Searching Text and Images in the Medical Domain

Allan Hanbury and Henning Müller

Allan Hanbury

- M.Sc. In Physics (University of Cape Town, South Africa)
- Ph.D. In Applied Mathematics (MINES ParisTech, France)
- Habilitation in Informatics (Vienna University of Technology, Austria)
- Senior Researcher at the Vienna University of Technology
- Scientific Coordinator of the Khresmoi project.
Vienna University of Technology

- Austria’s largest technical university
- 27000 students
- Faculty of Informatics
  - Over 1000 new student admissions per year
- Five Research Foci:
  - Computational Intelligence
  - Distributed and Parallel Systems
  - Media Informatics and Visual Computing
  - Computer Engineering
  - Business Informatics

Henning Müller

- Studies of medical informatics in Heidelberg, Germany (1992-97)
- Work at Daimler-Benz research, USA (1997-98)
- PhD in image processing, University of Geneva, Switzerland (1998-2002)
  - Work on artificial intelligence at Monash University, Melbourne, Australia (2001)
- Medical Informatics Service, University and Hospitals of Geneva (2002-)
- HES-SO, Business information system, Sierre (2007-)
- Coordinator of Khresmoi, organizer ImageCLEF
HES-SO Sierre (part of HES-SO)

- 2'000 students
  - Economy, tourism, business informatics
- Institute of business information systems
- Research in focused domains
  - Internet of things, RFID
  - Mobile applications
  - Energy, Green ICT
  - SAP Center
  - eHealth
  - Information retrieval and management

Khresmoi

Language Resources → Images → Books → Websites → Journals → Semantic Data → Information Answers → Queries Questions
Khresmoi partners

Visit the Khresmoi Stand!
## Course Contents

- **Introduction to Information Retrieval**
- Who searches for medical information and how do they search?
- Search in the medical domain
- Improving search in the medical domain (Discussion)
- Searching for medical images
- Who searches medical images and how do they search?
- Combining text and visual search
- Challenges for search in the medical domain (Discussion)
Contents

- Information Retrieval (IR)
- Indexing
- Queries
- Information Retrieval Models
  - Boolean Model
  - Ranking Model
- Advantages and Limitations
- Web Search
Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Key Characteristics:
- Unstructured information
- Separation of indexing and query time processing
- Strong empirical method
IR vs. Databases

- Structured vs. Unstructured Data
  - Structured data tends to refer to information in “tables”

<table>
<thead>
<tr>
<th>Employee</th>
<th>Manager</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>Jones</td>
<td>50000</td>
</tr>
<tr>
<td>Chang</td>
<td>Smith</td>
<td>60000</td>
</tr>
<tr>
<td>Ivy</td>
<td>Smith</td>
<td>50000</td>
</tr>
</tbody>
</table>

Typically allows numerical range and exact match (for text) queries, e.g.,

\[ \text{Salary} < 60000 \text{ AND Manager} = \text{Smith}. \]


Unstructured Information

- Text
- Images
- Music
- Videos

As opposed to

- Relational databases
- Lists of numbers
Semi-structured Data

- In fact almost no data is “unstructured”
- For example:
  - This slide has distinctly identified zones such as the *Title* and *Bullets*
  - Journal articles contain *Title, Abstract, Authors, …* sections
- Facilitates “semi-structured” search such as
  - *Title* contains data AND *Bullets* contain search


Separation of Indexing & Query Time

- IR is about large scale data collections
- The collection of information cannot be searched directly in interactive time
- Therefore we need to separate the process into:
  1. Offline (crawl/index) time processing
  2. Online query time processing
Empirical Method

- Need to show whether one system is better than another
- Better systems produce more relevant information
- We need reproducibility
- **Evaluation** is required
- Key evaluation measures:
  - Precision
  - Recall

---

Precision and Recall

- A query returns \( n \) ranked documents from a database of many.
- Each one is judged as relevant or not:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>3</td>
<td>NO</td>
</tr>
<tr>
<td>4</td>
<td>YES</td>
</tr>
<tr>
<td>5</td>
<td>NO</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>( n )</td>
<td>NO</td>
</tr>
</tbody>
</table>
Precision and Recall Concepts

- Precision = \[
\frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}\]
- Recall = \[
\frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}}\]

Retrieval Effectiveness

- **Precision**
  - How happy are we with what we’ve got?
  
  \[
  \text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of documents retrieved}}
  \]

- **Recall**
  - How much more we could have had?
  
  \[
  \text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}}
  \]
Search to the People!

- The Internet has democratised search
- Before the Web, computerised IR was usually done by specialised users, such as librarians and journalists
- The Internet is now accessed by 75% of the US adult population. 91% of those who use the Internet use Web search engines (Pew Internet survey 2008)

Conceptual Model for Search

- Documents
- Information Need
- Indexing
- Formulation
- Document Representation
- Query
- Retrieval Function
- Retrieved Documents
- Relevance Feedback, Query Reformulation, Query Expansion
- Further Analysis of the Documents
How an IR system DOES NOT work:
- The user types in a query
- Then the system scans through all documents and returns those that match the query

This would not allow rapid searching
- For this reason, the system first runs an indexing stage before any querying can be done
Aim of Indexing

- Storage of information in a way that supports efficient retrieval
- Two main points of consideration:
  - Accuracy of representation
  - Space and time efficiency
- The basic indexing process is pretty much the same for all search engines

Overview of Indexing Process

- Basic Concept

  I like to laugh. It is a tonic. It braces me up—makes me feel fine—and keeps me in prime mental condition. Laughter is a physiological necessity. The nerve system requires it. The deep, forceful chest movement in itself sets the blood to racing thereby livening up the circulation—which is good for us.

  Without a word, Mr. Stevens caught up the tray from the piano and glided away on his toe-points; whereupon Mr. Brimberly (being alone) became astonishingly agile and nimble all at once, diving down to straighten a rug here and there, rearranging chairs and tables; he even opened the window and hurled two half-smoked cigars far out into the night;

  It was always night on Martha, but Mark broke up his time into mornings, afternoons and evenings. Their life followed a simple routine. Breakfast, from vegetables and Mark's canned store. Then the robot would work in the fields, and the plants grew used to his touch.

  The untiring efforts of genius for over a century have succeeded in producing a musical instrument that falls little short of perfection. The whole edifice bears the same warm tinge of yellow that all those of good quality acquire from age in that pure climate.

I like to laugh. It is a tonic. It braces me up—makes me feel fine!—and keeps me in prime mental condition.

Laughter is a physiological necessity. The nerve system requires it. The deep, forceful chest movement in itself sets the blood to racing thereby livening up the circulation—which is good for us.

The whole edifice bears the same warm tinge of yellow that all those of good quality acquire from age in that pure climate.

The untiring efforts of genius for over a century have succeeded in producing a musical instrument that falls little short of perfection.
I like to laugh. It is a tonic. It braces me up—makes me feel fine!—and keeps me in prime mental condition. Laughter is a physiological necessity. The nerve system requires it. The deep, forceful chest movement in itself sets the blood to racing thereby livening up the circulation—which is good for us.

Without a word, Mr. Stevens caught up the tray from the piano and glided away on his toe-points; whereupon Mr. Brimberly (being alone) became astonishingly agile and nimble all at once, diving down to straighten a rug here and there, rearranging chairs and tables; he even opened the window and hurled two half-smoked cigars far out into the night; Their life followed a simple routine. Breakfast, from vegetables and Mark's canned store. Then the robot would work in the fields, and the plants grew used to his touch.

The whole edifice bears the same warm tinge of yellow that all those of good quality acquire from age in that pure climate. The untiring efforts of genius for over a century have succeeded in producing a musical instrument that falls little short of perfection.

Document Collection

Index


Inverted Index

• Default index structure in Information Retrieval
• Computationally very efficient. Scales well
• Words are sorted alphabetically to speed up access
• Frequency of a word in a document can also be stored

Inverted Index Construction

Documents to be indexed

Tokenizer

Token stream

Linguistic modules / Stemming

Modified tokens

Indexer

Inverted index


Tokenization

- **Input:** “Friends, Romans, Countrymen”
- **Output:** Tokens
  - *Friends*
  - *Romans*
  - *Countrymen*

A token is an instance of a sequence of characters

Each such token is now a candidate for an index entry, after further processing

Tokenization

- Issues in tokenization:
  - Finland’s capital →
    - Finland? Finlands? Finland’s?
  - Hewlett-Packard → Hewlett and Packard as two tokens?
    - state-of-the-art: break up hyphenated sequence.
    - co-education
    - lowercase, lower-case, lower case?
  - San Francisco: one token or two?
    - How do you decide it is one token?


Stemming

- Reduce terms to their “roots” before indexing
- “Stemming” suggests crude affix chopping
  - language dependent
  - e.g., automate(s), automatic, automation all reduced to automat.

- Approaches such as lemmatization also possible (e.g. am, are, is → be)
Conceptual Model for Search

Documents

Information Need

Document Representation

Indexing

Query

Formulation

Retrieved Documents

Retrieval Function

Relevance Feedback, Query Reformulation, Query Expansion

Further Analysis of the Documents

Types of Queries

- The type of query entered depends on what the search engine supports.

- Two main types of queries:
  - **Boolean**
    - Brutus AND Caesar
    - disabl! /p access! /s work-site work-place (employment /3 place)
  - **Free text queries**
    - Brutus Caesar
    - requirements disabled people access workplace
Search Interface

- Almost all IR systems are accessed through a search box
- There is usually also an advanced search option

Results

- Results are almost always viewed as a vertical list
Why are interfaces so simple?

- Search is a means towards some other end, rather than a goal in itself
- Search is a mentally intensive task
- Nearly everyone who uses the web uses search

- Therefore the interface should be non-distracting, non-intrusive and understandable

M. Hearst

Conceptual Model for Search

- Documents
  - Indexing
  - Document Representation
- Information Need
  - Formulation
- Query
- Retrieval Function
- Retrieved Documents
  - Relevance Feedback, Query Reformulation, Query Expansion
  - Further Analysis of the Documents
Information Retrieval Models

- The inverted index is used to access information about word presence and frequency in documents
- A retrieval model is a **mathematical**, potentially probabilistic, **model to rank retrieved documents**
- Tasks of IR models:
  - Process a query such that the result is specific (not too many hits and hits on topic) while being exhaustive (enough hits, good coverage)
  - Retrieve relevant documents while not retrieving non-relevant documents
  - Rank documents

Two Main Classes of IR Model

- **Boolean Retrieval Model**
- **Ranked Retrieval Model**
  - Vector space model (VSM)
  - BM25 / Okapi
  - Language Modelling
  - ...
Boolean Retrieval Model

- The **Boolean retrieval model** requires a query that is a Boolean expression:
  - Boolean Queries are queries using **AND, OR and NOT** to join query terms
  - Views each document as a **set of words**
  - Is precise: document matches condition or not.
  - Perhaps the simplest model on which to build an IR system
- Primary commercial retrieval tool for 3 decades

Advantages of Boolean Queries

- Precise: a document either matches a query or it does not
- Offers the user greater control and transparency over what is retrieved
- Good for expert users with precise understanding of their needs and the collection
Disadvantages of Boolean Models

- **Feast or Famine**
  - Boolean queries often result in either too few (zero) or too many (1000s) results.
  - It takes a lot of skill to come up with a query that produces a manageable number of hits.
  - AND gives too few; OR gives too many
  - Phrased another way: AND produces high precision but low recall; OR gives low precision but high recall

- **Difficult to rank output, some documents are more important than others**
  - Chronological order is often used

- **All terms are equally weighted**

- **Not good for the majority of users**

---

Two Main Classes of IR Model

- **Boolean Retrieval Model**
  - Extended Boolean Retrieval Model

- **Ranked Retrieval Model**
  - Vector space model (VSM)
  - BM25 / Okapi
  - Language Modelling
  - ...
Ranked Retrieval Models

- Rather than a set of documents satisfying a query expression, in ranked retrieval models, the system returns an ordering over the (top) documents in the collection with respect to a query.

- Free text queries: Rather than a query language of operators and expressions, the user’s query is just one or more words in a human language.

No more Feast or Famine Problem

- When a system produces a ranked result set, large result sets are not an issue.
  - The ranking already gives the user an idea of which documents are the best fit to the query.
  - The user doesn’t have to scan through 100s or 1000s of unranked results.

- Premise: the ranking algorithm works.
Scoring for Ranked Retrieval

- We wish to return the documents most likely to be useful to the searcher ranked highest.
- How can we rank-order the documents in the collection with respect to a query?
- Assign a score – say between 0 and 1 – to each document.
- This score measures how well document and query “match”.

Vector Space Model

- This is a simple model to calculate the similarity between documents, or between queries and documents.
- Vector representation doesn’t consider the ordering of words in a document.
- *John is quicker than Mary* and *Mary is quicker than John* have the same vectors.
- This is called the bag of words model.
Term-Document Count Vectors

- Consider the number of occurrences of a term in a document:
  - Each document is a count vector: a column below
  - In general very high dimensional vectors

<table>
<thead>
<tr>
<th>Term</th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>157</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brutus</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>232</td>
<td>227</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>


Document and Query Representation

- Each document is then a vector in a very high dimensional space
- The dimensionality is the number of words in the whole document collection
- The query is also represented as a vector in this space
Similarity

- The similarity between a query and a document is calculated as the angle between the query and document vectors.
- All documents can be ranked based on this similarity measure.


Some Details...

- Document representation vectors are usually not simply counts of words.
- Some weighting of word counts is usually applied so that words that occur in many/all documents receive lower weight, e.g. the, a, ...
### Advantages of the VSM

- Simple model based on linear algebra
- Term weights not binary
- Allows computing a **continuous degree of similarity** between queries and documents
- Allows ranking documents according to their possible relevance

### Limitations of the VSM

- Search keywords must precisely match document terms
- Semantic sensitivity; documents with similar context but different term vocabulary won't be associated
- The order in which the terms appear in the document is lost in the vector space representation
- Assumes terms are independent
- Weighting is intuitive but not very formal
Summary

- All ranked retrieval models try to rank according to the probability of relevance to the query
- Different models involve different weighting schemes
- Search engines usually go beyond a basic VSM, and allow search by e.g. phrases, wildcards or some (quasi-)Boolean operators

Open source search engines

- Lucene/SOLR
- Lemur/Indri
- MG4J
- Terrier
- ...

8/26/2012
Specificities of Internet Search

- Links are extremely common in web pages
- Internet search engines take advantage of these links
- How could this be done?

Web Search Basics

Link Analysis

- The most well known approach is **PageRank** (made famous by Google)
  - Every web page is assigned a PageRank score
  - Pages that are linked to by many pages have a higher score
  - Links are weighted by the PageRank of the linking pages
  - The final rank of a web page depends on a combination of features, such as similarity, term proximity, PageRank, ... (different per search engine)
Search Engine Optimisation

- Getting your web page to rank highly in a web search engine result list
- If you know how the search engine works, this can be done
- Constant battle between Search Engines and web page providers (Adversarial IR)
- Example:
  - Early web search engines relied heavily on the basic VSM to rank results
  - Repeating words gave a page a higher ranking (e.g. Repeating “maui resort” a few 100 times in white on a white background)
  - This no longer works!
Course Contents

- Introduction to Information Retrieval
- Who searches for medical information and how do they search?
- Search in the medical domain
- Improving search in the medical domain (Discussion)
- Searching for medical images
- Who searches medical images and how do they search?
- Combining text and visual search
- Challenges for search in the medical domain (Discussion)

End-Users of Health Information

- Physicians
- Specialists
- Nurses
- Medical Students
- Biomedical researchers
- Lay-people (general public)
- ...

8/26/2012
### Physician Information Needs

- Unrecognized Needs
- Recognized Needs
- Pursued Needs
- Satisfied Needs

### Unrecognized Needs

- Lack of awareness of the need
- Don’t know that new information is available
Recognized Needs

- Physicians recognise that they have an unmet information need

- Numbers from various studies:
  - Average of 2 unmet needs for every 3 patients (0.66 per patient) [CU85]
  - 1.4 questions per patient [OF91]

- Questions of type:
  - What is the cause of symptom X?
  - What is the dose of drug X?
  - How should I manage disease or finding X?
  - 69 in total [EO99]

Pursued Needs

- Physicians decided against pursuing answers for a majority of the unmet needs (from many studies)

- Most important reasons for not pursuing an answer [EO05]
  - Doubted existence of relevant information – 25%
  - Readily available consultation leading to referral rather than pursuit – 22%
  - Lack of time to pursue – 19%
  - Not important enough to pursue answer – 15%
  - Uncertain where to look for answer – 8%
Difficulties identified:

- **Time:**
  - Physicians search on average for less than 5 minutes, and seldom search for more than 10 minutes [HSV08].
  - The time taken to answer questions using MEDLINE averages 30 minutes [HH98], and the information found is often scattered over multiple articles, making PubMed searching MEDLINE impractical for intensive clinical use [HSV08]

- **Query language:**
  - Physicians tend to make simple queries, containing 2 to 3 terms on average [HSV08b], resulting in long lists of results (Boolean model of PubMed)

- **Language:**
  - Dutch-speaking physicians observed in the study [HSV08b] may have used erroneous English terms, resulting in poorer returned results

Satisfied Needs

- The information required is found

- The finding of relevant information could be improving as Internet affinity become more widespread
Where do physicians search for medical information?

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Search Engine</td>
<td>15</td>
<td>15</td>
<td>49</td>
<td>29</td>
<td>11</td>
</tr>
<tr>
<td>Medical research databases</td>
<td>8</td>
<td>12</td>
<td>21</td>
<td>35</td>
<td>24</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>8</td>
<td>21</td>
<td>31</td>
<td>33</td>
<td>7</td>
</tr>
<tr>
<td>Society websites</td>
<td>7</td>
<td>23</td>
<td>32</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>Hospital or University websites</td>
<td>25</td>
<td>37</td>
<td>32</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>Targeted/Area specialized websites</td>
<td>11</td>
<td>22</td>
<td>36</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>General health websites</td>
<td>17</td>
<td>31</td>
<td>38</td>
<td>38</td>
<td>19</td>
</tr>
<tr>
<td>Websites suggested by a colleague</td>
<td>15</td>
<td>36</td>
<td>42</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Point-of-Care databases</td>
<td>27</td>
<td>36</td>
<td>42</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>Pharmaceutical company websites</td>
<td>35</td>
<td>55</td>
<td>57</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Physician network communities</td>
<td>61</td>
<td>57</td>
<td>57</td>
<td>27</td>
<td>11</td>
</tr>
<tr>
<td>Medical forums/Blogs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialized medical search tool</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Survey: 560 participants

*Question asked: “How often do you use the following types of online resources to find online medical information?”

*Based on a sample where N= 558
### Other Groups

- Have different
- Needs
- Search behaviours
- ...

---

**Figure 3.4 How often does obtaining online information lead to the following actions?**

<table>
<thead>
<tr>
<th>Action</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conduct further research using other resources</td>
<td>10</td>
<td>18</td>
<td>45</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Request more information about a product or medication</td>
<td>16</td>
<td>38</td>
<td>35</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Modify a patient’s treatment or therapy</td>
<td>13</td>
<td>34</td>
<td>46</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Ask a colleague for his/her opinion</td>
<td>18</td>
<td>30</td>
<td>39</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Print out patient information/or recommend a website to the patient</td>
<td>21</td>
<td>29</td>
<td>35</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Recommend further tests based on symptoms</td>
<td>19</td>
<td>32</td>
<td>38</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Change a patient’s medication</td>
<td>17</td>
<td>37</td>
<td>40</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Recommend to a patient a behaviour change of habits</td>
<td>22</td>
<td>33</td>
<td>34</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>
Consumer Health Searchers

- Non-professionals can access large amount of health information on the Internet
- 61% of American Adults seek out health advice online
- Around a third of those surveyed admitted that they changed their thinking about how they should treat a condition based on what they found online (Pew Internet and American Life Project, June 2009)

Patients searching...

- The Internet is changing the doctor-patient relationship
- Want **empowered** patients but no Cyberchondria
  - But can they access information of high quality?
How often do you use the following types of online sources to find online health information?

- Search engines (Google, Yahoo, MSN, etc.)
- Web sites providing health information (Hospital, University, Pharma, Publishers)
- Wikipedia
- Specialized search tool (e.g., Medline Plus)
- Links from a health web site
- Web sites suggested by a health care provider
- Forums and blogs
- Portal web sites (Yahoo health, Google Health, MSN Health, etc.)
- Web sites suggested by a friend
- Facebook or other social networks

Course Contents

- Introduction to Information Retrieval
- Who searches for medical information and how do they search?
- Search in the medical domain
- Improving search in the medical domain (Discussion)
- Searching for medical images
- Who searches medical images and how do they search?
- Combining text and visual search
- Challenges for search in the medical domain (Discussion)
Health and Medical Information

Patient-specific information
- Narrative reports
- Structured data
- Images
- Radiology images
- -omics information

Knowledge-based information
- Journals
- Books
- Practice guidelines
- Taxonomies, vocabularies, ontologies, ...
- Language resources
- Websites, Web 2.0

• Primary – original research (in journals, books, reports, etc.)
• Secondary – summaries of research (in review articles, books, practice guidelines, etc.)

PubMed
- PubMed is an NLM search engine to search MEDLINE: http://www.pubmed.gov
- Pubmed uses a Boolean search model
- Results are returned in reverse chronological order
The Haynes 4S Model (EBM)

- An alternative knowledge-based information classification

- Systems
- Synopses
- Synteses
- Studies

Examples
- Computerised decision support systems
- Evidence based journal abstracts
- Cochrane reviews
- Original published articles in journals

Secondary literature
Primary literature
TRIP Database example

Medical vocabularies

- Many such vocabularies available:
  - Medical Subject Headings (MeSH) – literature
  - SNOMED CT – patient-specific information
  - ICD-10 – WHO International Classification of Diseases
  - RadLex – Radiology Lexicon
  - UMLS (Unified Medical Language System) - Metathesaurus

- Vocabularies can be seen as providing domain knowledge for search
Use of Vocabularies in IR

- Query suggestion
  - As the user types in a query, suggest terms from a vocabulary
  - NLM provides such a service for MeSH terms
Query Expansion

PubMed uses MeSH terms to expand queries

Translations:

<table>
<thead>
<tr>
<th>MeSH Term</th>
<th>Expanded Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>colon cancer</td>
<td>&quot;colonic neoplasms&quot;[MeSH Terms] OR &quot;colonic&quot;[All Fields] AND &quot;neoplasms&quot;[All Fields] OR &quot;colonic neoplasms&quot;[All Fields] OR &quot;colonic&quot;[All Fields] AND &quot;cancer&quot;[All Fields] OR &quot;colonic cancer&quot;[All Fields]</td>
</tr>
<tr>
<td>blocked nose</td>
<td>&quot;nasal obstruction&quot;[MeSH Terms] OR &quot;nasal&quot;[All Fields] AND &quot;obstruction&quot;[All Fields] OR &quot;nasal obstruction&quot;[All Fields] OR &quot;nasal&quot;[All Fields] AND &quot;nose&quot;[All Fields] OR &quot;blocked nose&quot;[All Fields]</td>
</tr>
</tbody>
</table>
Document annotation

- Find occurrences of words in documents and link them to the vocabulary
- Go beyond bag of words – allows queries like:
  - Find all documents that mention medication used in the treatment of cancer
- Difficulty: query languages tend to be complex, e.g. Mimir query

```
{Diabetes Insipidus} IN
{
  {{Section name="treatment"}} IN
  {{Document} OVER {{HONLabel targetAudience="Individuals"}}}
}
```

E.g. Exopatent: [http://exopatent.ontotext.com](http://exopatent.ontotext.com)

Annotation example

16196030
Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers. This paper describes the quantitative areal and laminar distribution of identified neuron populations projecting from areas of prefrontal cortex (PFC) to subcortical autonomic, motor, and limbic sites in the rat. Injections of the retrograde pathway tracer wheat germ agglutinin conjugated with horseradish peroxidase (WGA-HRP) were made into dorsal/ventral striatum (DS/VS), basolateral amygdala (BLA), mediodorsal thalamus (MD), lateral hypothalamus (LH), mediodorsal septum, dorsolateral periaqueductal gray, dorsal raphe, ventral tegmental area, parabrachial nucleus, nucleus tractus solitarius, rostral/caudal ventrolateral medulla, or thoracic spinal cord (SC). High-resolution flat-map density distributions of retrogradely labelled neurons indicated that specific prefrontal cortex (PFC) regions were differentially involved in the projections studied, with medial (m) prefrontal cortex (PFC) divided into dorsal and ventral sectors. The percentages that wheat germ agglutinin conjugated with horseradish peroxidase (WGA-HRP) retrogradely labelled neurons composed of the projection neurons in individual layers of infralimbic (IL), area 25, prelimbic (PL), area 24, and dorsal anterior cingulate (ACC; area 24b) cortices were calculated. Among layer 5 pyramidal cells, approximately 27.4% in infralimbic (IL)/prelimbic (PL)/ACC cortices projected to lateral hypothalamus (LH), 22.9% in infralimbic (IL)/ventral prelimbic (PL) to VS, 18.3% in ACC dorsal prelimbic (PL) to VS, 13% in lateral hypothalamus (LH)/prelimbic (PL) to basolateral amygdala (BLA), and 37% of layer 6 pyramidal cells in infralimbic (IL)/prelimbic (PL)/ACC project to mediodorsal thalamus (MD). Data for other projection pathways are given. Multiple retrograde fluorescent tracing studies indicated that moderate populations (~5%) of layer 5 prefrontal cortex (PFC) neurons projected to lateral hypothalamus (LH)/VS, lateral hypothalamus (LH)/spinal cord (SC), or VS/basolateral amygdala (BLA). The data provide new quantitative information concerning the density and distribution of neurons involved in identified projection pathways from defined areas of the rat prefrontal cortex (PFC) to specific subcortical targets involved in dynamic goal-directed behavior.
Classification constraint

Know from the labels and ontology information if a classification of organs in an image is possible

Multilingual search

Map terms in many languages into the vocabulary

Example: http://www.wrapin.org

diabetes, autoimmune →
("diabète de type i" OR "diabète auto-immun" OR "diabète insulinodépendant" OR "diabète juvénile")

Allow browsing through related terms
Search Engines

- About 70% of the top websites with information on oral cancers gathered by Google and Yahoo searches had serious deficiencies [LC09]
  - websites failed to attribute authorship, cite sources and report conflicts of interest.

- On the first page of results, “lawyers were the most common sponsors of websites retrieved by the terms cerebral palsy (52%), birth trauma (48%), and shoulder dystocia (43%)” [KCB08]

Wikipedia

- Wikipedia articles appear in the top 10 results for more than 70% of medical queries in four different search engines tested in [LV09]

- Whereas Wikipedia medical articles have been found to be accurate, they are also often incomplete.

  - E.g. a study on drug information comparing Wikipedia to the Medscape Drug Reference [CPK08] found that “no factual errors were found in Wikipedia, whereas 4 answers in Medscape conflicted with the answer key.” However, “Wikipedia was able to answer significantly fewer drug information questions (40.0%) compared with MDR (82.5%).”

  - An advantage of Wikipedia was that “there was a marked improvement in Wikipedia over time, as current entries were superior to those 90 days prior.”
Various criteria for the quality of health web pages have been put forward.

E.g. Health on the Net is an NGO that certifies health web pages satisfying the HONcode Principles: http://www.healthonnet.org

Semi-automatic certification

Have a search engine that searches certified pages

HONcode principles

1. Authoritative
   - Indicate the qualifications of the authors
2. Complementarity
   - Information should support, not replace, the doctor-patient relationship
3. Privacy
   - Respect the privacy and confidentiality of personal data submitted to the site by the visitor
4. Attribution
   - Cite the source(s) of published information, date and medical and health pages
5. Justifiability
   - Site must back up claims relating to benefits and performance
6. Transparency
   - Accessible presentation, accurate email contact
7. Financial disclosure
   - Identify funding sources
8. Advertising policy
   - Clearly distinguish advertising from editorial content
Reference


References