Subject: Information Retrieval (IR)
Code: 32420
Institution: Escuela Politécnica Superior
Degree: Master's program in Research and Innovation in Information and Communications Technologies (I^2-ICT)
Level: Master
Type: Elective [computational intelligence]
ECTS: 6

COURSE GUIDE: Information Retrieval (IR)

Academic year: 2016-2017

Program: Master's program in Research and Innovation in Information and Communications Technologies (I^2-ICT)
Center: Escuela Politécnica Superior
University: Universidad Autónoma de Madrid

Last modified: 2016/05/10
Status: Approved 2014/06/10

1. ASIGNATURA / COURSE (ID)

Recuperación de Información
Information Retrieval (IR)

1.1. Programa / program

Máster Universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones (I^2-TIC)

Master in Research and Innovation in Information and Communications Technologies (I^2-ICT) [Officially certified]

1.2. Course code

32420

1.3. Course areas

Computer Science and Artificial Intelligence

1.4. Tipo de asignatura / Course type

Optativa [itinerario: Inteligencia computacional]
Elective [itinerary: Computational Intelligence]

1.5. Semester

First semester

1.6. Credits

6 ETCS

1.7. Language of instruction

The lecture notes are in English. The lectures are mostly in Spanish. Some of the lectures and seminars may be in English.

1.8. Recommendations / Related subjects

Knowledge of probability and statistics at an introductory level is useful to follow the course.

Related subjects are:
- Aprendizaje automático: teoría y aplicaciones [Machine Learning: theory and applications]
- Minería Web [Web Mining]
- Métodos bayesianos aplicados [Applied Bayesian Methods]

1.9. Lecturers

Add @uam.es to all email addresses below.

Lectures and labs:

Dr. Pablo Castells Azpilicueta (coordinator)
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-415
Tel.: +34 91 497 2016
E-mail: pablo.castells
Web: http://ir.ii.uam.es/castells

Dr. Fernando Díez Rubio
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-409
Tel.: +34 91 497 2213
E-mail: fernando.diez
Web: http://ir.ii.uam.es/~fdiez

1.10. Objetivos de la asignatura / Course objectives

La recuperación de información designa la ciencia y tecnología que subyace a los sistemas modernos de acceso a la información, tales como los motores de búsqueda Web y los sistemas de recomendación. La asignatura capacita al estudiante en el manejo, aplicación y evaluación de los métodos y tecnologías de recuperación de información no estructurada, incluyendo sus fundamentos teóricos, aplicados y metodológicos. Se aprenden técnicas orientadas al diseño, implementación y mejora de motores y algoritmos de búsqueda en texto (modelización, indexación, algoritmos de ranking) y sistemas de recomendación.

Information retrieval encompasses the science and technology underlying modern information access and delivery technologies, such as search engines and recom-

mender systems. In this course the student learns to handle, apply and evaluate unstructured information retrieval methods and technologies. The course provided an understanding of the theoretical foundations, as well as applied and methodological skills. The student shall learn techniques for the design, implementation and enhancement of text search algorithms and recommender systems.

At the end of each unit, the student should be able to:

UNIT BY UNIT SPECIFIC OBJECTIVES	
UNIT 1.- Fundamentals	
1.1.	Understand the nature of information retrieval, in contrast to data retrieval; understand fundamental notions such as information need, relevance and ranking, and the main variables involved in information retrieval problems; describe the main steps and components involved in an information retrieval process.
1.2.	Understand the mathematical formulation of the vector-space information retrieval model, as a basis for the development of effective information retrieval algorithms; run the algorithms and compute their output both manually and programmatically on examples and test data.
1.3.	Apply principled evaluation methodologies to assess the quality of information retrieval systems; understand formal effectiveness metrics and the theories underneath; compute metric values on example queries; design information retrieval experiments on both off-line public collections and live systems.
UNIT 2.- Advanced search	
2.1.	Formalize information retrieval tasks on a probabilistic basis; apply Bayesian methods and computational statistics techniques to develop effective information retrieval models; understand state of the art probabilistic approaches, along with the differences, commonalities, and equivalences between them.
2.2.	Define and apply dimensionality reduction techniques to text corpora, and incorporate them into information retrieval processes.
2.3.	Understand the distinctive specifics of the Web as an information retrieval domain; develop specific components for Web search, such as crawlers, and effective ranking algorithms that leverage the graph structure of the WWW to enhance the quality and effectiveness of search engines' output.
UNIT 3.- Personalized Information Retrieval	
3.1.	Devise personalization strategies that automatically adapt search results to the particularities and interest of individual users, by biasing the query, the graph-based scoring, or the final results ranking, using implicit or explicit evidence of user interests.
3.2.	Understand the automatic recommendation task and its different formulation variants; define and implement state of the art recommendation algorithms, and compute their output both manually and programmatically on examples and test data; define and compute specific evaluation metrics and evaluation paradigms to assess the effectiveness of recommender systems.
UNIT 4.- Advanced topics	
4.1.	Get up to date with latest advances in information retrieval, such as novelty and diversity in information retrieval, distributed retrieval, or learning to rank.

1.11. Course contents

1. Fundamentals

1.1. Introduction
- 1.1.1. Fundamental concepts
- 1.1.2. Components of a search engine

1.2. Basic information retrieval models
- 1.2.1. Boolean model
- 1.2.2. Vector-space model

1.3. Evaluation
- 1.3.1. The Cranfield paradigm
- 1.3.2. Information retrieval metrics
- 1.3.3. Test collections
- 1.3.4. Statistical significance

2. Advanced search

2.1. Probabilistic information retrieval models
- 2.1.1. The probability ranking principle
- 2.1.2. Binary independence model
- 2.1.3. Statistical language models

2.2. Latent factors
- 2.2.1. Algebraic Latent Semantic Analysis
- 2.2.2. Probabilistic latent semantic analysis
- 2.2.3. Latent Dirichlet allocation

2.3. Web search
- 2.3.1. The Web search scenario for search engines
- 2.3.2. Crawling and indexing
- 2.3.3. PageRank and HITS

3. Personalized information retrieval

3.1. Personalized search
- 3.1.1. Personalized PageRank
- 3.1.2. Personalized re-ranking and query refinement
- 3.1.3. Relevance feedback

3.2. Recommender Systems
- 3.2.1. The recommendation task
- 3.2.2. Neighbor-based collaborative filtering
- 3.2.3. Matrix factorization
- 3.2.4. Recommender system evaluation

4. Advanced topics

4.1. Novelty and diversity in information retrieval

4.2. Learning to rank

4.3. Time-aware recommendation

1.12. Course bibliography

Primary textbooks

1. Information Retrieval: Implementing and Evaluating Search Engines. S. Büttcher, C. L. A. Clarke, G. V. Cormack. MIT Press, 2010.
2. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Cambridge University Press, 2008.
3. Modern Information Retrieval – The Concepts and Technology behind Search. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Addison-Wesley, 2011.
4. Recommender Systems Handbook. F. Ricci, L. Rokach, B. Shapira, P. B. Kantor (Eds.). Springer Verlag, 2011.

Auxiliary readings

5. Search Engines – Information Retrieval in Practice. W. Bruce Croft, Donald Metzler, Trevor Strohman, Pearson, 2010.
6. Statistical Language Models. ChengXiang Zhai, Morgan & Claypool, 2009.
7. Google's PageRank and Beyond: The Science of Search Engine Rankings. Amy N. Langville and Carl D. Meyer. Princeton University Press, 2006.
8. Recuperación de Información: un enfoque práctico y multidisciplinar. F. Cacheda, J. M. Fernández Luna, J. Huete (editores). Ra-Ma, 2011.

1.13. Coursework and evaluation

The course involves lectures, exercises and lab assignments, and a written exam.

In both the ordinary and the extraordinary exam period it is necessary to have a pass grade (≥ 5) in the final exam, and average pass grades in both the exercises and lab assignments.

- In the ordinary exam period, the grade will be determined according to the following scheme:
  - 20 % Exercises
  - 30 % Lab assignments
  - 50 % Written exam
- In case of a fail grade in the ordinary exam period, in the extraordinary exam period, the student has the opportunity to:
  - Turn in all the exercises with corrections
  - Turn in all the lab assignments with corrections

  If the student does not turn in some of these items, the grades used will be the corresponding to the ordinary exam period.

  The grade will be determined by:
  - 20 % Exercises [only if the exercises are turned in]
  - 30 % Lab assignments [only if the lab assignments are turned in]
  - 50 % Written exam [mandatory]