COM6115: Text Processing

Information Retrieval: Document Indexing — Manual

Mark Hepple

Department of Computer Science
University of Sheffield

Overview

- Definition of the information retrieval problem
- Approaches to document indexing
  - manual approaches
  - automatic approaches
- Automated retrieval models
  - boolean model
  - ranked retrieval methods (e.g. vector space model)
- Term manipulation:
  - stemming, stopwords, term weighting
- Web Search Ranking
- Evaluation

Issues in IR

- How can I formulate a query?
  - query type: normally keywords, could be natural language
- How are the documents represented?
  - indexing
- How does the system find the best-matching document?
  - retrieval model
- How does the system find it *efficiently*?
- How are the results presented to me?
  - unsorted list, ranked list, clusters
- How do we know whether the system is any good?
  - evaluation

Indexing

The task of finding terms that describe documents well

- Manual:
  - indexing by humans (using fixed vocabularies)
  - labour and training intensive
- Automatic:
  - Term manipulation (certain words count as the same term)
  - Term weighting (certain terms are more important than others)
  - Index terms must only derive from text

Manual Indexing

- Large vocabularies (several thousand items)
  - Dewey Decimal System
  - Library of Congress Subject Headings
  - ACM – subfields of CS
  - MeSH – Medical Subject Headings

Example: Manual Indexing — ACM

ACM Computing Classification System (1998)
B Hardware
B.3 Memory structures
B.3.0 General
B.3.1 Semiconductor Memories (NEW) (was B.7.1)
Dynamic memory (DRAM) (NEW)
Read-only memory (ROM) (NEW)
Static memory (SRAM) (NEW)
B.3.2 Design Styles (was D.4.2)
Associative memories
Cache memories
Interleaved memories
Mass storage (e.g., magnetic, optical, RAID)
Primary memory
Sequential-access memory

Example: Manual Indexing — MeSH

MeSH — Medical Subject Headings

- a very large *controlled vocabulary* for describing/indexing medical documents, e.g. journal papers and books
- provides a *hierarchy* of descriptors (a.k.a. *subject headings*)
  - assigned to documents to describe their content
- hierarchy has a number of *top-level* categories, e.g.:
  - Anatomy [A]
  - Organisms [B]
  - Diseases [C]
  - Chemicals and Drugs [D]
  - Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
  - Psychiatry and Psychology [F]
  - Biological Sciences [G]

    ...

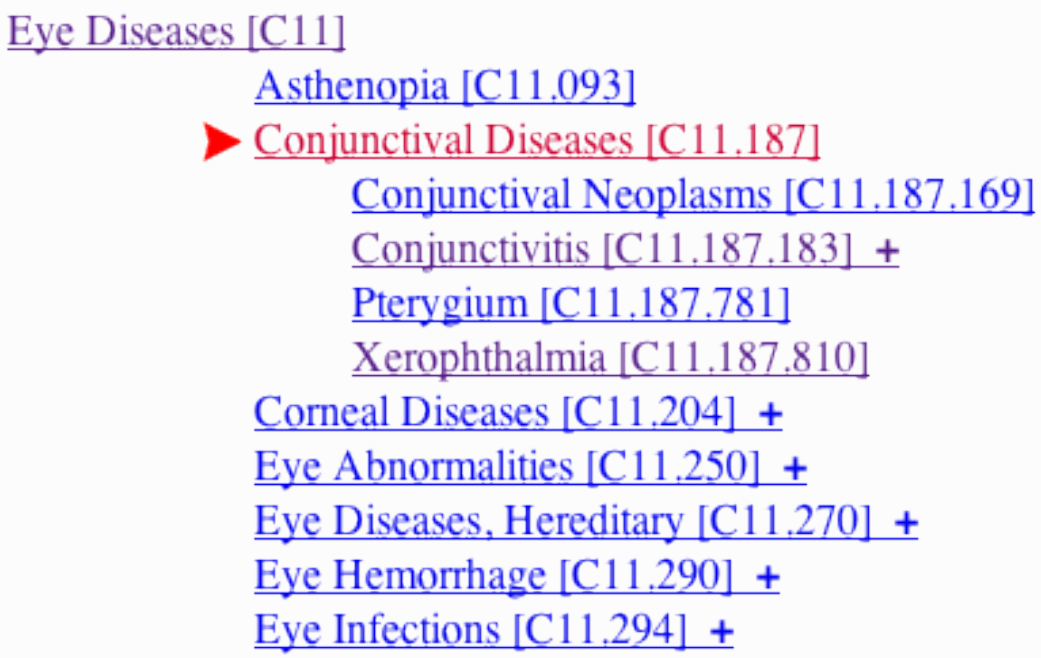
Example: Manual Indexing — MeSH (contd)

- And a number of subcategories (more specific/detailed terms):

  - Diseases [C]
    - **MeSH** *C01* --- bacterial infections and mycoses
    - **MeSH** *C02* --- virus diseases
    - **MeSH** *C03* --- parasitic diseases
    - **MeSH** *C04* --- neoplasms
    - **MeSH** *C05* --- musculoskeletal diseases
    - **MeSH** *C06* --- digestive system diseases
    - **MeSH** *C07* --- stomatognathic diseases
    - **MeSH** *C08* --- respiratory tract diseases
    - **MeSH** *C09* --- otorhinolaryngologic diseases
    - **MeSH** *C10* --- nervous system diseases
    - **MeSH** *C11* --- eye diseases
    - **MeSH** *C12* --- urologic and male genital diseases
    - **MeSH** *C13* --- female genital diseases and pregnancy complications
    - **MeSH** *C14* --- cardiovascular diseases

Example: Manual Indexing — MeSH (contd)

- And a number of subsubcategories (even more specific/detailed terms):



Example: Manual Indexing — MeSH (contd)

- And a number of subsubsubcategories (yet again more specific/ detailed terms):

Eye Diseases [C11]
Conjunctival Diseases [C11.187]
Conjunctival Neoplasms [C11.187.169]
Conjunctivitis [C11.187.183]
Conjunctivitis, Allergic [C11.187.183.200]
Conjunctivitis, Bacterial [C11.187.183.220] +
Conjunctivitis, Viral [C11.187.183.240] +
Keratoconjunctivitis [C11.187.183.394] +
Reiter Syndrome [C11.187.183.749]
Pterygium [C11.187.781]
Xerophthalmia [C11.187.810]

Example: Manual Indexing — MeSH (contd)

- MEDLINE — Medical Literature Analysis and Retrieval System Online
  - ◇ international database of literature for medicine and the life sciences
  - ◇ includes papers from ≈5600 different sources (mostly journals), in various languages
  - ◇ database now holds records for ≈26 million papers
- Each MEDLINE article indexed with 10-15 descriptors from MeSH
  - ◇ papers accessed by PubMed search engine interface, using MeSH terms (and other terms, e.g. author name, etc)
  - ◇ by default, all descriptors below a given one in the hierarchy are also included in search

Manual Indexing

- Advantages:
  - ◇ High precision searches
  - ◇ Works well for closed collections (books in a library)
- Problems:
  - ◇ Searchers need to know terms to achieve high precision
  - ◇ Labellers need to be trained to achieve consistency
    - Not feasible to expect this from all content creators on the web
  - ◇ Collections are dynamic $\rightarrow$ schemes change constantly

Reading

- Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval. New Yorl: ACM Press, 1999.
- C. Manning, P. Raghavan and H. Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.
- I.H. Witten, A. Moffat and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.