270028 - CAIM - Massive Information Search and Analysis

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science
Academic year: 2018
Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2010). (Teaching unit Optional)
ECTS credits: 6  
Teaching languages: Catalan, Spanish

Teaching staff

Coordinator: - Ramon Ferrer Cancho (rferrericancho@cs.upc.edu)
Others: - Javier Béjar Alonso (bejar@cs.upc.edu)
- Marta Arias Vicente (marias@cs.upc.edu)
- Ricard Gavaldà Mestre (gavalda@cs.upc.edu)

Prior skills

In general, all those that are acquired in the required prior courses.
Specifically:
- To know and use comfortably basic concepts of linear algebra, discrete mathematics, probability and statistics.
- To program comfortably in object-oriented languages, including inheritance between classes.
- To know the main data structures to access information efficiently and their implementations (lists, hashing, trees, graphs, heaps). To be able to use them to build efficient programs. To be able to analyze the execution time and memory used by an algorithm of average difficulty. To have an idea of the difference in time to access main memory and disk.
- To know the main elements of a relational database and SQL-like access language.

Requirements

- Prerequisite PE
- Prerequisite BD
- Corequisite PROP
Learning objectives of the subject

1. Understand the problems associated with storage and information retrieval, in particular with information in textual form.
2. Understand that effective search and information retrieval is closely related to the organization and description of this information.
3. To know and understand the structure, architecture and functioning of the web, and elements related to it: indices, search engines, crawlers, among others.
4. To know and understand the descriptive parameters of complex networks and the algorithms to analyze their structure.
5. Recognizing the opportunities for using massive information to an organization’s goals, and choose the most appropriate methods, tools, and procedures.
6. Be able to decide the information retrieval techniques that may be effective in a specific information system, especially those of textual type.
7. Be able to evaluate the effectiveness and usefulness of an information retrieval system, according to several criteria.
8. To implement the main techniques learned during the course.
9. Know how to use, adapt and extend open-source software.
# Study load

<table>
<thead>
<tr>
<th>Total learning time: 150h</th>
<th>Theory classes:</th>
<th>15h</th>
<th>10.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Practical classes:</td>
<td>15h</td>
<td>10.00%</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes:</td>
<td>30h</td>
<td>20.00%</td>
</tr>
<tr>
<td></td>
<td>Guided activities:</td>
<td>6h</td>
<td>4.00%</td>
</tr>
<tr>
<td></td>
<td>Self study:</td>
<td>84h</td>
<td>56.00%</td>
</tr>
</tbody>
</table>
## Content

### Introduction

**Degree competences to which the content contributes:**

**Description:**

Need of search and analysis techniques of massive information. Search and analysis vs. databases. Information retrieval process. Preprocessing and lexical analysis.

### Models of information retrieval

**Degree competences to which the content contributes:**

**Description:**


### Implementation: Indexing and searching

**Degree competences to which the content contributes:**

**Description:**

Inverse and signature files. Index compression. Example: Efficient implementation of the rule of the cosine measure with tf-idf. Example: Lucene.

### Evaluation in information retrieval

**Degree competences to which the content contributes:**

**Description:**

Recall and precision. Other performance measures. Reference collections. Relevance feedback and query expansion.

### Web search

**Degree competences to which the content contributes:**

**Description:**


### Architecture of massive information processing systems

**Degree competences to which the content contributes:**
### Network analysis

**Degree competences to which the content contributes:**

**Description:**
Descriptive parameters and characteristics of networks: degree, diameter, small-world networks, among others. Algorithms on networks: clustering, community detection and detection of influential nodes, reputation, among others.

### Information Systems based on massive information analysis. Combination with other technologies.

**Degree competences to which the content contributes:**

**Description:**
## Planning of activities

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours: 26h 30m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction and models of information retrieval</td>
<td>Theory classes: 5h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 6h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 13h 30m</td>
</tr>
</tbody>
</table>

**Description:**
2 theory hours, 2 problem hours, and 4 lab hours on the topics "Introduction" and "Models of information retrieval". See the description in the Teaching Methodology section.

**Specific objectives:**
1, 2, 6

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours: 13h 30m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation and evaluation</td>
<td>Theory classes: 3h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 6h 30m</td>
</tr>
</tbody>
</table>

**Description:**
2 theory hours, 2 problem hours, and 4 lab hours on the topics "Implementation" and "Evaluation". See the description in the Teaching Methodology section.

**Specific objectives:**
2, 7, 8, 9

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours: 24h 30m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching the web</td>
<td>Theory classes: 4h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 6h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 12h 30m</td>
</tr>
</tbody>
</table>

**Description:**
2 theory hours, 2 problem hours, and 4 lab hours on the topic "Web search". See the description in the Teaching Methodology section.

**Specific objectives:**
3, 5, 9

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours: 2h</th>
</tr>
</thead>
<tbody>
<tr>
<td>First partial exam</td>
<td>Guided activities: 2h</td>
</tr>
<tr>
<td></td>
<td>Self study: 0h</td>
</tr>
</tbody>
</table>

**Description:**
Partial exam of the first part of the course.
### Specific objectives:
1, 2, 3, 5, 6, 7

| Architecture of web search systems | Hours: 24h 30m  
Theory classes: 4h  
Practical classes: 2h  
Laboratory classes: 6h  
Guided activities: 0h  
Self study: 12h 30m |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td>2 theory hours and 6 lab hours on the topic &quot;Architecture&quot;. See the description in the Teaching Methodology section.</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>3, 6, 8, 9</td>
</tr>
</tbody>
</table>

| Network analysis                  | Hours: 22h 30m  
Theory classes: 4h  
Practical classes: 2h  
Laboratory classes: 4h  
Guided activities: 0h  
Self study: 12h 30m |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td>4 theory hours and 6 lab hours on the topic &quot;Network Analysis&quot;. See the description in the Teaching Methodology section.</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>4, 6, 7, 8, 9</td>
</tr>
</tbody>
</table>

| Information systems based on massive information analysis | Hours: 21h 30m  
Theory classes: 3h  
Practical classes: 2h  
Laboratory classes: 4h  
Guided activities: 0h  
Self study: 12h 30m |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td>Theory, problems, and labs on this topic. The emphasis is on practical cases in the problems and lab sessions. See the description in the Teaching Methodology section.</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>5, 6, 7, 9</td>
</tr>
</tbody>
</table>
270028 - CAIM - Massive Information Search and Analysis

<table>
<thead>
<tr>
<th>Second partial exam or final exam</th>
<th>Hours: 15h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description:</td>
<td></td>
</tr>
<tr>
<td>The student chooses between an exam of the second part of the course or an exam on the whole course.</td>
<td></td>
</tr>
<tr>
<td>Specific objectives:</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3, 4, 5, 6, 7, 8, 9</td>
<td></td>
</tr>
</tbody>
</table>

The course will include the following evaluation events:

- Reports of laboratory sessions, which will be delivered within a time limit for each session (generally around 2 weeks). We will compute a weighted average from the grade of these laboratory reports, which we refer to as L.

- A mid-term exam, covering material seen until the exam is done. Let P1 be the grade obtained in this exam.

- The day of the final exam, each student will choose between two options mutually exclusive: 1) do a second partial exam covering what was not covered in the mid-term exam (we refer to the grade of this exam as P2), or 2) do a final exam covering the whole course (whose grade we refer as F). There is no requirement in the grade of P1 to chose between options 1) and 2).

The four grades L, P1, P2, and F are between 0 and 10. The final grade is computed as:

$$0.4 \times L + \max(0.3 \times P1 + 0.3 \times P2, 0.6 \times F).$$

As to the competency grade associated to Autonomous Learning, it will be computed as follows:

- For the i-th laboratory report submitted: the value Ri will be 1 if the report has been submitted within the deadline and enough effort has been put into the report. Ri will be 0 otherwise. Let Rsum be the sum of all the individual Ri values (which can add up to k if there are k lab sessions).

- Some questions in partial or final exams, marked appropriately, will focus on topics that are only partially covered during theory and problem-solving sessions, and that therefore students must prepare on their own. Let E be the weighted average of such questions, scaled to the interval [0.1].

Let S be the value of \((Rsum/k + E)/2\), which lies between 0 and 1.

The competency grade is:

- D if S is less than 0.5
- C if S lies between 0.5 and 0.599
- B if S lies between 0.6 and 0.799
- A if S is 0.8 or more.
Bibliography

Basic:


Others resources:

Hyperlink

https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/?refid