UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Course guide
270028 - CAIM - Massive Information Search and Analysis

Last modified: 13/07/2023

Unit in charge: Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science.

Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2010). (Optional subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan, Spanish

LECTURER

Coordinating lecturer: RAMON FERRER CANCHO

Others: Primer quadrimestre:
ALBERT CALVO IBAÑEZ - 11, 12
RAMON FERRER CANCHO - 11, 12, 13
IGNASI GÓMEZ SEBASTIÀ - 13

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 BD
- Prerequisite PE
- Corequisite PROP

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CCO2.5. To implement information retrieval software.
CSI2.3. To demonstrate knowledge and application capacity of extraction and knowledge management systems .
CSI2.6. To demonstrate knowledge and capacity to apply decision support and business intelligence systems.

Generical:
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

TEACHING METHODOLOGY

- Theory lectures. Before each class, students must have read the notes and materials on the topic to be discussed in class, which will be announced with enough time to prepare. Students will also have at their disposal a questionnaire with basic questions to see if a basic degree of understanding has been reached. In class, the teacher will present the main points, assuming that the student has done the job indicated and has tried to answer the questionnaire; difficulties found by students will be discussed in class collectively.

- Problem-solving sessions. Teachers and students will discuss and compare the solutions to problems provided by the teacher with sufficient time before each class. Discussions can be made collectively in class or individually between teacher and student. The teacher will assume that the students have spent a reasonable amount of time trying to solve these exercises, and priority will be given to those who have done so.

- Laboratory sessions. Before each class, students are assumed to have read the script of practical work to be developed during the session. During class, students will do the work specified in the script with the guidance of the teacher. In many cases, students will probably need extra time to finish the work. For most lab sessions the students will have to write a short report and/or deliver files associated with it (output files and code).

- Personal work. Every type of classroom activity involves a certain amount of personal work. Additionally, some topic or topics of the course could have no theory classes or exercises associated; students must study these on their own, and can take advantage the directed activities' sessions to assess whether they have learnt them sufficiently or not.

Given that the course appears as part of two different degree specialties, the activities proposed (in theory, problem-solving or lab sessions) may slightly differ for different types of students, always ensuring fairness regarding difficulty or workload.

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 themain techniques learned during the course.
9.Know how to use, adapt and extend open-source software.

STUDY LOAD

Type	Hours	Percentage
Hours small group	30,0	20.00
Hours large group	15,0	10.00
Self study	84,0	56.00
Hours medium group	15,0	10.00
Guided activities	6,0	4.00

Total learning time: 150 h

CONTENTS

Introduction

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

Description:
Formal definition and basic concepts: abstract models of documents and query languages. Boolean model. Vector model. Latent Semantic Indexing.

Implementation: Indexing and searching

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

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

Web search

Description:
Ranking and relevance in the web. The PageRank algorithm. Crawling. Architecture of a simple web search system.

Architecture of massive information processing systems

Description:
Scalability, high performance, and fault tolerance: the case of massive web searchers. Distributed architectures. Example: Hadoop.

Network analysis

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.

Description:
Search Engine Optimization. Joint use of IR techniques with Data Mining and Machine Learning. Recommender Systems.

ACTIVITIES

Introduction and models of information retrieval

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

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 25h 30m
Theory classes: 4h
Practical classes: 2h
Laboratory classes: 6h
Self study: 13h 30m

Implementation and evaluation

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

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 14h
Theory classes: 2h 30m
Practical classes: 1h
Laboratory classes: 4h
Self study: 6h 30m

Searching the web

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

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 22h
Theory classes: 2h 30m
Practical classes: 1h
Laboratory classes: 6h
Self study: 12h 30m

First partial exam

Description:
Partial exam of the first part of the course.

Specific objectives:
1, 2, 3, 5, 6, 7

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 10h
Guided activities: 2h
Self study: 8h

Architecture of web search systems

Description:
2 theory hours and 6 lab hours on the topic "Architecture". See the description in the Teaching Methodology section.

Specific objectives:
3, 6, 8, 9

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 23h 30m
Theory classes: 3h 30m
Practical classes: 1h 30m
Laboratory classes: 6h
Self study: 12h 30m

Network analysis

Description:
4 theory hours and 6 lab hours on the topic "Network Analysis". See the description in the Teaching Methodology section.

Specific objectives:
4, 6, 7, 8, 9

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 21h 30m
Theory classes: 3h 30m
Practical classes: 1h 30m
Laboratory classes: 4h
Self study: 12h 30m

Information systems based on massive information analysis

Description:
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.

Specific objectives:
5, 6, 7, 9

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 18h 30m
Theory classes: 1h 30m
Practical classes: 0h 30m
Laboratory classes: 4h
Self study: 12h 30m

Second partial exam or final exam

Description:
The student chooses between an exam of the second part of the course or an exam on the whole course.

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 8, 9

Related competencies :
G7. AUTONOMOUS LEARNING: to detect deficiencies in the own knowledge and overcome them through critical reflection and choosing the best actuation to extend this knowledge. Capacity for learning new methods and technologies, and versatility to adapt oneself to new situations.

Full-or-part-time: 15h
Guided activities: 3h
Self study: 12h

GRADING SYSTEM

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*L + maximum(0.3*P1+0.3*P2, 0.6*F).

The choice of final exam implies P2 = 0. The choice of second parcial implies F = 0.
No attending the day of the final / second partial exam implies that the student has chosen second partial exam.

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:
- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Harlow [etc.]: Addison-Wesley / Pearson, 2011. ISBN 9780321416919.
- Manning, C.D.; Raghavan, P; Schütze, H. Introduction to information retrieval. New York: Cambridge University Press, 2008. ISBN 9780521865715.
- McCandless, M.; Hatcher, E.; Gospodnetic, O. Lucene in action. 2nd ed. Greenwich, Conn: Manning, 2010. ISBN 9781933988177.
- Croft, W.B.; Metzler, D.; Strohman, T. Search engines: information retrieval in practice. Pearson, 2010. ISBN 9780131364899.
- Russell, M.A.; Klassen, M. Mining the social web: data mining from Facebook, Twitter, and LindedIn, Instagram, GitHub, and more. 3rd ed. Sebastopol, [California]: O'Reilly Media, 2018. ISBN 9781491973509.

RESOURCES

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