

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](#) [X](#) [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](#)

Tags

- [Arizona](#) ×
- [Reflection](#) ×
- [Desert](#) ×

[Add a tag](#)

Comments



[alexei_322](#) pro says:

thumbs up! great shot

Posted 31 months ago. ([permalink](#) | [delete](#))

Add your comment

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 wasn't able to move her right arms and legs) 2 months before the study. She went through a slow recovery with a lot of physical therapy and drugs. She was recovering some of her movements but suddenly all the improvement stopped. We performed an MRI that showed the changes expected for a lesion of that time (2 months old) but also showed an 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 through 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: T1 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 appear bright and surround 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 are fluid related to synovitis or inflammation of the bursae 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**

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)

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 -- micro-
spheres made up of fat molecules -- that can be used to re-
toxicity of certain drugs administered intravenously to treat
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%

ot

brief records, document set: 1

0 WSJ (weight = 274)

(1) drug (1) cancer (1) tested

0 WSJ (weight = 268)

(1) drugs (1) cancer (1) testin

0 WSJ (weight = 263)

) drug (1) developed (1) cancer

0 WSJ (weight = 263)

nt (1) test (1) drug (1) cancer:

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

Web data based search

The image shows a composite screenshot of a web search interface. At the top left is the Yahoo! logo with navigation links like 'NEW', 'COOL', and 'RANDOM'. Below it is a search bar with 'Search' and 'Options' buttons. A vertical list of categories is on the left, including Arts, Business and Economy, Computers and Internet, Education, Entertainment, Government, Health, News, Recreation and Sports, Reference, Regional, and Science. To the right is the AltaVista logo with a search bar containing 'the Web' and a 'Submit' button. Below that is a Google! BETA logo with a search bar containing 'the Web' and 'Google Search' and 'I'm feeling lucky' buttons. At the bottom are links for 'Special Searches' (Stanford, Linux), 'Help' (About Google!, Company Info, Logos), and a subscription form with 'Subscribe' and 'Archive' buttons. The copyright notice 'Copyright ©1998 Google Inc.' is at the very bottom.

- [Arts](#) -- *Humanities, Photography, Architecture, ...*
- [Business and Economy \[Xtra!\]](#) -- *Directory, Investments, Classifieds, ...*
- [Computers and Internet \[Xtra!\]](#) -- *Internet, WWW, Software, Multimedia, ...*
- [Education](#) -- *Universities, K-12, Courses, ...*
- [Entertainment \[Xtra!\]](#) -- *TV, Movies, Music, Magazines, ...*
- [Government](#) -- *Politics [Xtra!], Agencies, Law, Military, ...*
- [Health \[Xtra!\]](#) -- *Medicine, Drugs, Diseases, Fitness, ...*
- [News \[Xtra!\]](#) -- *World [Xtra!], Daily, Current Events, ...*
- [Recreation and Sports](#)
- [Reference](#) -- *Libraries*
- [Regional](#) -- *Countries*
- [Science](#) -- *CS, Biology*

Search and Display the Results in

Search Submit

Search the web using Google!

Special Searches
[Stanford Search](#)
[Linux Search](#)

Help
[About Google!](#)
[Company Info](#)
[Google! Logos](#)

Get Google!
updates monthly.

 [Archive](#)

Copyright ©1998 Google Inc.

Online shopping

The screenshot shows the Macy's website interface for women's sweaters. At the top, there is a navigation bar with the Macy's logo, links for 'ORDER TRACKING', 'STORES', 'WEDDING REGISTRY', and 'SHIPPING TO'. Below this is a search bar with the placeholder text 'Search or enter web ID' and a 'SHOP BY DEPARTMENT' dropdown menu. The main content area is titled 'Macy's / Women / Sweaters' and 'Cashmere Shop' with a sub-header '1044 items in Sweaters'. A 'Filter By' sidebar on the left includes sections for 'Offers' (Clearance/Closeout, Last Act, Sales & Discounts), 'Sweater Style', 'Size Range', 'Size', 'Brand', 'Color' (with a color palette), 'Sleeve Length', 'Price', and 'Discount Range'. The main product grid displays three items: 'Charter Club Pure Cashmere Solid Crewneck Sweater in Regular & Petite Sizes, Created for Macy's' (USD 139.00, 5 stars, 880 reviews), 'Charter Club Pure Cashmere Turtleneck Sweater in Regular & Petite Sizes, Created for Macy's' (USD 139.00, 5 stars, 327 reviews), and 'Charter Club Pure Cashmere V-neck Sweater in Regular & Petite Sizes, Created for Macy's' (USD 139.00, 5 stars, 958 reviews). Each item includes a product image, a color selection palette, and a 'More Like This' link. A 'NEED IT FASTER?' banner is visible on the right side of the page.

The screenshot shows the Macy's website interface for men's shirts. At the top, there is a navigation bar with the Macy's logo, links for 'ORDER TRACKING', 'STORES', 'WEDDING REGISTRY', and 'SHIPPING TO'. Below this is a search bar with the placeholder text 'Search or enter web ID' and a 'SHOP BY DEPARTMENT' dropdown menu. The main content area is titled 'Macy's / Men / Shirts' and 'Polo Shirts' with a sub-header '2691 items in Shirts'. A 'Filter By' sidebar on the left includes sections for 'Offers' (Clearance/Closeout, Last Act, Sales & Discounts), 'Shirt Type', 'Size', 'Brand', 'Shirt Fit', 'Color', 'Fabric', 'Pattern', 'Customers Top Rated', and 'Price'. The main product grid displays three items: 'Weatherproof Vintage Men's Heathered Henley Special Savings' (USD 44.00, Sale USD 30.80, 5 stars, 60 reviews), 'Weatherproof Vintage Men's Indigo Denim Shirt Special Savings' (USD 60.00, Sale USD 42.00, 5 stars, 22 reviews), and 'Weatherproof Vintage Men's Plaid Flannel Shirt Special Savings' (USD 60.00, Sale USD 42.00, 5 stars, 44 reviews). Each item includes a product image, a color selection palette, and a 'More Like This' link. A 'NEED IT FASTER?' banner is visible on the right side of the page.

Medical domain

(B) Free text query →

(A) Image drop area →

Current query →

(C) Recognized medical terms

(D) Knowledge-based assisted expansion

(E) Case-based search result

The screenshot displays a medical search interface. At the top, a search bar contains several tags: 'painless x', 'hematuria x', 'abdominal x', 'tomography, spiral computed x', 'renal mass x', and 'pelvis x'. Below the search bar is a green button labeled '+ Add images (JPG only)...' and a blue 'Search' button. A yellow tooltip box titled 'Also matches' lists related terms: 'spiral volumetric ct', 'tomography, helical computed', 'spiral ct', 'helical ct', 'tomography, spiral volumetric computed', 'helical computed tomography', and 'spiral computed tomography'. Below the search bar, the 'Results for:' section shows the query text: 'painless hematuria abdominal tomography spiral computed renal mass left renal pelvis ureter'. Two medical images are displayed in a grid. Below the images, a case report is shown with the title 'Massive hematuria due to a congenital renal arteriovenous malformation mimicking a renal pelvis tumor: a case report'. The authors listed are Sountoulides, P; Zachos, I; Paschalidis, K; Asouhidou, I; Fotiadou, A; Bantis, A; Palasopoulou, M; and Podimatas, T. The introduction text reads: 'Introduction Congenital renal arteriovenous malformations (AVMs) are very rare benign lesions. They are more common in women and rarely manifest in elderly people. In some cases they present with massive hematuria. Contemporary'. Three small medical images are shown at the bottom right of the case report.

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 wasn't able to move her right arms and legs) 2 months before the study. She went through a slow recovery with a lot of physical therapy and drugs. She was recovering some of her movements but suddenly all the improvement stopped. We performed an MRI that showed the changes expected for a lesion of that time (2 months old) but also showed an 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 through 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.

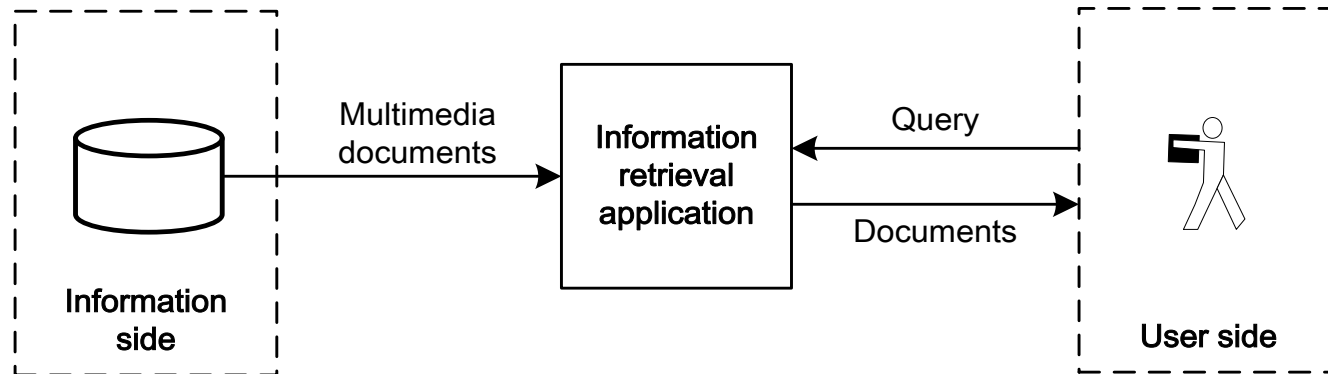


?

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 aggregation 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

Data dimensions

	1	...		n
	...			
Documents	m			

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

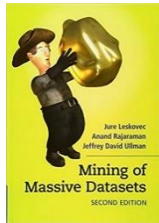
- Part 1 – Metric spaces and efficient search
 - Social-media data representation
 - Hashing similar documents
- Part 2 – Web data categorization and recognition
 - Information categorization
 - Information extraction
- Part 3 – Graphs
 - PageRank
 - Graph mining
- Part 4 – Learning embeddings
 - Recommendation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

Course plan

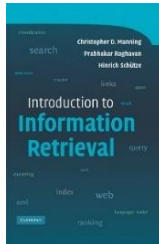
Web Search				
Week	#	Lecture	In-class labs	
10-Sep-18	1	Introduction		
17-Sep-18	2	Social-media data representation	Inf. Extraction	Environment setup + project introduction
24-Sep-18	3	Hashing similar documents		Project
01-Oct-18	4	Information categorization		Project
08-Oct-18	5	Information extraction		Project
15-Oct-18	6	Web graph analysis		Checkpoint 1
22-Oct-18		Checkpoint 1 discussion		
29-Oct-18	7	Mining data graphs	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		
03-Dec-18	11	Case study: Dbpedia entity linking	Embeddings	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