Searching Text and Images in the Medical Domain

Allan Hanbury and Henning Müller

"The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528"

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

3

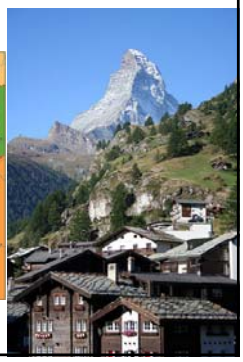
Henning Müller

- 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

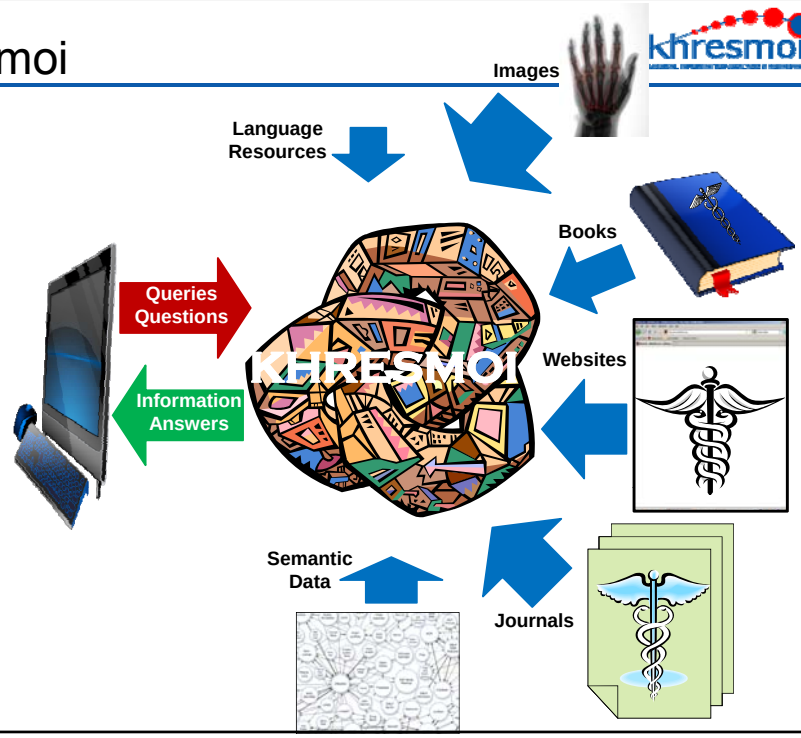
4

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



Khresmoi partners

Visit the Khresmoi Stand!

8

Course Contents

- Introduction to Information Retrieval
- Who searches for medical information and how do they search?
- Search in the medical domain

Allan

- 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

Henning

- Challenges for search in the medical domain (Discussion)

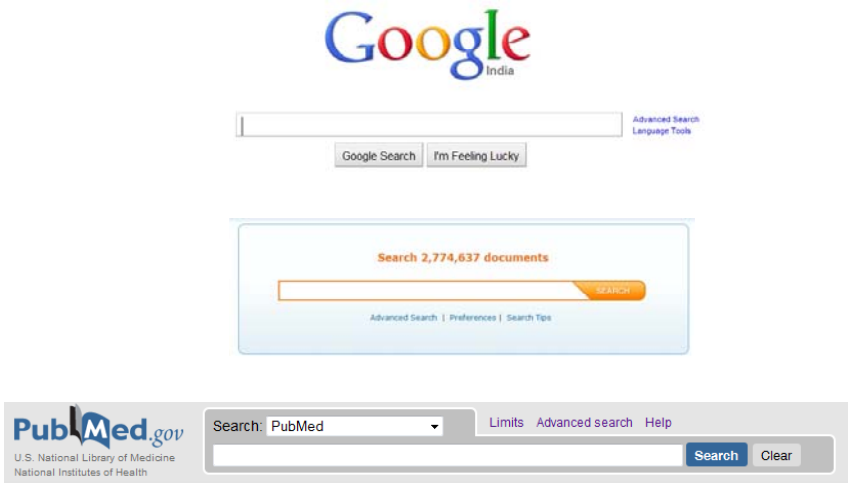
Course Contents

- 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)

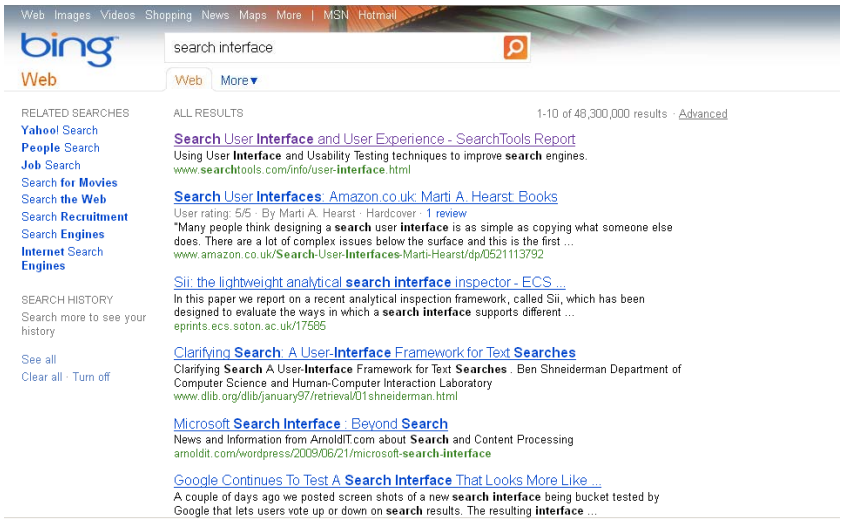
Contents

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

11



12



13

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

14

IR vs. Databases

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

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,
Salary < 60000 AND Manager = Smith.

From: http://nlp.stanford.edu/IR-book/ 15

Unstructured Information

- Text
- Images
- Music
- Videos

As opposed to

- Relational databases
- Lists of numbers

16

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

From: http://nlp.stanford.edu/IR-book/ 17

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

18

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

19

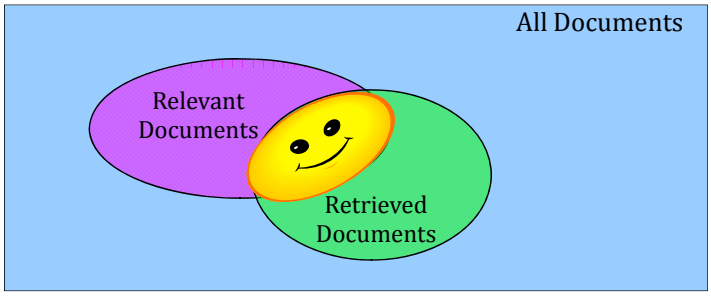
Precision and Recall

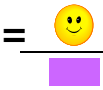
- A query returns *n* ranked documents from a database of many.
- Each one is judged as relevant or not:

Rank	Relevant
1	YES
2	YES
3	NO
4	YES
5	NO
…	
n	NO

20

Precision and Recall Concepts



- Precision =  Recall = 

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?

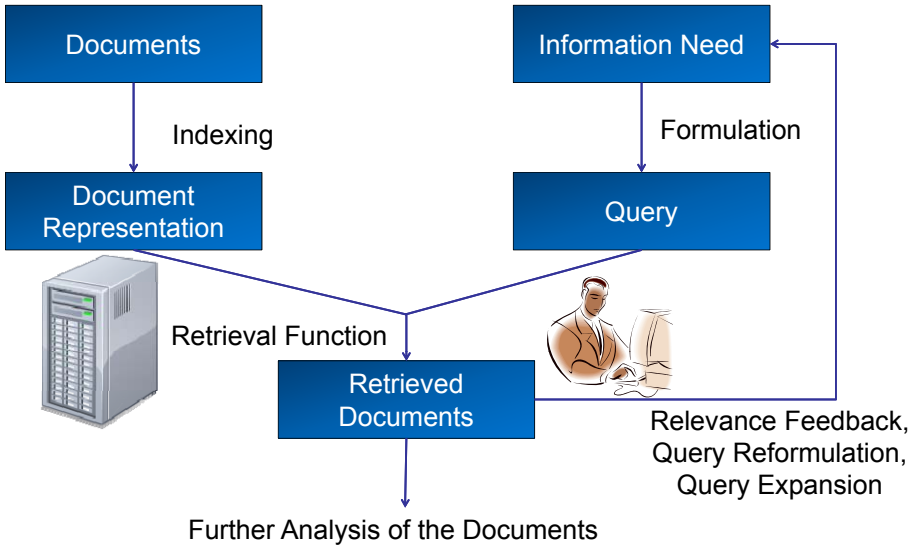
$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}}$$

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)

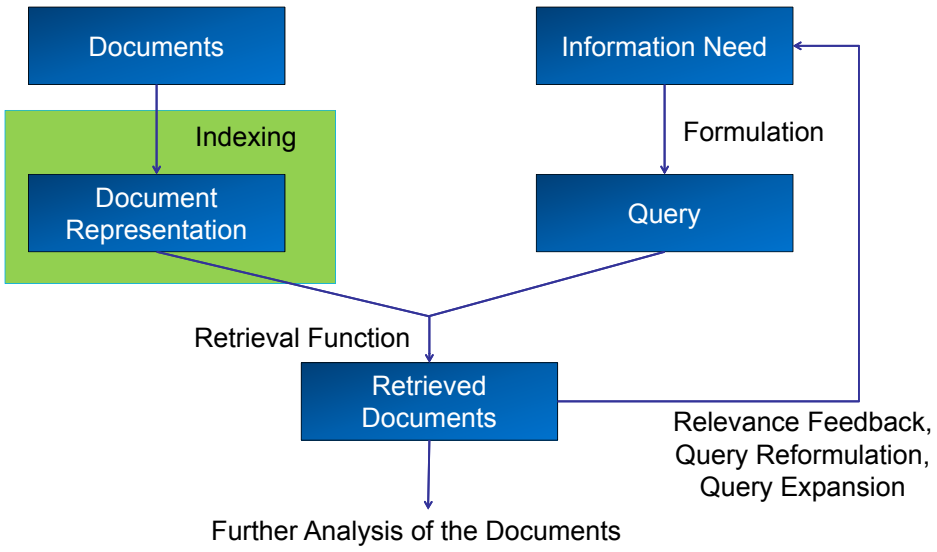
23

Conceptual Model for Search



24

Conceptual Model for Search



25

Indexing

- 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

26

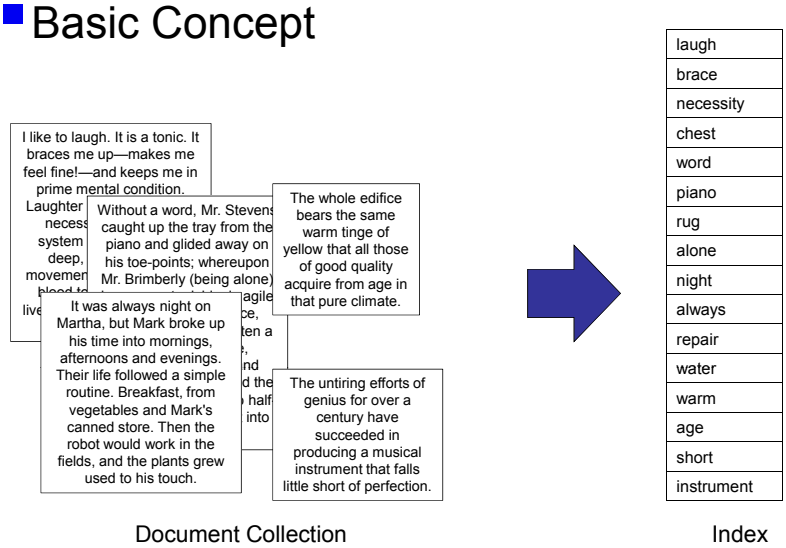
Aim of Indexing

- 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

27

Overview of Indexing Process

- Basic Concept



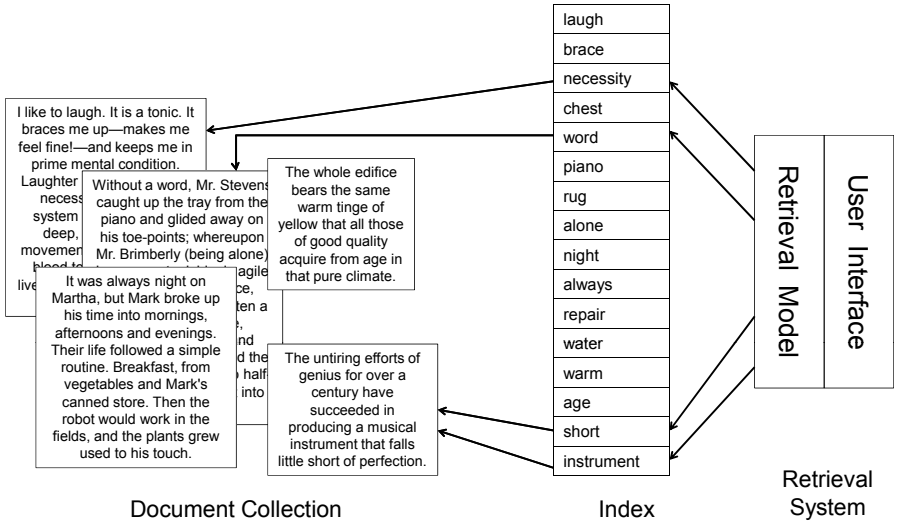
Document Collection

Index

From: http://nlp.stanford.edu/IR-book/

28

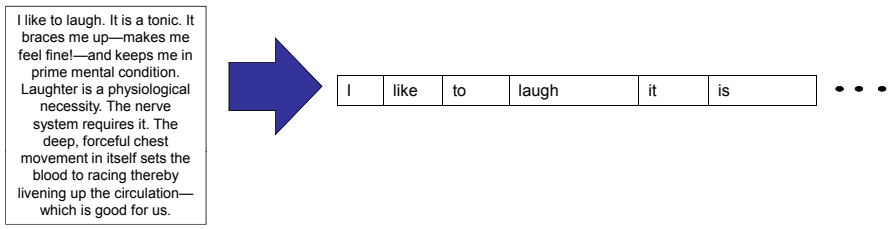
Overview of Indexing Process



From: http://nlp.stanford.edu/IR-book/ 29

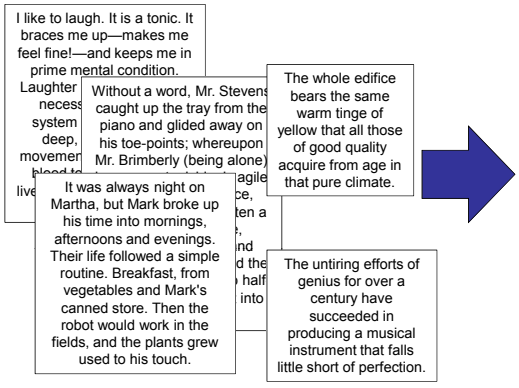
Document Representations (Glimpse)

- Represent documents via the complete set of terms



From: http://nlp.stanford.edu/IR-book/ 30

Index Creation



D1	I	like	to	laugh	it
D2	without	a	word	mr	stevens
D3	the	whole	edifice	bears	the
D4	it	was	always	night	on
D5	the	untiring	efforts	of	genius

Document Collection

Index

From: http://nlp.stanford.edu/IR-book/ 31

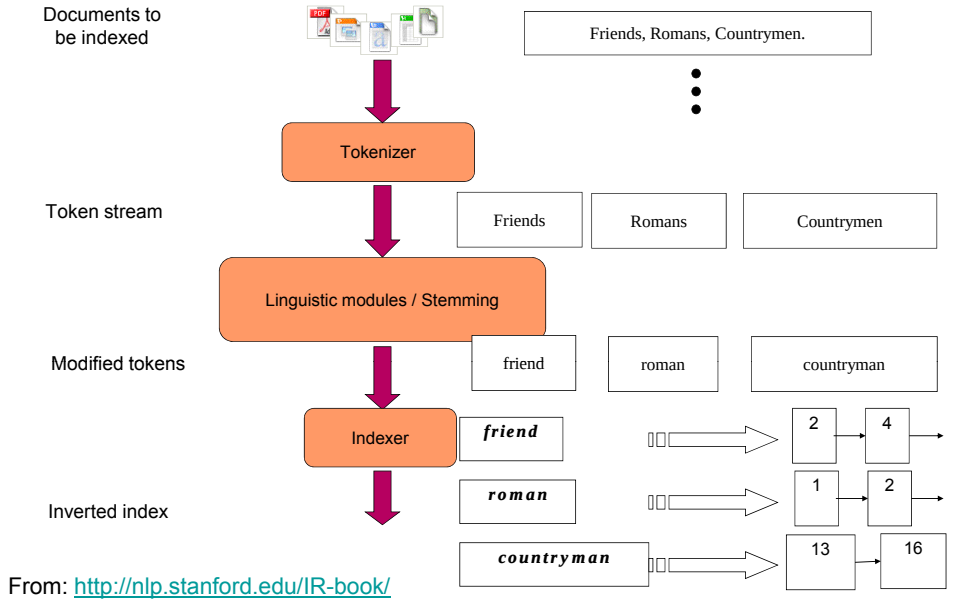
Inverted Index

- 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

From: http://nlp.stanford.edu/IR-book/

I	D1
like	D1
to	D1
laugh	D1
it	D1, D4
without	D2
a	D2
word	D2
mr	D2
stevens	D2
the	D3, D5
whole	D3
edifice	D3
bears	D3
was	D4
always	D4
night	D4
on	D4
untiring	D5
efforts	D5
of	D5
genius	D5

Inverted Index Construction



From: http://nlp.stanford.edu/IR-book/

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

From: http://nlp.stanford.edu/IR-book/ 35

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?

From: http://nlp.stanford.edu/IR-book/ 36

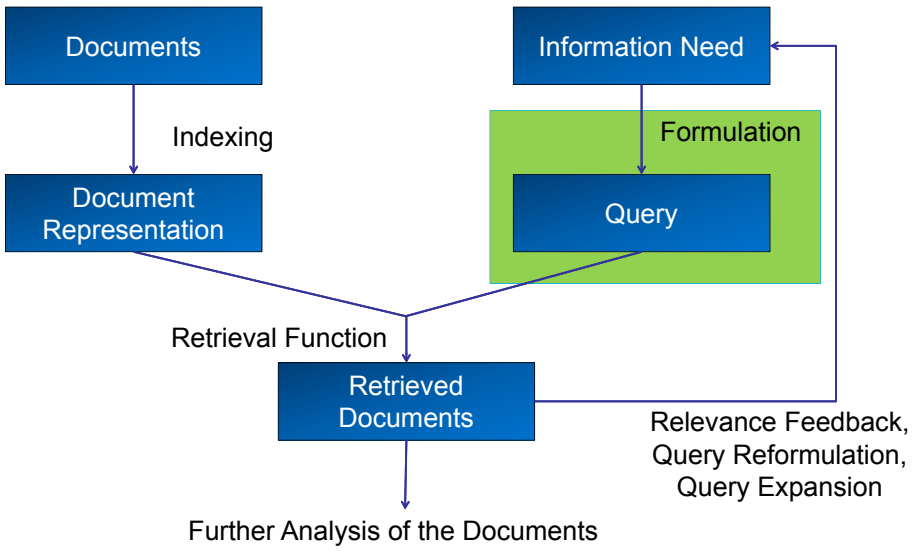
Stemming

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

for example compressed and compression are both accepted as equivalent to compress.	for exampl compress and compress ar both accept as equival to compress

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

Conceptual Model for Search



38

Types of Queries

- The type of query entered depends on what the search engine supports.
- Two main types of queries: Search
  - 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

39

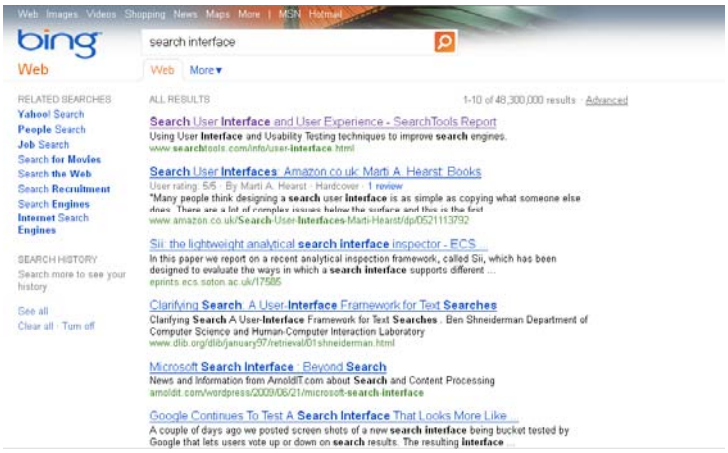
Search Interface

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

40

Results

- Results are almost always viewed as a vertical list



41

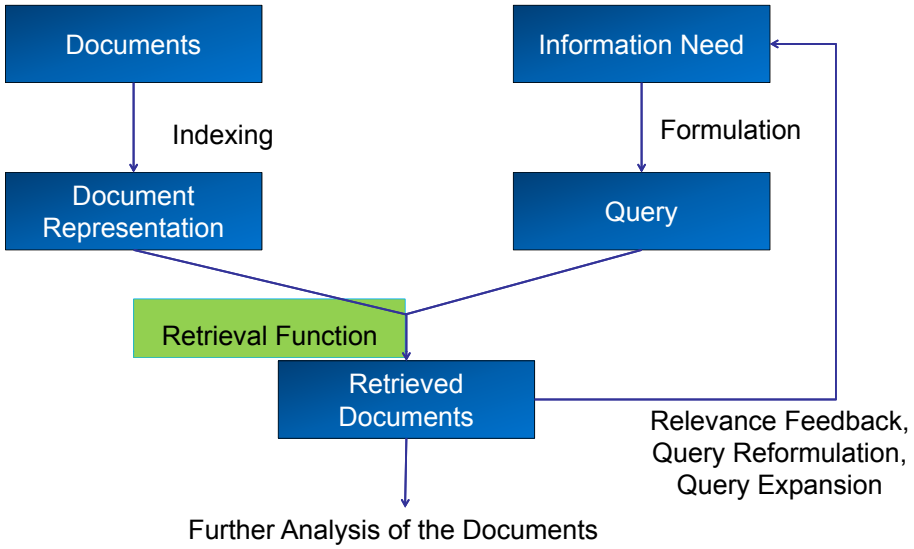
Why are interfaces so simple?

- 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

42

Conceptual Model for Search



43

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

44

Two Main Classes of IR Model

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

45

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

46

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

47

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

48

Two Main Classes of IR Model

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

49

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

50

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

51

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"

52

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

53

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

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

From: http://nlp.stanford.edu/IR-book/ 54

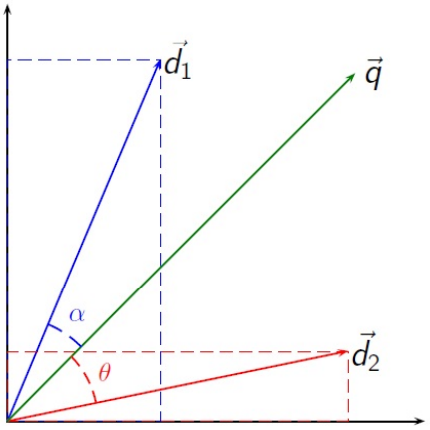
Document and Query Representation

- 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

55

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



From: http://nlp.stanford.edu/IR-book/ 56

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, ...

57

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

58

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

59

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

60

Open source search engines

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

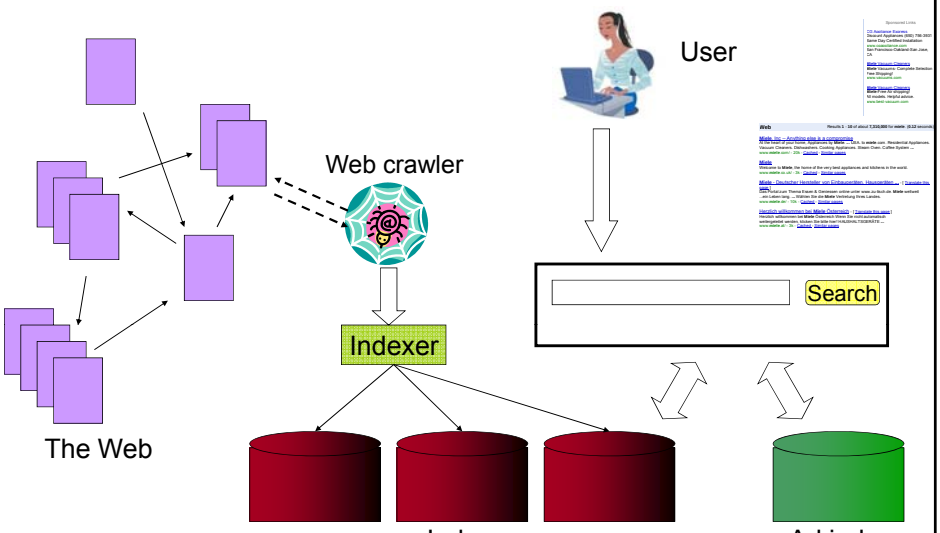
61

Specificities of Internet Search

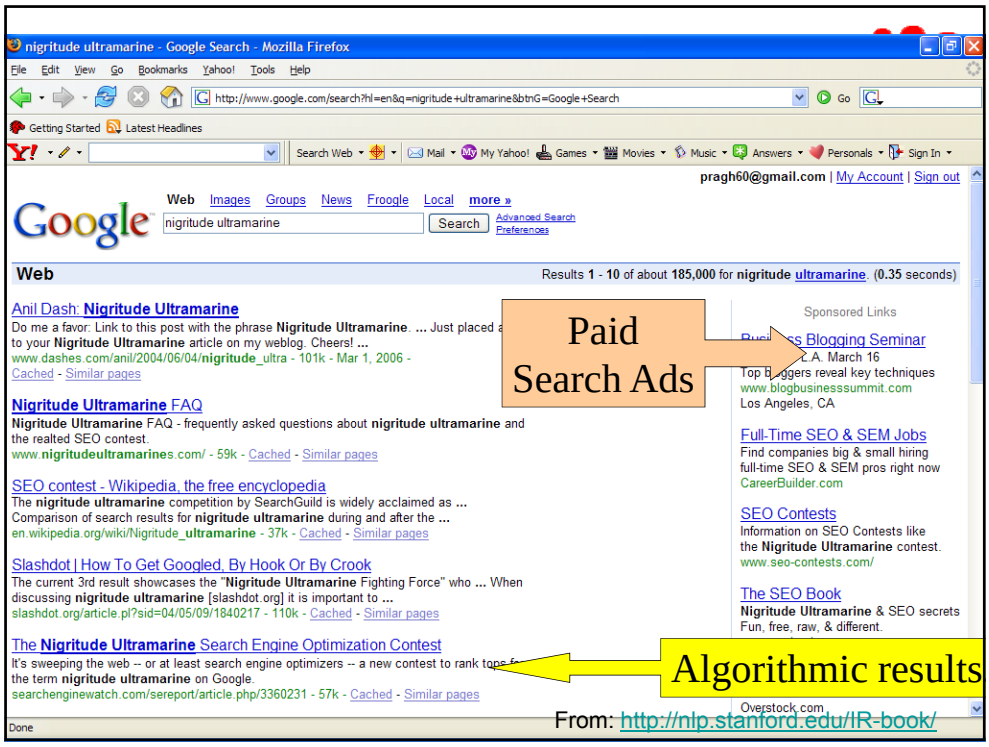
- Links are extremely common in web pages
- Internet search engines take advantage of these links
- How could this be done?

62

Web Search Basics



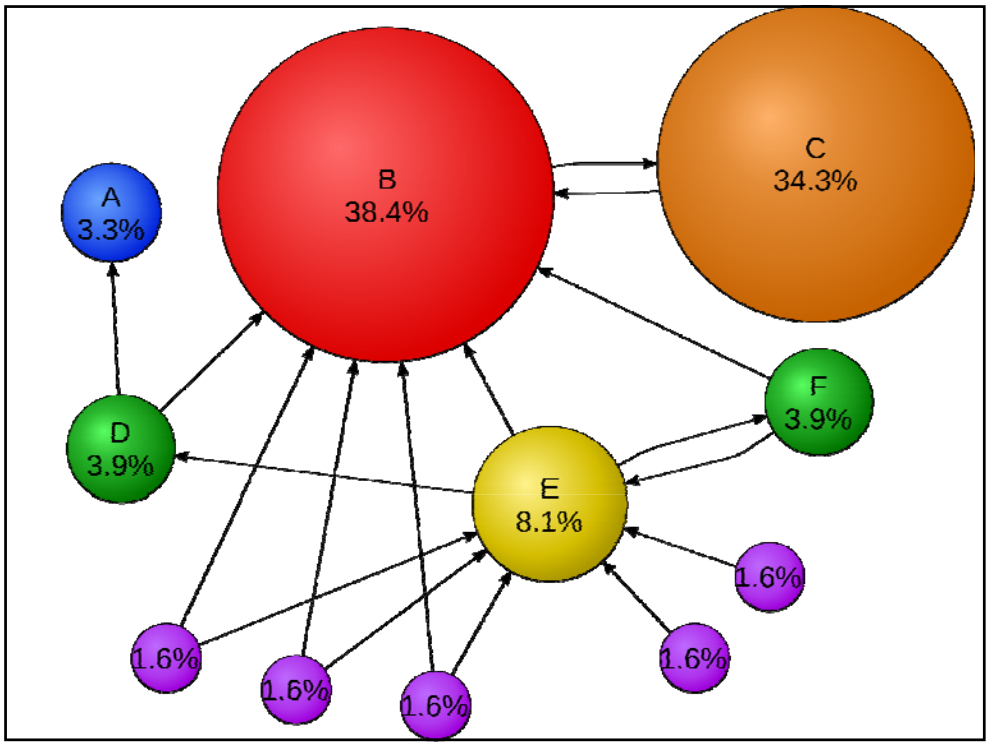
From: http://nlp.stanford.edu/IR-book/



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)

66



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 pages a higher ranking (e.g. Repeating “maui resort” a few 100 times in white on a white background)
  - This no longer works!

68

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)
- ...

70

Physician Information Needs

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

71

Unrecognized Needs

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

72

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]

73

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%

74

- 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

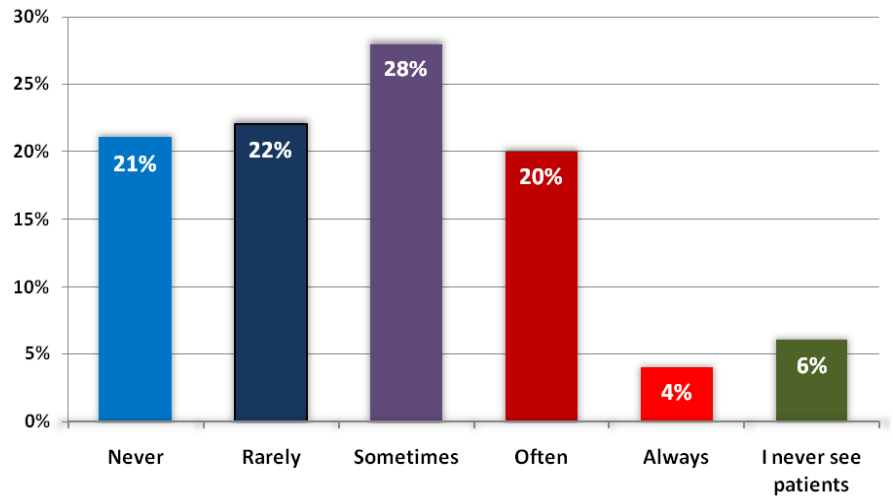
75

Satisfied Needs

- The information required is found
- The finding of relevant information could be improving as Internet affinity become more widespread

76

Figure 2.3 How often do you access the Internet during a medical/patient consultation?*

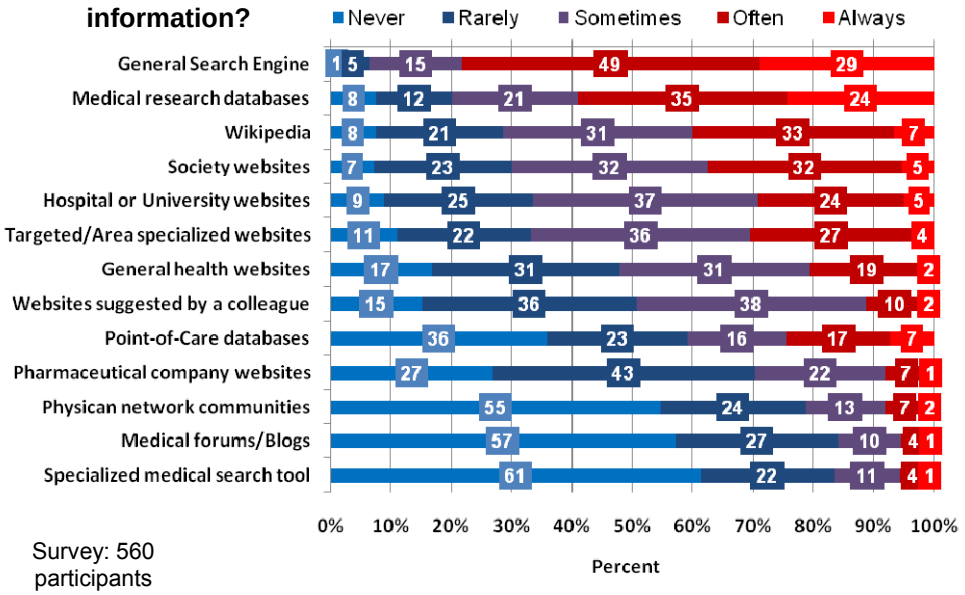


* Based on a sample where N= 558

77

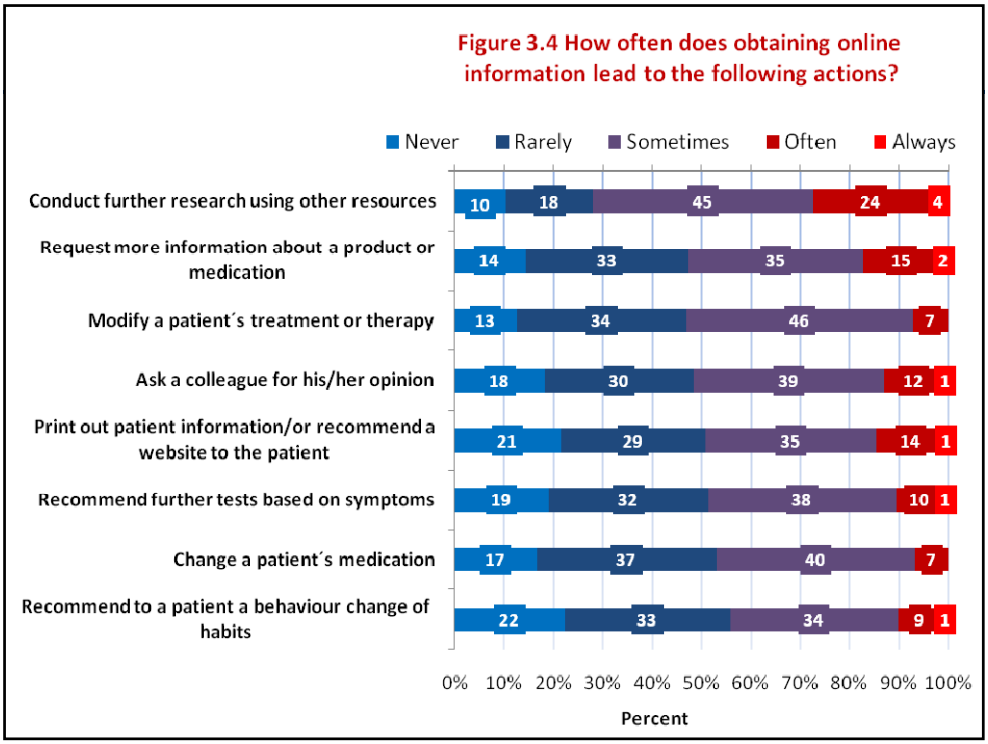
Where do physicians search for medical information?

Figure 4.1 Frequency of use of online resources ?*



Survey: 560 participants

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



Other Groups

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

83

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)

84

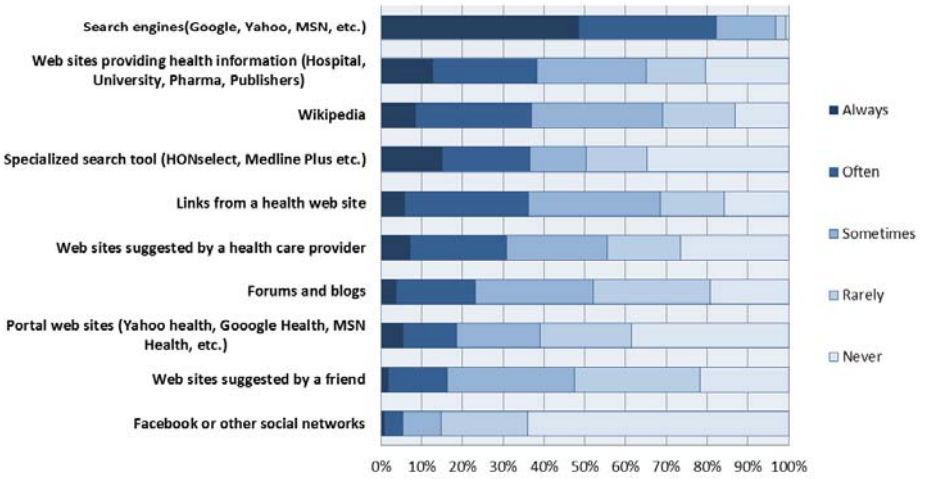
Patients searching...

- The Internet is changing the doctor-patient relationship
- Want empowered patients but no Cyberchondria
  - But can they access information of high quality?

86

General public information sources

How often do you use the following types of online sources to find online health information?



87

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

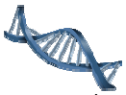

Structured data


Images

EHR/EMR


Radiology images

PACS


-omics information

Microsoft HealthVault

Personal Health Records (PHR)

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.)

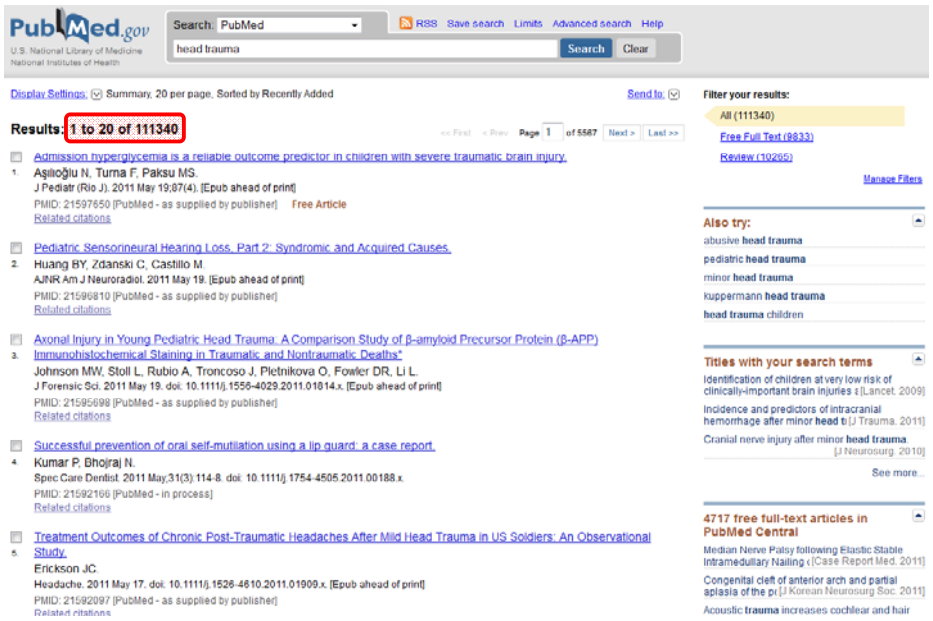
89

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

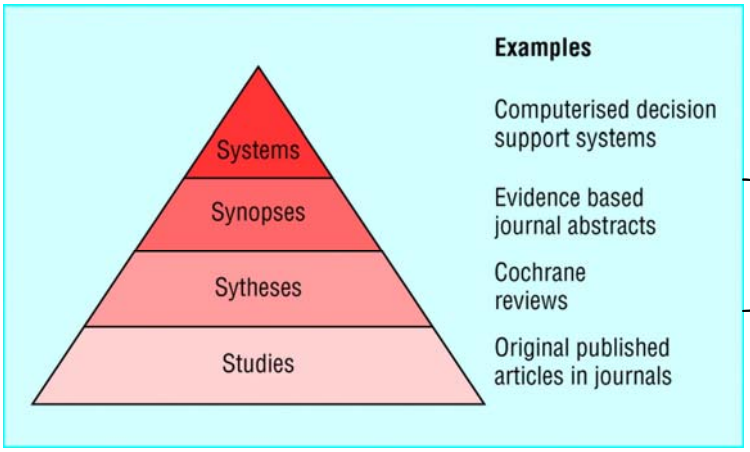
90

PubMed



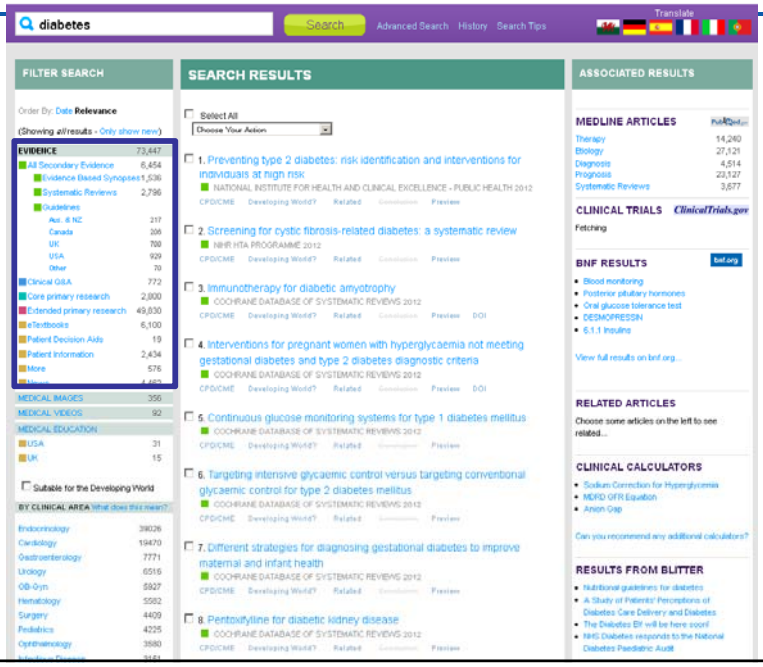
The Haynes 4S Model (EBM)

- An alternative knowledge-based information classification



92

TRIP Database example



Medical vocabularies

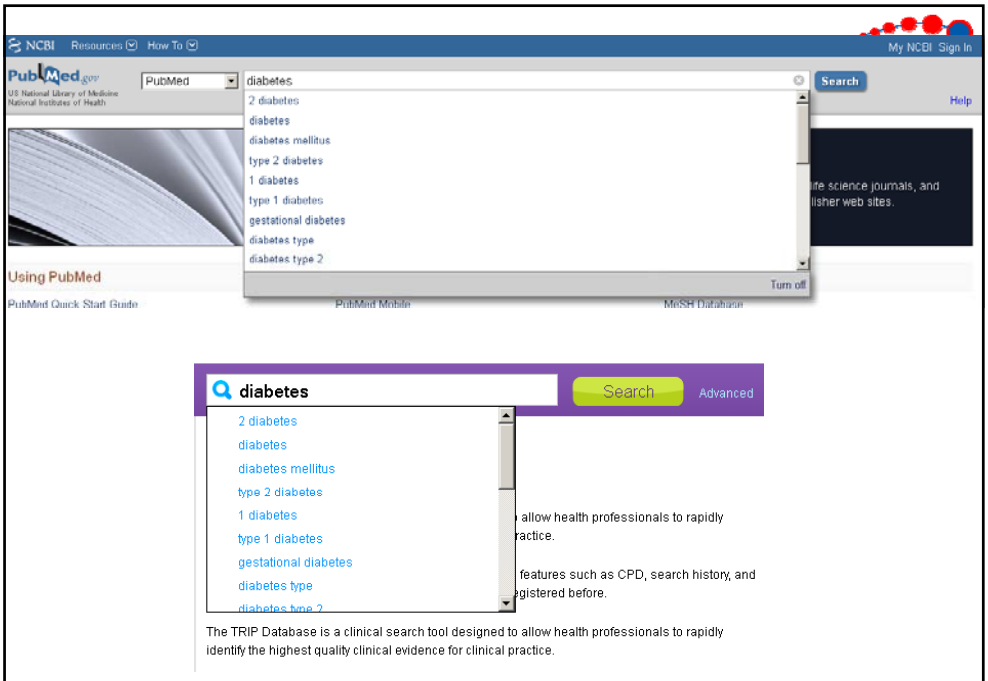
- Many such vocabularies available:
  - Medical Subject Headings (MeSH) – literature
  - SNOMED CT – patient-specific information
  - ICD-10 – WHO International Classification of Diseases
  - CPT – Current Procedural Terminology
  - RadLex – Radiology Lexicon
  - UMLS (Unified Medical Language System) - Metathesaurus
- Vocabularies can be seen as providing domain knowledge for search

94

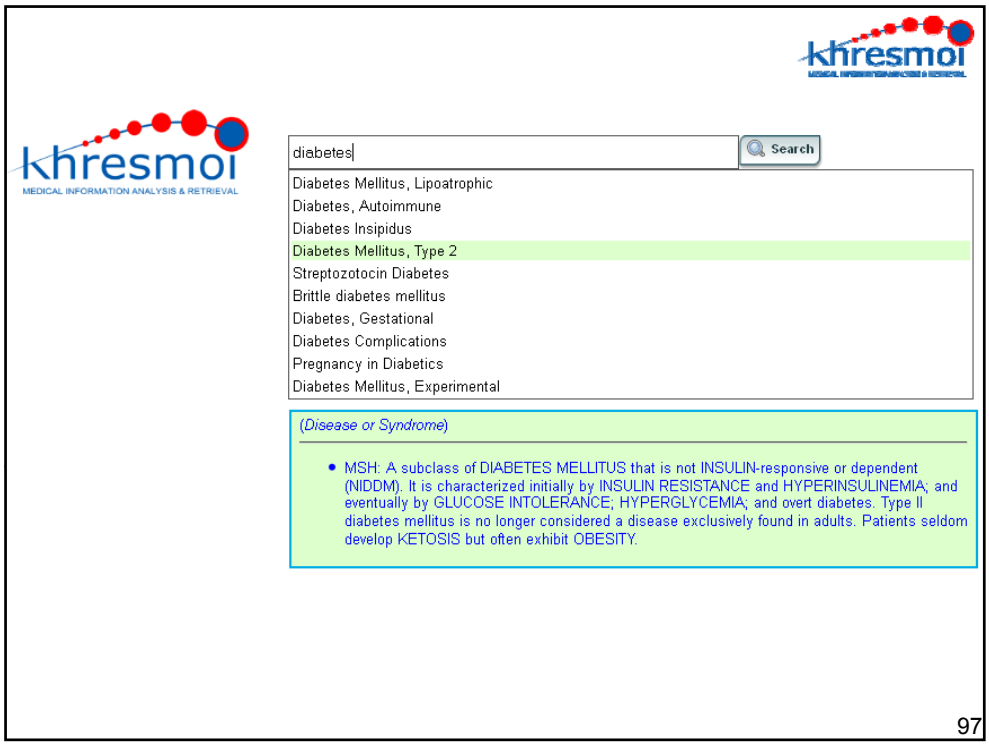
Use of Vocabularies in IR

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

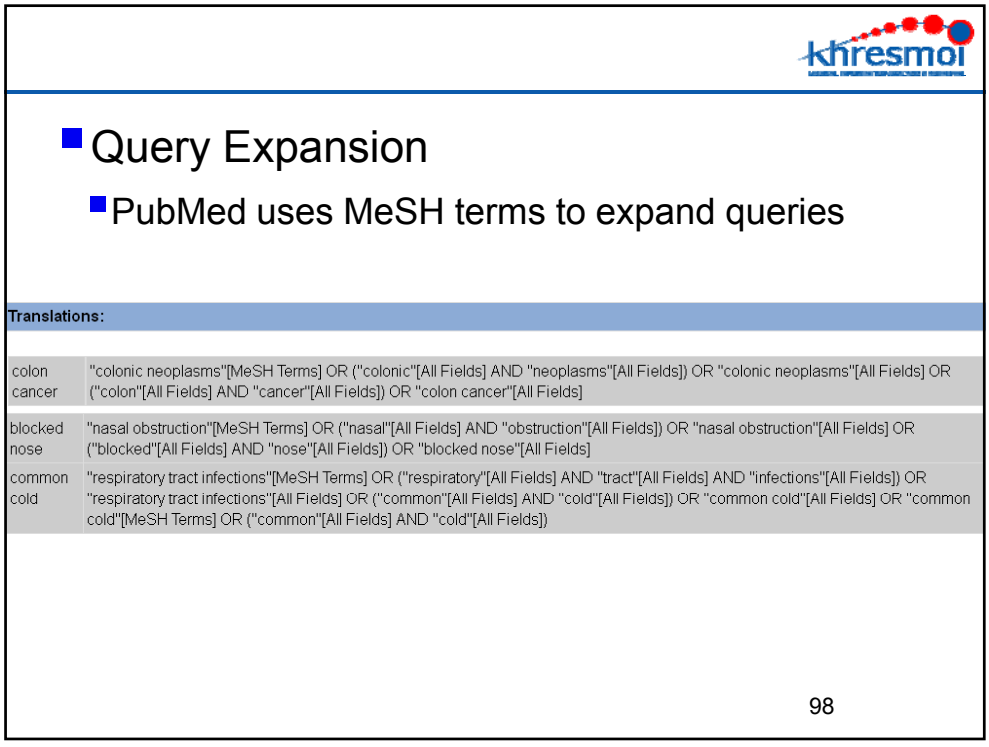
95



96


khresmoi
MEDICAL INFORMATION ANALYSIS & RETRIEVAL
diabetes
Search
Diabetes Mellitus, Lipoatrophic
Diabetes, Autoimmune
Diabetes Insipidus
Diabetes Mellitus, Type 2
Streptozotocin Diabetes
Brittle diabetes mellitus
Diabetes, Gestational
Diabetes Complications
Pregnancy in Diabetics
Diabetes Mellitus, Experimental
(Disease or Syndrome)
MSH: A subclass of DIABETES MELLITUS that is not INSULIN-responsive or dependent (NIDDM). It is characterized initially by INSULIN RESISTANCE and HYPERINSULINEMIA; and eventually by GLUCOSE INTOLERANCE; HYPERGLYCEMIA; and overt diabetes. Type II diabetes mellitus is no longer considered a disease exclusively found in adults. Patients seldom develop KETOSIS but often exhibit OBESITY.

97


khresmoi
Query Expansion
PubMed uses MeSH terms to expand queries
Translations:
colon cancer
"colonic neoplasms"[MeSH Terms] OR ("colonic"[All Fields] AND "neoplasms"[All Fields]) OR "colonic neoplasms"[All Fields] OR ("colon"[All Fields] AND "cancer"[All Fields]) OR "colon cancer"[All Fields]
blocked nose
"nasal obstruction"[MeSH Terms] OR ("nasal"[All Fields] AND "obstruction"[All Fields]) OR "nasal obstruction"[All Fields] OR ("blocked"[All Fields] AND "nose"[All Fields]) OR "blocked nose"[All Fields]
common cold
"respiratory tract infections"[MeSH Terms] OR ("respiratory"[All Fields] AND "tract"[All Fields] AND "infections"[All Fields]) OR "respiratory tract infections"[All Fields] OR ("common"[All Fields] AND "cold"[All Fields]) OR "common cold"[All Fields] OR "common cold"[MeSH Terms] OR ("common"[All Fields] AND "cold"[All Fields])

98

- 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

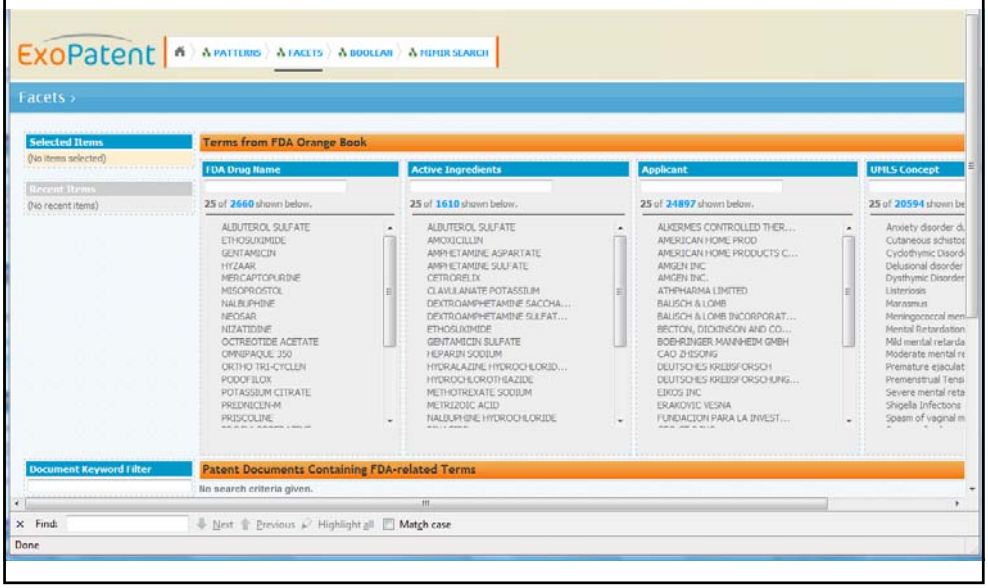
99

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), mediolateral 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 32), and dorsal anterior cingulate (ACd; area 24b) cortices were calculated. Among layer 5 pyramidal cells, approximately 27.4% in infralimbic(IL) / prelimbic(PL) /ACd cortices projected to lateral hypothalamus(LH) , 22.9% in infralimbic(IL) /ventral prelimbic(PL) to VS, 18.3% in ACd/dorsal prelimbic(PL) to DS, and 8.1% in areas infralimbic(IL) / prelimbic(PL) to basolateral amygdala(BLA) ; and 37% of layer 6 pyramidal cells in infralimbic(IL) / prelimbic(PL) /ACd projected to mediodorsal thalamus(MD) . Data for other projection pathways are given. Multiple dual retrograde fluorescent tracing studies indicated that moderate populations (<9%) of layer 5 m 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.

100

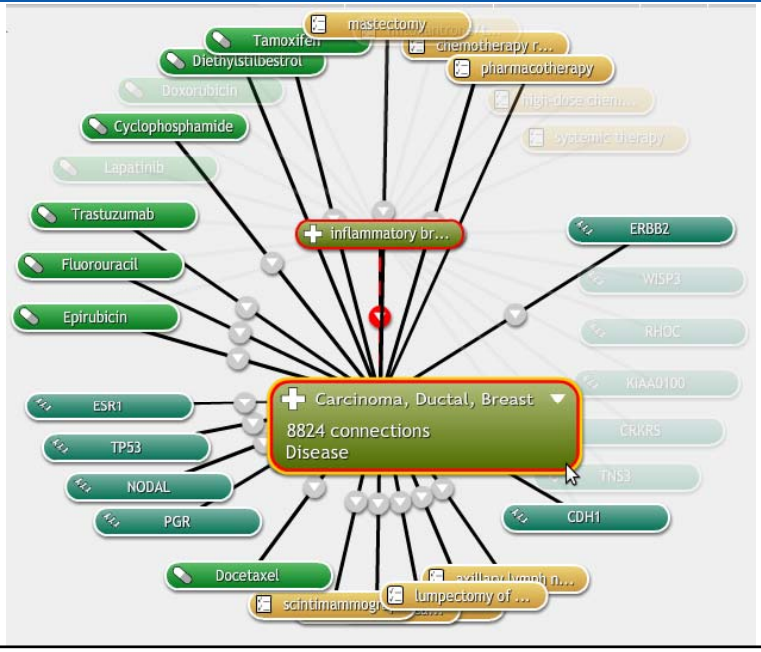
http://exopatent.ontotext.com



- 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

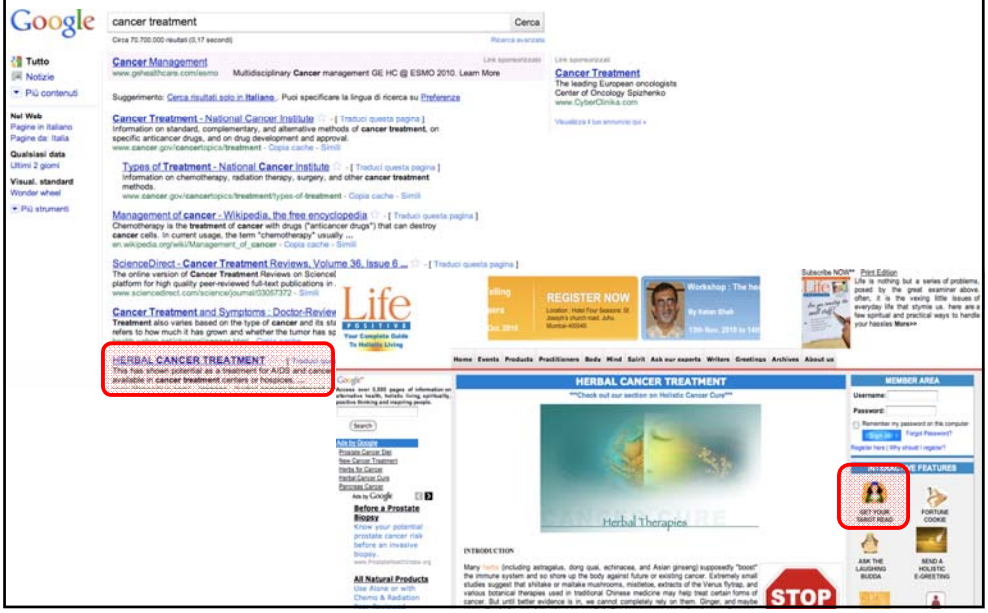
102

Coreminer.com



103

Information Trustability



Search Engines

- About 70% of the top websites with information on oral cancers gathered by Google and Yahoo searches had serious deficiencies [LC09]
  - web sites 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]

105

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.”

106

Codes of Conduct

- 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

107

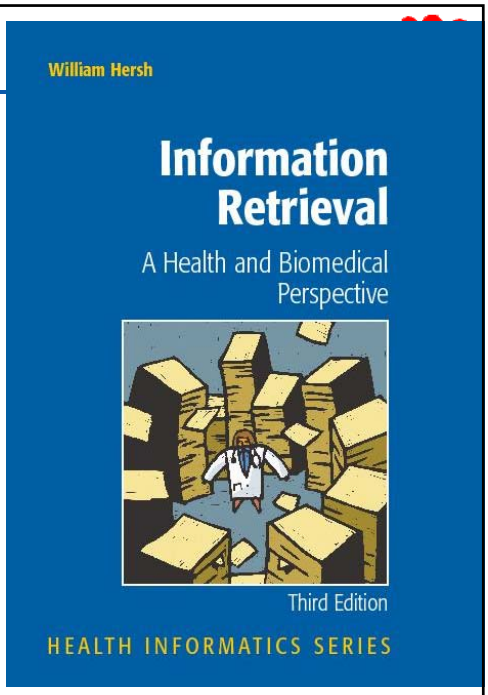
HONcode principles

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

108

Reference

- William Hersh, M.D., Information Retrieval: A Health and Biomedical Perspective, Third Edition, Springer, 2009



110

References

[CPK08] K. A Clauson, H. H. Polen, M. N. Kamel Boulos, J. H. Dzenowagis, Scope, Completeness, and Accuracy of Drug Information in Wikipedia, The Annals of Pharmacotherapy, Volume 42, No. 12, pages 1814-1821, 2008

[CU85] D. Covell, G. Uman, et al, Information needs in office practice: are they being met? Annals of Internal Medicine, 103:596-599, 1985

[EO99] J. Ely, J. Osheroff, et al., Analysis of questions asked by family doctors regarding patient care, British Medical Journal, 319(7206):358-61, 1999

[EO05] J. Ely, J. Osheroff, Answering Physicians' Clinical Questions: Obstacles and Potential Solutions, J Am Med Inform Assoc., 12(2): 217–224, 2005.

[HH98] W. R. Hersh, D. H. Hickam, How Well Do Physicians Use Electronic Information Retrieval Systems? A Framework for Investigation and Systematic Review, Journal of the American Medical Association, 280:15, 1998

[HSV08] A Hoogendam, A. F. H. Stalenhoef, P. F de Vries Robbé, A. J. P. M. Overbeke, Answers to Questions Posed During Daily Patient Care Are More Likely to Be Answered by UpToDate Than PubMed, J Med Internet Res, Volume 10, Number 4, 2008.

[HSV08b] A. Hoogendam, A. F. H. Stalenhoef, P. F. de Vries Robbé, A. J. P. M. Overbeke, Analysis of queries sent to PubMed at the point of care: Observation of search behaviour in a medical teaching hospital, BMC Medical Informatics and Decision Making 2008, Volume 8, Number 42, 2008

[KCB08] A. J. Kamal, Y. W. Cheng, A. S. Bryant, M. E. Norton, B. L. Shaffer, A. B. Caughey, Google obstetrics: who is educating our patients?, American Journal of Obstetrics & Gynecology, Volume 198, Number 6, June 2008.

[LC09] P. López-Jornet, F. Camacho-Alonso, The quality of Internet sites providing information relating to oral cancer, Oral Oncology, 2009.

[LV09] M. R. Laurenta, T. J. Vickers, Seeking Health Information Online: Does Wikipedia Matter?, Journal of the American Medical Informatics Association, Volume 16, pages 471-479, 2009

[OF91] J. Osheroff, D. Forsythe, et al., Physicians' information needs: analysis of questions posed during clinical teaching, Annals of Internal Medicine, 114:576-581, 1991

111