)</td>
<td>70 (31%)</td>
</tr>
<tr>
<td>Correct</td>
<td>41 (13%)</td>
<td>27 (12%)</td>
<td>45 (20%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Incorrect</th>
<th>Correct</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>18%</td>
<td>18%</td>
<td>.61</td>
</tr>
<tr>
<td>Precision</td>
<td>28%</td>
<td>29%</td>
<td>.99</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>All</th>
<th>Medical</th>
<th>NP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>18%</td>
<td>18%</td>
<td>20%</td>
</tr>
<tr>
<td>Precision</td>
<td>29%</td>
<td>30%</td>
<td>26%</td>
</tr>
</tbody>
</table>
Work followed on by others

- Physicians and nurse consultants searching full-text and MEDLINE resource – both improved with searching (Westbrook, 2005)
- Physicians using self-chosen resource improved minimally (McKibbon, 2006)
- Physician searching improved more with textbook than Google or MEDLINE (Thiele, 2010)
- Physicians had modest improvement with searching, no difference between Pubmed and Clinical Queries (McKibbon, 2013)

Including study of non-clinicians

  - Correct answering 61.2% before searching and 82.0% after
  - Confidence not associated with correctness
- Van Duersen (2012) – older and less educated searchers have lower search skills although younger searchers more likely to use nonrelevant search results and unreliable sources
- Taylor (2012) – same attributes of younger ("millenial generation") searchers seen in general
Back to batch evaluation: domain-specific IR

- TREC Genomics Track
- ImageCLEFmed
- TREC Medical Records Track

TREC Genomics Track (Hersh, 2009)

- Based on use case of exploding research in genomics and inability to biologists to know all that might impact work
- First TREC track devoted to “domain-specific” retrieval, with focus on IR systems for genomics researchers
- History
  - 2004-2005 – focus on ad hoc retrieval and document categorization
  - 2006-2007 – focus on passage retrieval and question-answering as means to improve document retrieval
Lessons learned (Hersh, 2009)

• Ad hoc retrieval
  – Modest benefit for techniques known to work well in general IR, e.g., stop word removal, stemming, weighting
  – Query term expansion, especially domain-specific and/or done by humans, helped most

• QA
  – Most consistent benefit from query expansion and paragraph-length passage retrieval

• For all experiments, big problem (as always) was lack of detailed description and use of low-performing baselines

Image retrieval – ImageCLEF medical image retrieval task

• Biomedical professionals increasingly use images for research, clinical care, and education, yet we know very little about how they find them

• Developed test collection and exploration of information needs motivating use of image retrieval systems (Hersh, 2006; Hersh, 2009; Müller, 2010)

• Started with ad hoc retrieval and added tasks
  – Modality detection
  – Case finding
TREC Medical Records Track

• Adapting IR techniques to medical records
• Use case somewhat different – want to retrieve records and data within them to identify patients who might be candidates for clinical studies
• Motivated by larger desire for “secondary use” of clinical data (Safran, 2007)
• Opportunities facilitated by growing incentives for “meaningful use” of EHRs in the HITECH Act (Blumenthal, 2011; Blumenthal, 2011), aiming toward the “learning healthcare system” (Friedman, 2010; Smith 2012)

Challenges for secondary use of clinical data

• EHR data does not automatically lead to knowledge (Hersh, 2013; Hersh, 2013)
  – Data quality and accuracy is not a top priority for busy clinicians
  – Patients get care in many places, so record may be incomplete
  – Data provenance often a concern; where does data come from?
  – Best evidence for medical tests and treatments comes from experiments, i.e., evidence-based medicine
Challenges for informatics research with medical records

• Has always been easier with knowledge-based content than patient-specific data due to a variety of reasons
  – Privacy issues
  – Task issues
• Facilitated with development of large-scale, de-identified data set from University of Pittsburgh Medical Center (UPMC)
• Launched in 2011, repeated in 2012 (Voorhees, 2012)
Some issues for test collection

- De-identified to remove protected health information (PHI), e.g., age number → range
- De-identification precludes linkage of same patient across different visits (encounters)
- UPMC only authorized use for TREC 2011 and TREC 2012 but nothing else, including any other research (unless approved by UPMC)

Easy and hard topics

- Easiest – best median bpref
  - 105: Patients with dementia
  - 132: Patients admitted for surgery of the cervical spine for fusion or discectomy
- Hardest – worst best bpref and worst median bpref
  - 108: Patients treated for vascular claudication surgically
  - 124: Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma
- Large differences between best and median bpref
  - 125: Patients co-infected with Hepatitis C and HIV
  - 103: Hospitalized patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis
  - 111: Patients with chronic back pain who receive an intraspinal pain-medication pump
Failure analysis for 2011 topics (Edinger, 2012)

<table>
<thead>
<tr>
<th>Reasons for Incorrect Retrieval</th>
<th>Number of Visits</th>
<th>Number of Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visits Judged Not Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic terms mentioned as future possibility</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Topic symptom/condition/procedure done in the past</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>All topic criteria present but not in the time/sequence</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>Most, but not all, required topic criteria present</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Topic terms denied or ruled out</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>Notes contain very similar term confused with topic term</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>Non-relevant reference in record to topic terms</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>Topic terms not present—unclear why record was ranked high</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Topic present—record is relevant—disagree with expert judgment</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>Visits Judged Relevant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic not present—record is not relevant—disagree with expert judgment</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Topic present in record but overlooked in search</td>
<td>103</td>
<td>27</td>
</tr>
<tr>
<td>Visit notes used a synonym or lexical variant for topic terms</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Topic terms not named in notes and must be inferred</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Topic terms present in diagnosis list but not visit notes</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Results for 2012

<table>
<thead>
<tr>
<th>Run</th>
<th>mNDCG</th>
<th>mAP</th>
<th>P(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLMmmr1q1</td>
<td>0.690</td>
<td>0.386</td>
<td>0.739</td>
</tr>
<tr>
<td>uDE1SUM</td>
<td>0.578</td>
<td>0.286</td>
<td>0.592</td>
</tr>
<tr>
<td>unstream2</td>
<td>0.547</td>
<td>0.275</td>
<td>0.557</td>
</tr>
<tr>
<td>oshuMdata10a</td>
<td>0.526</td>
<td>0.250</td>
<td>0.611</td>
</tr>
<tr>
<td>itgpe1</td>
<td>0.524</td>
<td>0.224</td>
<td>0.519</td>
</tr>
<tr>
<td>UDAutoMed123</td>
<td>0.517</td>
<td>0.236</td>
<td>0.528</td>
</tr>
<tr>
<td>wogDrMedQR6</td>
<td>0.509</td>
<td>0.231</td>
<td>0.553</td>
</tr>
<tr>
<td>NICTAUC05</td>
<td>0.487</td>
<td>0.216</td>
<td>0.517</td>
</tr>
</tbody>
</table>
What approaches did (and did not) work?

• Best results in 2011 and 2012 obtained from NLM group (Demner-Fushman, 2011)
  – Top results from manually constructed queries using Essie domain-specific search engine (Ide, 2007)
  – Other automated processes fared less well, e.g., creation of PICO frames, negation, term expansion, etc.
• Best automated results in 2011 obtained by Cengage (King, 2011)
  – Filtered by age, race, gender, admission status; terms expanded by UMLS Metathesaurus
• Benefits of approaches commonly successful in IR provided small or inconsistent value
  – Document focusing, term expansion, etc.

Conclusions and future directions

• Evaluation must focus on real-world
  – Use cases
  – Collections and topics
• Use cases should focus on tasks of clinicians, researchers, and other specific roles
• Collections should reflect type and quantity of information appropriate to use cases