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...

Comments

alexei_322, pro says:
thumbs up! great shot
Posted 31 months ago (permalink | delete)

Add your comment

Tags

- Arizona
- Reflection
- Desert

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 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 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.

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 inflammation of the bursae of the knee in a patient with a sport-related injury.

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.
How to search multimedia information?

• Richness of multimedia information

• Expressiveness of the user query
Vestar Inc. said it received a patent covering the development of certain liposomes. The process lengthens the shelf life of liposomes -- microscopic spheres made up of fat molecules -- that can be used to reduce the toxicity of certain drugs administered intravenously to treat cancer 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...
Web data based search
Online shopping
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”

- **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
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
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

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

- **Web Search**
- **Graphs**
- **Embeddings**
- **Inf. Extraction**
References

• Slides and articles provided during classes.

• Books:


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