Information Retrieval Systems

Dr. Adrian Mondry
and
Dong Peng
Evolution of Medical Information Retrieval

Information retrieval is the science and practice of identification and efficient use of recorded media.

Initially restricted to biomedical literature, it now includes databases of images, patient data etc.
1879: *Index Medicus* by Dr John Shaw Billings

Journal articles were indexed by author name and subject headings and aggregated into bound volumes.

1966: the Medical Literature Analysis and Retrieval System (*MEDLARS*) by the National Library of Medicine (NLM)

- Predecessor of *Medline*.
- Stores only abstracted information from each article, e.g. author names, article title, journal source, publication date due to limited computing power and disk storage.
- Assign to each article a number of terms from its MeSH thesaurus.

1980s: full-text databases emerges due to growing computing power and less cheaper disk storage.

Early 1990s: With the advent of the World Wide Web and the exponentially increasing power of computers and networks, vast quantities of medical information from multiple sources were available over the global internet.
The Information-Retrieval Process


- Indexing
- Query formulation
- Retrieval
- Evaluation and Refinement
Content

Can be defined as media developed to convey information or knowledge.

Can be original (I.e. journal papers, conference proceedings etc.) or synoptic (I.e. textbook, review article, etc).

Bibliographic content is abstracted from full text.

Meta-analysis of research: combines findings from original texts
Objective: produce smallest, most efficient representation of original content.

Indexing of bibliographic content

- **Medline** is the world’s premiere bibliographic information source developed by NLM. It contains references to 10 million biomedical journal articles published in 3,500 journals since 1966.

- Multiple-indexed source:
  - Information abstracted from the publication, e.g. authors’ names, article title, publication date.
  - Information added by a human indexer, e.g. subject headings (MeSH), publication type.

  - **MeSH**
    - Developed by NLM to represent important concepts in biomedicine.
    - A collection of 18,000 subheadings grouped into 16 trees.
Indexing

Full text indexing
No terms are assigned by human indexer.
Term frequency and term locations are used in the indexing method.
The most common method *Automated Indexing (vector-space model)* was pioneered by Salton in the 60s but only achieved widespread use in the 90s.

Document preprocessing:
1. extract unique words.
2. eliminate non-content-bearing “stop words”, e.g. a, an, the
3. stem word to their roots.
4. count the number of occurrence of each word in each document.

Assigning weight to each word to enhance discrimination between various document vectors.

\[
weight_{ij} = TF_{ij} \cdot IDF_i
\]

Inverse Document Frequency
Term Frequency

\[
IDF_i = \log\left(\frac{\text{# of documents}}{\text{# of documents with term } i}\right) + 1
\]

\[
TF_{ij} = \log\left(\text{frequency of term } i \text{ in document } j\right) + 1
\]
Query Formulation

Objective: to translate an information need into a high quality query.

Boolean Queries
- Boolean operators: AND, OR, NOT.
- Wildcard characters: # (end with), : (start with).
- Field qualification for multiple indexed databases (e.g. author, subject).
- Term explosion is allowed by most Medline systems (e.g. Hypertension, Malignant).
- Text-word searching is allowed by modern Medline systems, i.e. words searched through the title and the abstract.

Natural Language Queries
- Syntactic: removal of stop words, identification of common phrases, stemming of words to their roots.
- Semantic: identifying a word’s semantic type (e.g. diagnosis, treatment), expanding the query to include synonyms of the entered terms, identifying indexes in which individual words should be searched.
Retrieval

Objective: to compare queries against index to create results.

Matching:
Queries are compared against index and a result set is created.

Ranking:
Chronological
Alphabetic
Relevance ranking: common in natural language IR systems

Display:
usually, contents rather than items or attributes is displayed
Evaluation

• Focusing on measuring the retrieval of relevant documents.
• Two parameters: recall and precision.

\[
\text{recall} = \frac{\text{# documents retrieved and relevant}}{\text{# relevant documents in database}}
\]

\[
\text{precision} = \frac{\text{# documents retrieved and relevant}}{\text{# number of documents retrieved}}
\]

• Why low recall and precision?
  – From indexing
    • Inconsistency in human indexers.
    • Problems with word-based indexing: context, polysemy, synonymy, content.
  – From retrieval
    • E.g. excess AND in queries

Sensitivity
analogous

Positive predictive value
Trends in Medical Information Retrieval Systems

Aggregation of content from diverse information sources.

Example: MedWeaver
A web-based application that integrates functions from 3 systems:

- **DXplain**: a computer-based decision support system, to provide a summary description of a disease and possible diagnoses when a patient’s clinical findings are entered into the system.
- Assisted literature searches of MEDLINE.
- Links to clinically relevant web sites.

Advantages:

- Allows needs-based rather than source-based system design.
- Content creation and maintenance are distributed across organizations.
MedWeaver
Trends in Medical Information Retrieval Systems

Distributed Hypermedia.

Most texts not peer-reviewed, written for lay people

Suggested elements for web-pages to identify qualifications and disclosure of conflicts of interest: Authorship, Attribution, Disclosure, Currency

Problem of document structure => low retrieval performance

Solutions:

Creation of authoritative sites, backed by qualified institutions, e.g. professional associations, government agencies

Combining content from multiple sources in one service, eg. Medscape, ovid technologies, MD consult

Creation of searchable pointers to medical resources, e.g. medweb, medical matrix, cliniweb...
Trends in Medical Information Retrieval Systems

Common Medical Vocabulary - the Unified Medical Language System.

Many coding schemes in medicine, e.g. MeSH for indexing literature, ICD for diagnosis coding, SNOMED for clinical information coding

Several initiatives around to create common vocabularies, e.g. the NLM’s Unified Medical Language System (UMLS)

ULMS’s main component is the Metathesaurus which “translates” between vocabularies.

Other important component: Information sources map, guiding to databases most likely to provide the output needed.
Online CATs Websites

Crawler

Full-text CAT

Noun Phraser

CAT Descriptor

Indexer

Available CAT Search Engines

Meta-Search Engine

Categorized documents

Matched results

Algorithm - Extension to CATs??
Future Challenges in Medical Information Retrieval

Clinicians need high-quality, trusted information in the delivery of health care. Outdated information needs to be archived dynamically.

Information must be organized and indexed effectively for easy retrieval, to increase recall and precision of information retrieval.

Information must be delivered on a platform that is convenient and reliable. More and more IR system will become web-based.
Questions you should know to answer

What is primary, what is synoptic content?

What are the two major types of indexing and their respective advantages and disadvantages?

What are the benefits and limitations of the internet and WWW platforms for clinical information retrieval systems?