Web Search

Course presentation

João Magalhães

How to search multimedia information?

- Textual and visual data can communicate a wide variety of information that are critical for several decision processes.

- Temporal and spatial structure adds organization and usability to information.

- Non-structured data (language and vision) puts a heavy complexity burden on standard data structures.

Funny reflex...

Share This

ADD NOTE SEND TO GROUP ADD TO SET BLOG THIS ALL SIZES ORDER PRINTS ROTATE EDIT PHOTO DELETE

Uploaded on August 3, 2006
by jmc_mag

jmc_mag's photostream

Best ones (Set)

You are at the first photo.

11 items

browse

Comments

alexei_322 pro says:

thumbs up! great shot
Posted 31 months ago. (permalink | delete)

Add your comment

Tags

- Arizona
- Reflection
- Desert

Add a tag

Additional Information

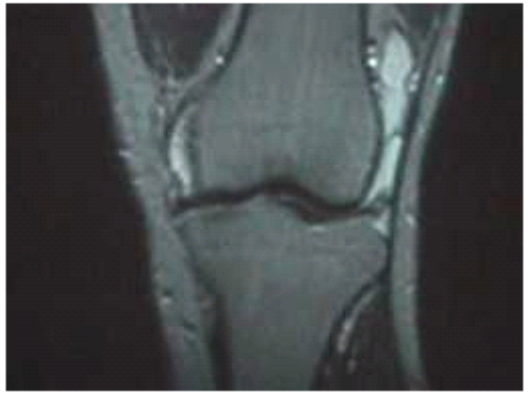
- All rights reserved (edit)
- Anyone can see this photo (edit)
- Add to your map
- Taken with a Canon EOS Digital Rebel XT.
  More properties
- Taken on July 15, 2006 (edit)
- Photo stats
- Viewed 26 times (Not including you)
- Edit title, description, and tags

Example MRI Scan Reports

This patient had a sudden loss of her motor functions (she wasnt able to move her right arms and legs) 2 months before the study. She went thru a slow recovery with lot physical therapy and drugs. She was recovering some of her movements but suddenly all the improvement stop. We performed an MRI that showed the changes expected for a lesion of that time (2 months old) but also showed and increase in the size of the ventricular system(where the Cerebrospinal fluid or CSF flows) that was causing hydrocephalus. Due to this finding, the patient went thru another surgery and had a shunt valve installed, the last word we had from one of her relatives is that she is again on recovery.

The *official* report included this: T 1 coronal SE (spin echo) sequence that shows an area of infarction in the left parietal lobe. Also enlargement of the ventricular system is observed.

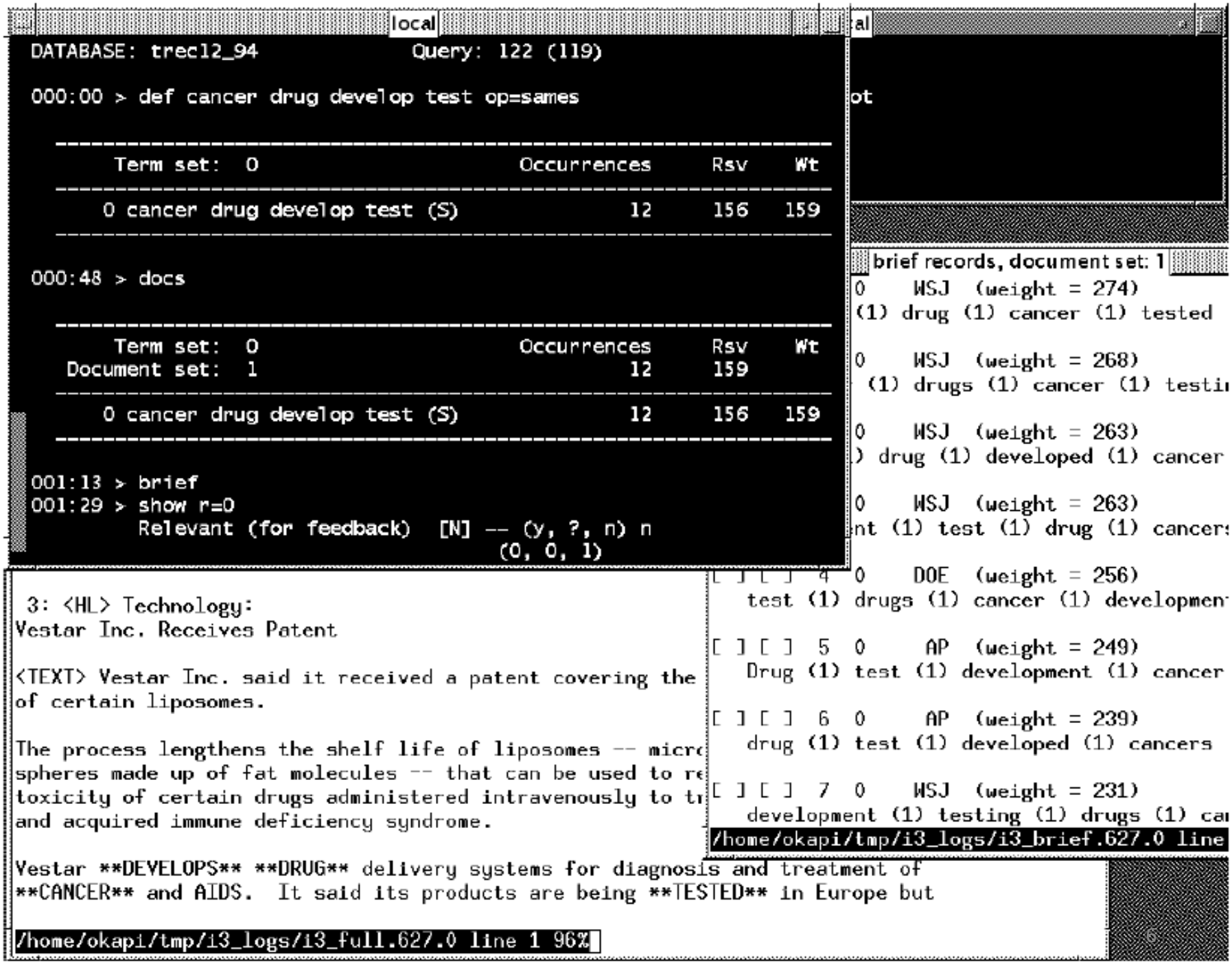
A 30 year old male that after a soccer game came with swelling of the knee. A meniscal tear was suspected. The MRI confirmed the lesion and also showed important swelling within the knee. The appearance of any structure is easily disclosed in MRI. Here you can actually *see* the bones, ligaments, soft tissues and the fluid collections that appears bright and at surrounds the knee.



The *official* report included this: T2 coronal Se (spin echo) sequence of the knee. The bright (white) rounded images that surround the knee is fluid related to synovitis or inflamation of the bursaes of the knee in a patient with a sport-related injury.

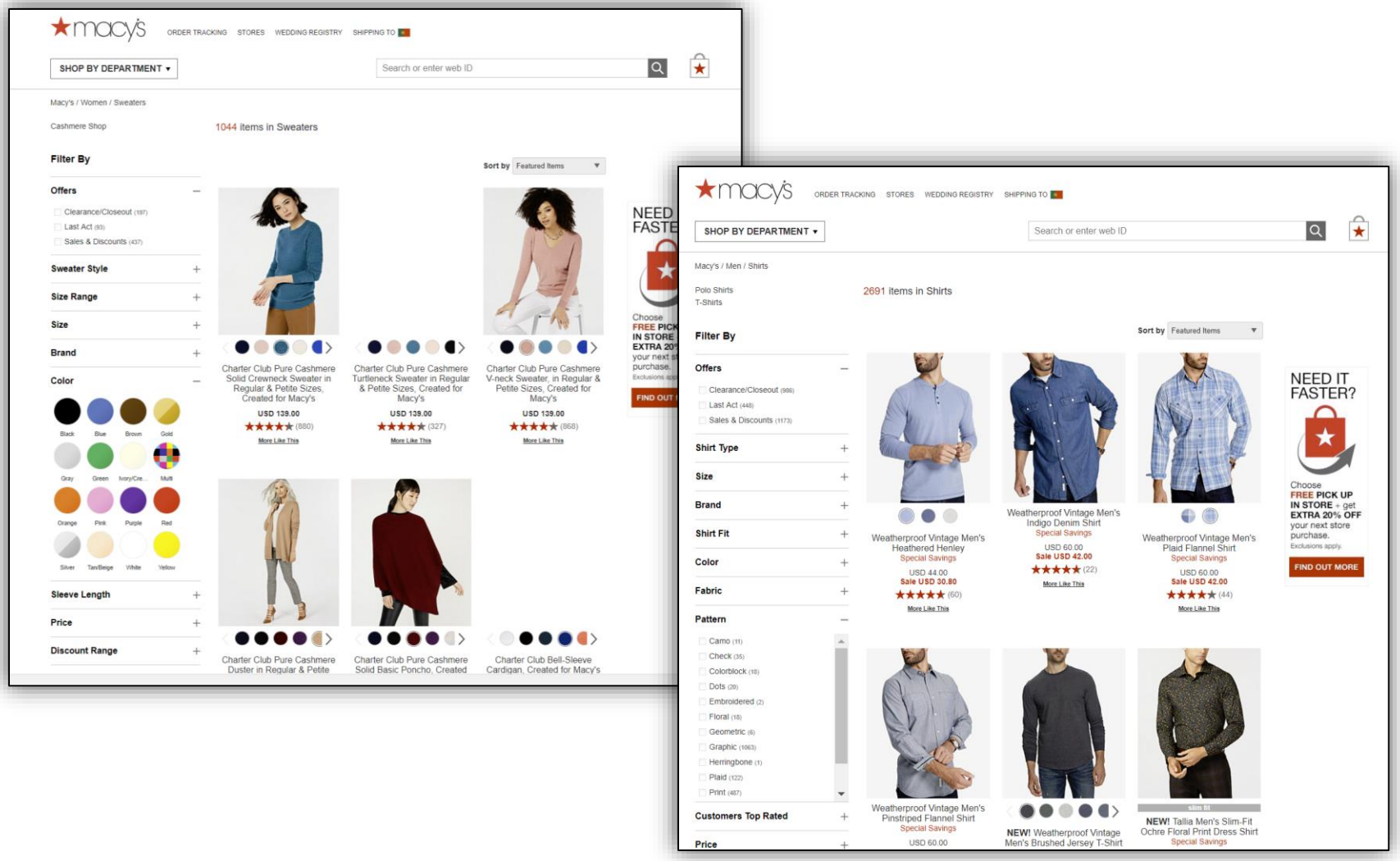
How to search multimedia information?

- **Richness of multimedia information**
- **Expressiveness of the user query**

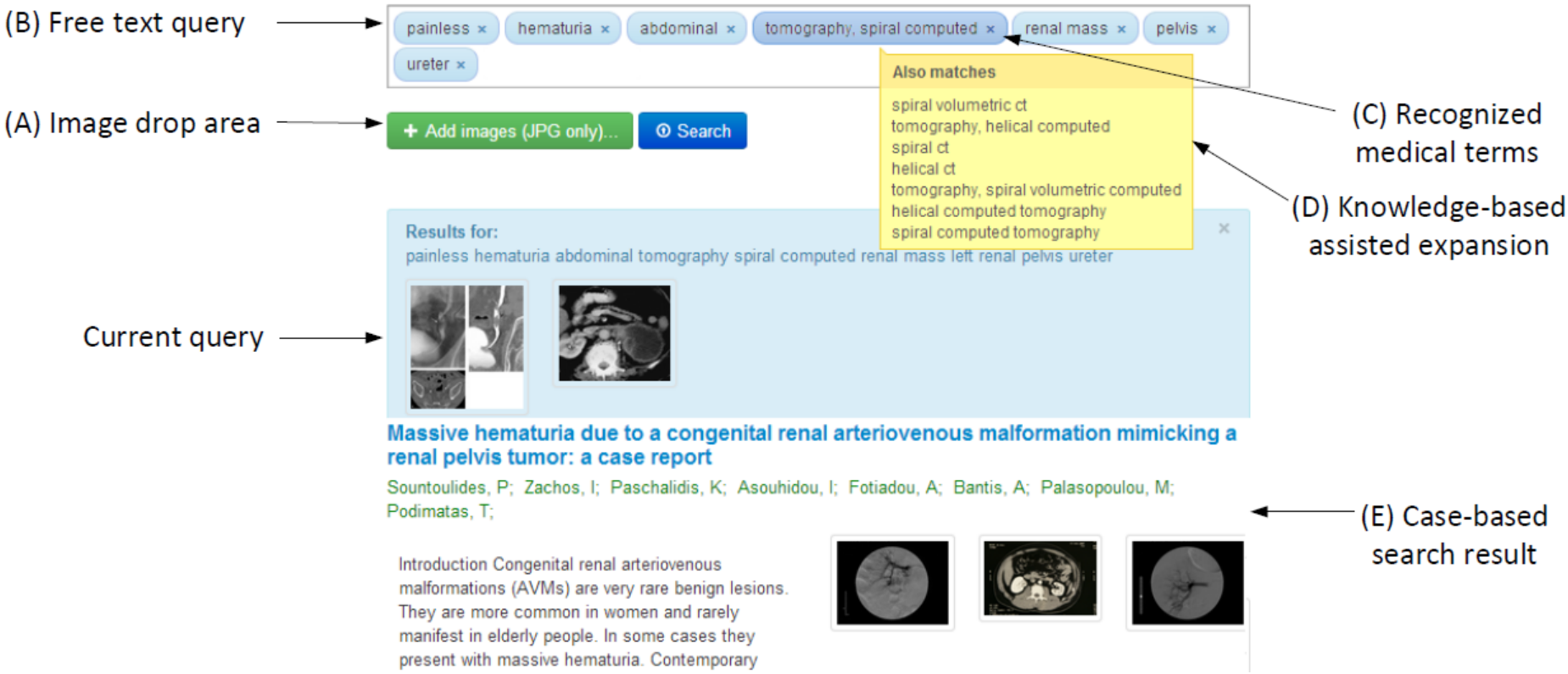

local
DATABASE: trec12_94 Query: 122 (119)
000:00 > def cancer drug develop test op=sames
Term set: 0 Occurrences Rsv Wt
0 cancer drug develop test (S) 12 156 159
000:48 > docs
Term set: 0 Occurrences Rsv Wt
Document set: 1 12 159
0 cancer drug develop test (S) 12 156 159
001:13 > brief
001:29 > show r=0
Relevant (for feedback) [N] -- (y, ?, n) n
(0, 0, 1)
brief records, document set: 1
0 WSJ (weight = 274)
(1) drug (1) cancer (1) tested
0 WSJ (weight = 268)
(1) drugs (1) cancer (1) testi
0 WSJ (weight = 263)
) drug (1) developed (1) cancer
0 WSJ (weight = 263)
nt (1) test (1) drug (1) cancer
[ ] [ ] 4 0 DOE (weight = 256)
test (1) drugs (1) cancer (1) developmen
[ ] [ ] 5 0 AP (weight = 249)
Drug (1) test (1) development (1) cancer
[ ] [ ] 6 0 AP (weight = 239)
drug (1) test (1) developed (1) cancers
[ ] [ ] 7 0 WSJ (weight = 231)
development (1) testing (1) drugs (1) ca
/home/okapi/tmp/i3_logs/i3_brief.627.0 line
3: <HL> Technology:
Vestar Inc. Receives Patent
<TEXT> Vestar Inc. said it received a patent covering the
of certain liposomes.
The process lengthens the shelf life of liposomes -- micr
spheres made up of fat molecules -- that can be used to re
toxicity of certain drugs administered intravenously to tr
and acquired immune deficiency syndrome.
Vestar **DEVELOPS** **DRUG** delivery systems for diagnosis and treatment of
CANCER and AIDS. It said its products are being **TESTED** in Europe but
/home/okapi/tmp/i3_logs/i3_full.627.0 line 1 96%

Web data based search

Online shopping



Medical domain



What makes a good search application?

- **Efficiency**: application replies to user queries without noticeable delays.
  - 1 sec is the "limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer"
    - Miller, R. B. (1968). Response time in man-computer conversational transactions. *Proc. AFIPS Fall Joint Computer Conference* Vol. 33, 267-277.

- **Effectiveness**: application replies to user queries with relevant answers.
  - This depends on the interpretation of the user query and the stored information.

Information extraction

- This stage deals with the extraction of the information to be made searchable
- Extract meaningful words, pairs of words or n-grams
- Extract images and their main characteristics
- Link visual characteristics and text data

This patient had a sudden loss of her motor functions (she wasnt able to move her right arms and legs) 2 months before the study. She went thru a slow recovery with lot physical therapy and drugs. She was recovering some of her movements but suddenly all the improvement stop. We performed an MRI that showed the changes expected for a lesion of that time (2 months old) but also showed and increase in the size of the ventricular system(where the Cerebrospinal fluid or CSF flows) that was causing hydrocephalus. Due to this finding, the patient went thru another surgery and had a shunt valve installed, the last word we had from one of her relatives is that she is again on recovery.

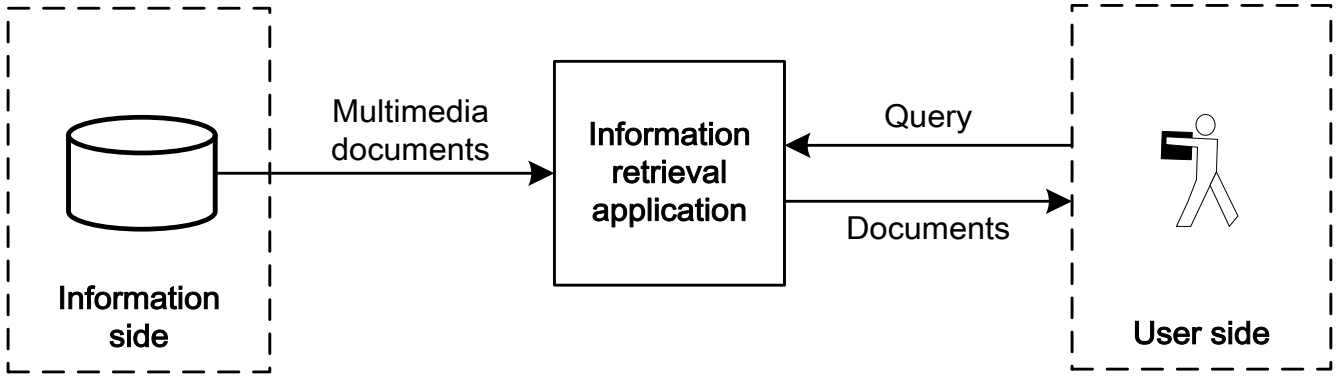
The *official* report included this: T 1 coronal SE (spin echo) sequence that shows an area of infarction in the left parietal lobe. Also enlargement of the ventricular system is observed.

?

Querying

- Conversion of the user query into the internal search space
  - Parsing
- Usage history
  - Cookies, profiles, etc.
- User intention
  - What type of task is the user doing?

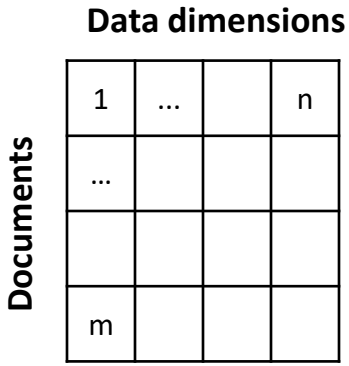
Relevance vs similarity



What is the best [search space + dissimilarity function] to compute the relevance of documents for a given user information need?

Indexing

- This stage creates an index to quickly locate relevant documents



- An index can be an agregation of several data structures (e.g. several B-trees)

- High-dimensional data can not be indexed by standard data structures, they require special hashing methods and data structures.

- The distribution of the index pages across a cluster improves the search engine responsiveness

Ranking and browsing

- Once the user query is converted into the internal search space...
  - The ranking function sorts the information according to its relevance to the user query
- Ranking functions should model the human notion of relevance
  - We don’t really know the mathematical form of the human notion of similarity.
- Browsing similar data requires specific algorithms for matching information on the target search space.

Course program

- Part 1 – Metric spaces and efficient search
  - Social-media data representation
  - Hashing similar documents
- Part 2 – Web data categorization and recognition
  - Information categorization
  - Information extraction
- Part 3 – Graphs
  - PageRank
  - Graph mining
- Part 4 – Learning embeddings
  - Recommendation
  - Word embeddings
  - Cross-modal spaces

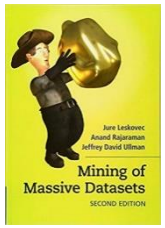
Course plan

Web Search				
Week	#	Lecture		In-class labs
10-Sep-18	1	Introduction		
17-Sep-18	2	Social-media data representation	Inf. Extraction	Environment setup + project introduction
24-Sep-18	3	Hashing similar documents		Project
01-Oct-18	4	Information categorization		Project
08-Oct-18	5	Information extraction		Project
15-Oct-18	6	Web graph analysis		**Checkpoint 1**
22-Oct-18		**Checkpoint 1 discussion**	Graphs	Project
29-Oct-18	7	Mining data graphs		Project
05-Nov-18	8	Recommendation algorithms		Project
12-Nov-18	9	Word embeddings		Project
19-Nov-18	10	Cross-modal search spaces		**Checkpoint 2**
26-Nov-18		**Checkpoint 2 discussion**	Embeddings	Project
03-Dec-18	11	Case study: Dbpedia entity linking		Project
10-Dec-18		Test		Project
17-Dec-18		-		**Project submission**

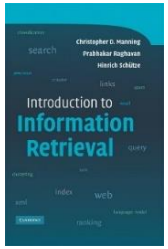
References

- Slides and articles provided during classes.

- Books:



Jure Leskovec, Anand Rajaraman, Jeff Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2011.



C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008.

Course grading

- 40% theoretical part (1 test or 1 exam)
- 60% for a 3 parts project (groups of 2 to 3 students):
  - 25% Submission 1
  - 25% Submission 2
  - 50% Final submission
- Additional rules:
  - Minimum individual grade: 8
  - Minimum grade on the labs or theory: 9
  - You may use one sided A4 sheet handwritten by you with your notes
    - It must be handed at the end of the test.

Project grading

- Scoring:
  - Implement. correctness 30%
  - Results analysis 30%
  - Critical discussion 40%
- Report:
  - Maximum of 8 pages.
  - No cover page.
  - Must include graphs, tables, etc.

- Report organization:
  - Introduction
  - Algorithms
  - Implementation
  - Evaluation
    - Dataset description
    - Baselines
    - Results analysis
  - Critical discussion
  - References

Summary

- Web Search course context
- Course objectives and plan
- Grading
- Labs