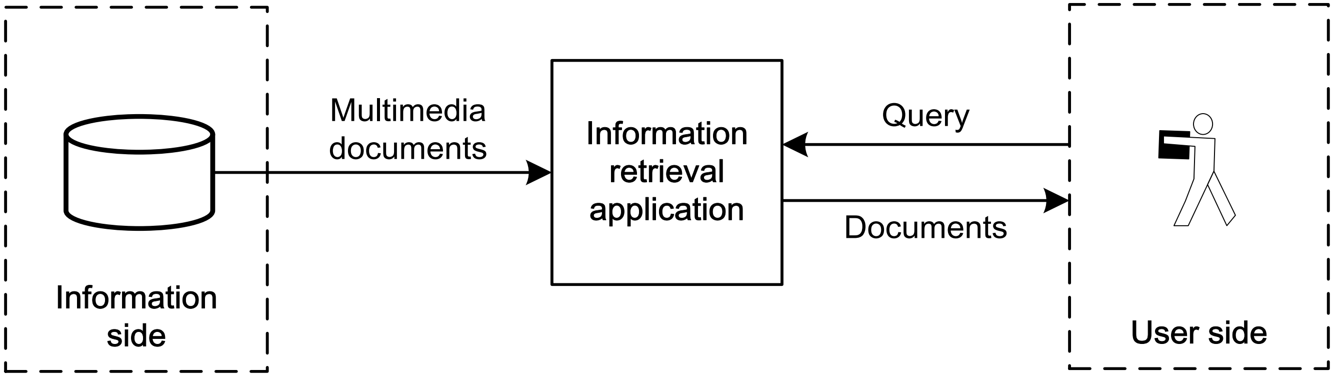
Information Retrieval

Course presentation

João Magalhães

Relevance vs similarity



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

What makes a good search application?

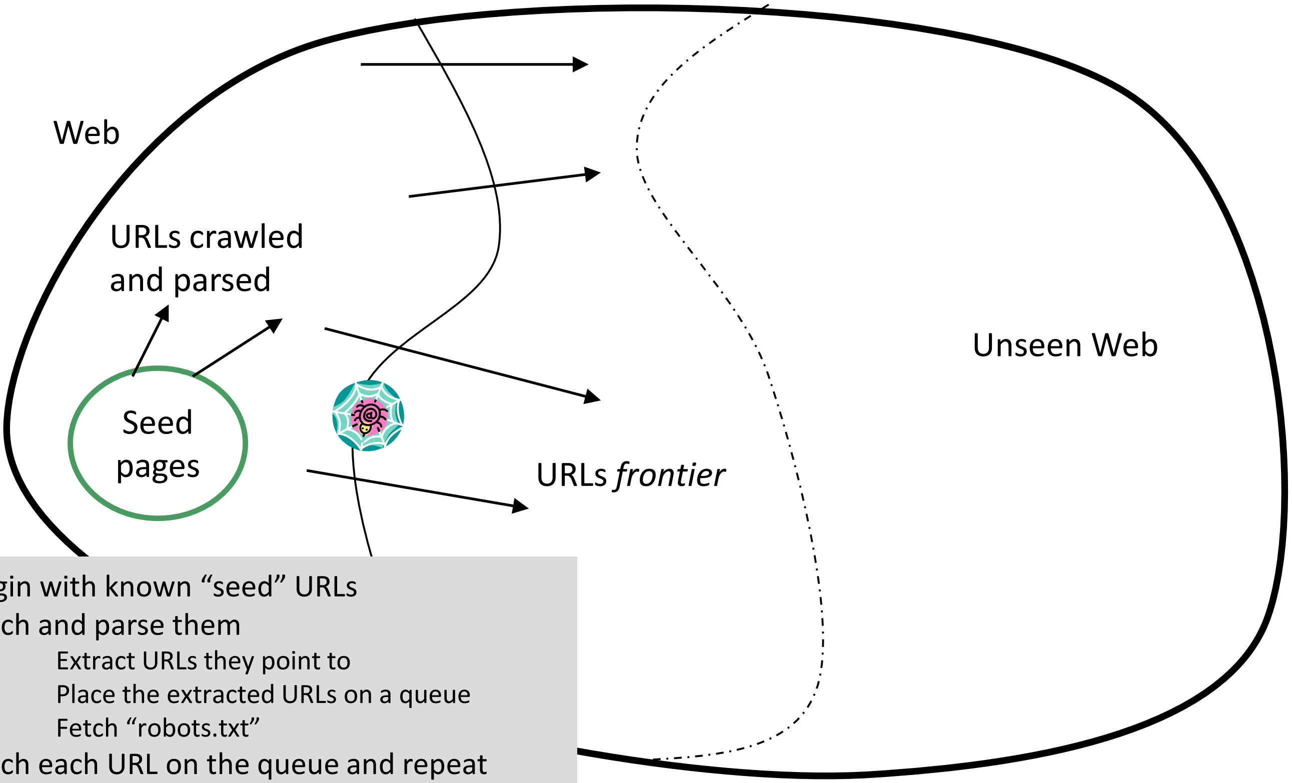
- **Efficiency**: application replies to user queries without noticeable delays.
  - 1 sec is the “limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer”
    - Miller, R. B. (1968). Response time in man-computer conversational transactions. *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.

The tasks of a search application

- **Collect** data for storage
  - Crawler
- Analyse collected data and compute the **relevant information**
  - Information analysis
- Store data in an **efficient** manner
  - Indexing
- Process **user** information needs
  - Querying
- Find the documents that best **match** the user information need
  - Ranking

Web crawling



Begin with known “seed” URLs

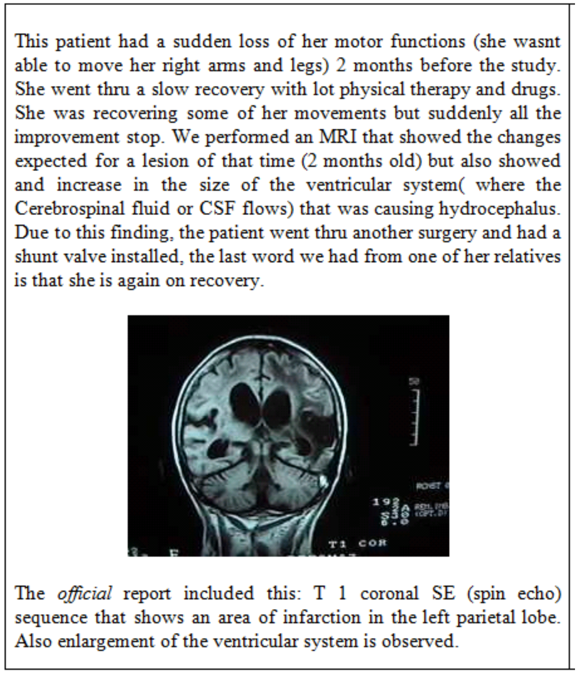
Fetch and parse them

- Extract URLs they point to
- Place the extracted URLs on a queue
- Fetch “robots.txt”

Fetch each URL on the queue and repeat

Information analysis

- 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.

Indexing

- This stage creates an index to quickly locate relevant documents

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

- Index compression is used to reduce the amount of space and the time needed to compute similarities

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

Querying

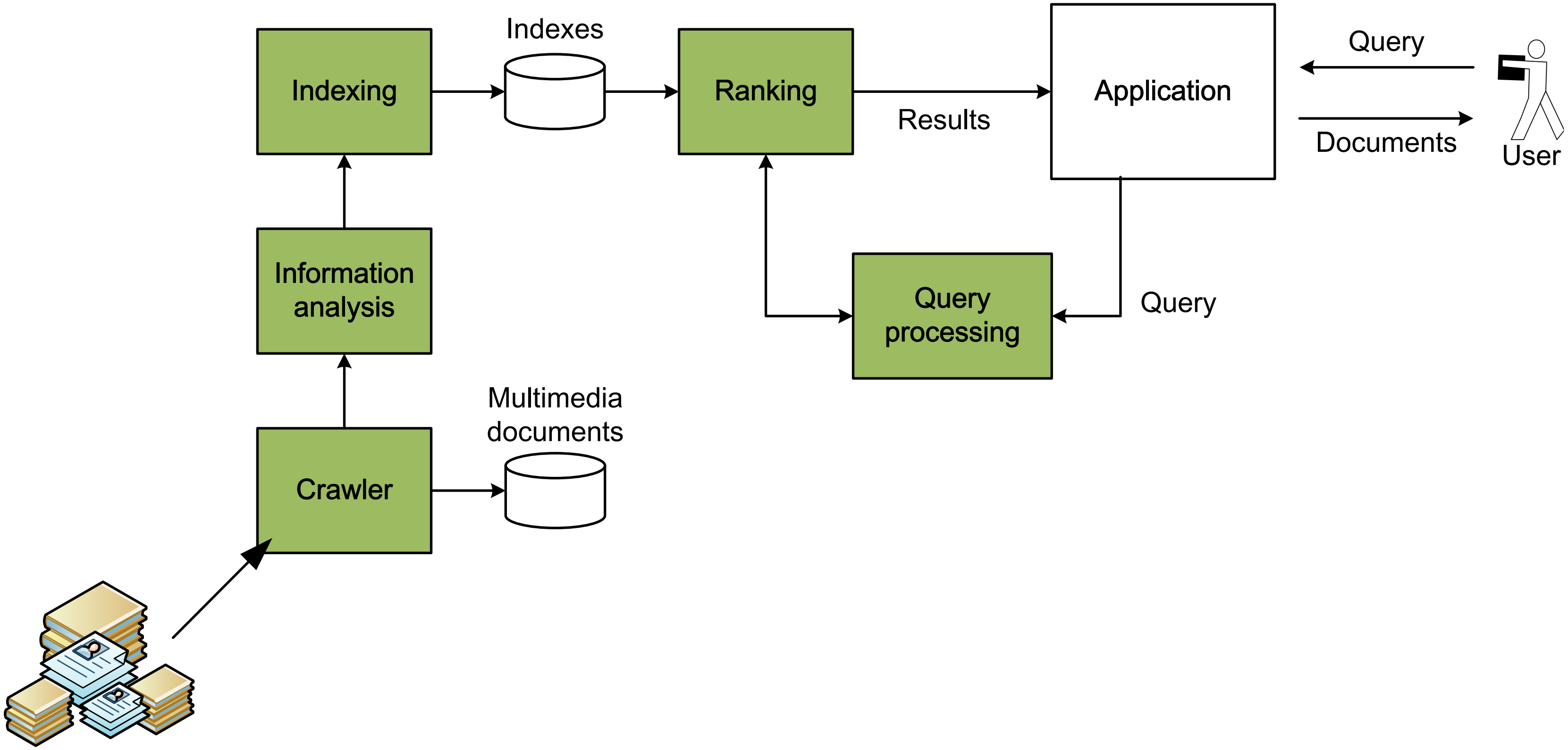
- Conversion of the user query into the internal search space
    - Parsing
- Usage history
    - Cookies, profiles, etc.
- User intention
    - What type of task is the user doing?

Ranking

- 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...

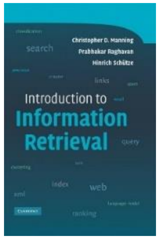
Putting all together...



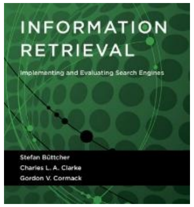
References

- Slides and articles provided during classes.

- Books:



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



Stefan Buettcher, Charles L. A. Clarke, Gordon V. Cormack, “Information Retrieval: Implementing and Evaluating Search Engines”, The MIT Press, 2010.

Course grading

- The course has two mandatory components:
  - Theoretical part (1 test or 1 exam): 40% **(minimum grade > 9.0)**
  - Labs (groups of 3 students): 60% **(minimum grade > 9.0)**

- Theory test/exam:
  - Test: 12 December
  - Exam: date to be defined

- Additional rules:
  - You may use one sided A4 sheet handwritten by you with your notes.
  - It must be handed at the end of the test.

- Individual mini-lab grading **(minimum grade > 8.0)**
  - 30% implementation + 20 % report + 20% questions + 30% discussion

Laboratories: News search

- Implement a search engine to search online news.
- Understand the roles of each component of a search engine in the performance of the search results.
- Labs are done incrementally. Each week new functionalities will be added to the initial implementation.
- There will be 4 mini-labs throughout the semester.
  - The submission date of each mini-lab is three days after the last lab class of the corresponding mini-lab.

Schedule

Information Retrieval				
Week	#	Lectures	In-class labs	
12-Sep-18	1	Introduction		
19-Sep-18	2	Basic techniques (Lucene examples)	Lab 1	Environment setup
26-Sep-18	3	Evaluation		Text pre-processing, VSM
03-Oct-18	4	Retrieval models: LM + BIM + BM25		Evaluation scripts
10-Oct-18	5	Implementation of Ret Models	Lab 2	Retrieval models
17-Oct-18	6	Query processing and taxonomies		Retrieval models
24-Oct-18		**Reports discussion**	Lab 3	Query expansion
31-Oct-18	7	Information duplicates		Query expansion
07-Nov-18	8	Multiple fields and rank fusion		Query expansion
14-Nov-18	9	-	Lab 4	Ranking multiple fields
21-Nov-18	10	Static and distributed indexing		Ranking multiple fields
28-Nov-18	11	Efficient query processing		Ranking multiple fields
05-Dec-18	12	Elasticsearch vs Lucene		Ranking multiple fields
12-Dec-18		**Test + Reports discussion**		

Summary

- “Information Retrieval” course context
- Course objectives and plan
- Grading
- Labs