Information Retrieval (Search)

What is Biomedical & Health Informatics?
William Hersh
Copyright 2019
Oregon Health & Science University

Information retrieval (IR)

- Field concerned with organization and retrieval of predominantly text-based information
  - But multimedia (e.g., images, sounds, video, etc.) and more complex databases are increasingly a part
- When I began work in this area (circa 1989), few physicians or scientists and virtually no patients had done an on-line search
  - Now everyone is searching!

8.1   2

1

More history: Who said the following and when?

- "It has become increasingly difficult to keep abreast of and to assimilate the investigative reports which accumulate day after day. My friend … was ill at ease because he felt unable to control even the area of his own discipline; one suffocates, he once told me, through exposure to the massive body of rapidly growing information."

Answer

- The previous was said by Bernhard von Langenbeck, a German surgeon, in 1872
- Another insightful quote comes from Herbert Simon (1971), an early artificial intelligence (AI) researcher,
  - "What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."
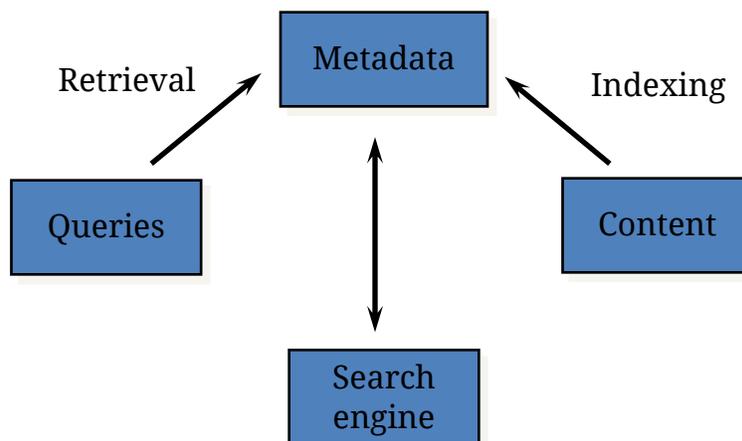
IR process and field

- Overview of IR process
- Field of IR
- Pertinence of IR to health and biomedicine

IR system

Retrieval

Metadata

Indexing

Queries

Content

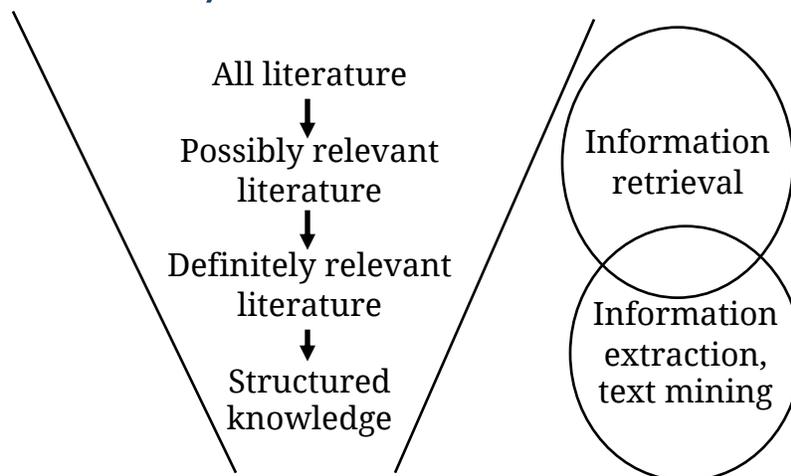Search engine

The intellectual tasks of IR

- Indexing
  - Assigning metadata to content items
  - Can assign
    - Subjects (terms) – words, phrases from controlled vocabulary
    - Attributes – e.g., author, source, publication type
- Retrieval
  - Most common approaches are
    - Boolean – use of AND, OR, NOT
    - Natural language – words common to query and content

8.1   7

IR also a growing part of "knowledge discovery" from scientific literature

All literature

↓

Possibly relevant literature

↓

Definitely relevant literature

↓

Structured knowledge

Information retrieval

Information extraction, text mining

8.1   8

Major challenges in IR

- We have gone from information paucity to information overload
- Many topics we want to search on have multiple ways to be expressed
  - e.g., diseases, genes, symptoms, etc.
- The converse is a problem too: Many words and terms used to express topics have multiple meanings
- Balancing open access vs. providing for cost of production and maintenance

8.1   9

IR is now "mainstream"

- Internet (and likely search engine) use is now ubiquitous
  - Not only in developed countries (Perrin, 2015) but across world – http://www.internetworldstats.com/stats.htm
- 71% of Internet users (59% of US adults) have searched for health information, with 35% using it for self-diagnosis (Fox, 2013)
- "Search engine optimization" (SEO) is a key function used by many companies and organizations (Moz, 2015)
  - https://moz.com/beginners-guide-to-seo
  - Some are lucky, e.g., last name of "Hersh"

Daily 73%

Less than daily 13%

| 21 | 42 | 10 | 6 | 7 |

Almost constantly — Several times a day — About once a day — Several times a week — Less often

WORLD INTERNET USAGE AND POPULATION STATISTICS
MARCH, 2019 - Updated

World Regions	Population (2019 Est.)	Population % of World	Internet Users 31 Mar 2019	Penetration Rate (% Pop.)	Growth 2000-2019	Internet Users %
Africa	1,320,038,716	17.1 %	492,762,185	37.3 %	10,815 %	11.2 %
Asia	4,241,972,790	55.0 %	2,197,444,783	51.8 %	1,822 %	50.1 %
Europe	829,173,007	10.7 %	719,365,521	86.8 %	584 %	16.4 %
Latin America / Caribbean	658,345,826	8.5 %	444,493,379	67.5 %	2,360 %	10.1 %
Middle East	258,356,867	3.3 %	173,542,069	67.2 %	5,183 %	4.0 %
North America	366,496,802	4.7 %	327,568,127	89.4 %	203 %	7.5 %
Oceania / Australia	41,839,201	0.5 %	28,634,278	68.4 %	276 %	0.7 %
WORLD TOTAL	7,716,223,209	100.0 %	4,383,810,342	56.8 %	1,114 %	100.0 %

8.1   10

The Web has changed the nature of search

- Three major uses (Broder, 2002)
  - Informational – seeking information (39-48%)
  - Navigational – looking for a specific page, e.g., a home page (20-24%)
  - Transactional – perform transactions, e.g., on-line purchasing (30-36%)
- We are in the era of "adversarial" search – there is content we do not want to retrieve (Castillo, 2011; Smith, 2014)
  - Some of the content we might not want to retrieve is "fake news," which came to the fore in 2016 (Holan, 2016)
- Growing privacy concerns about tracking our searching (Huesch, 2013; Libert, 2015)

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)
- Clinicians cannot keep up – average of 75 clinical trials and 11 systematic reviews published each day (Bastian, 2010)
- 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" – text search over electronic health record notes (Metzger, 2012)

Use is ubiquitous among physicians (Google/Manhattan Research, 2012)

- Most have multiple devices – 99% with a desktop or laptop, 84% with a smartphone, and 54% with a tablet
- Spend twice as much time using online resources as print resources
- Even physicians aged 55+ heavy users – 80% own a smartphone, 84% use search engines daily, and 9 hours per week is spent online for professional purposes
- Search engine use a daily activity – 84%, with average of six searches done per day and 94% using Google
- When looking for clinical or treatment information, about a third click first on sponsored listings from a search
- About 93% say they take action based on searching – everything from pursuing more information to sharing with a patient or colleague to changing treatment decisions
- On smartphones, searching is preferred over mobile apps – 48% of use time with a search engine, 34% with mobile apps, and 18% going to specific Web sites in a browser or with a bookmark
- Spend about 6 hours per week watching online video, with about half of that time spent for professional purposes

8.1   13

What kind of health information do consumers search for? (Fox, 2011)

| Health topic	% searching
Specific disease or medical problem	66%
Certain medical treatment or procedure	56%
Doctors or other health professionals	44%
Hospitals or other medical facilities	36%
Health insurance – private or government	33%
Food safety or recalls	29%
Environmental health hazards	22%
Pregnancy and childbirth	19%
Medical test results	16%

8.1   14

7

How to find more information about IR in health and biomedicine

- From me!
- Hersh WR, *Information Retrieval: A Health and Biomedical Perspective*, Third Edition, 2009
  - Web site: www.irbook.info
- Chapters in other books, e.g., Shortliffe (2014), Sanchez-Mendiola (2014)
- OHSU BMI 514 – Information Retrieval
- Plenty of other books, journals, and other sources

William Hersh

Information Retrieval

A Health and Biomedical Perspective

Third Edition

HEALTH INFORMATICS SERIES
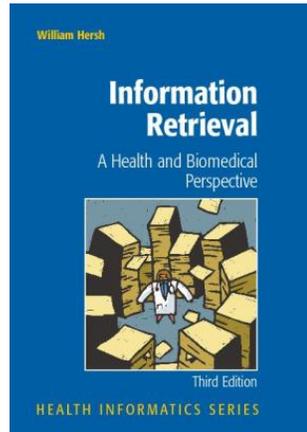
8.1   15

Why is IR pertinent to health and biomedicine?

- Growth of knowledge has long surpassed human memory capabilities
- Clinicians have frequent and unmet information needs
- Researchers must frequently update their knowledge in new areas quickly
- Primary literature on a given topic can be scattered and hard to synthesize
- Non-primary literature sources are often neither comprehensive nor systematic
- Web is increasingly used as source of health and biomedical information

8.1   16