Subject: Web Mining (WM)
Code: 32423
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: Web Mining (WM)

Academic year: 2015-2016

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: 2015/05/12
Status: Approved June 8th 2015

1. ASIGNATURA / COURSE (ID)

Minería Web
Web Mining (WM)

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

32423

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

Second semester

1.6. Credits

6 ETCS

1.7. Language of instruction

The lecture notes and the assignments and exam statements are in English. The lectures are mostly in Spanish. Some of the lectures and seminars may be in English. All the students' work can be presented in either Spanish or 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]
- Recuperación de Información [Information Retrieval]
- Métodos Bayesianos aplicados [Applied Bayesian Methods]

1.9. Lecturers

Add @uam.es to all email addresses below.

Dr. Alejandro Bellogín (coordinator)
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-434
Tel.: +34 91 497 2256
E-mail: alejandro.bellogin
Web: http://www.eps.uam.es/~abellogin

Dr. Iván Cantador
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-418
Tel.: +34 91 497 2215
E-mail: ivan.cantador
Web: http://www.eps.uam.es/~cantador

Dr. Irene Rodríguez
Departamento de Ingeniería Informática
Escuela Politécnica Superior
Office: B-424
Tel.: +34 91 497 2358
E-mail: irene.rodriguez
Web: http://www.eps.uam.es/~irodriguez

1.10. Objetivos de la asignatura / Course objectives

La asignatura capacita al estudiante en el conocimiento y manejo de técnicas y tecnologías del ámbito de la minería de datos, análisis y clasificación, y métodos estadísticos aplicados al contexto de la Web, las redes sociales y otros entornos emergentes de la Web social y semántica. En esta asignatura se estudian la extracción y procesado de datos en la Web, la minería de texto, el análisis de logs Web, la minería de opinión, la extracción y explotación de datos semánticos (semi-) estructurados, la minería de contenidos creados por usuarios, y las redes sociales: análisis y métricas para redes sociales, redes de mundo pequeño, descubrimiento de comunidades, acoplamiento bibliográfico, predicción de enlaces, búsqueda de expertos, análisis de influencia, y fenómenos de propagación.

In this course the student learns the knowledge and use of techniques and technologies in the scope of data mining, analysis, classification, and statistical methods, applied in the context of the Web, social networks, and emerging media in the Social Semantic Web. The course contents include Web data extraction and processing, text mining, Web log analysis, opinion mining, extraction and exploitation of semantic (semi-) structured data, user-created content mining, and social networks: social network analysis and metrics, small world networks, community discovery, bibliographic coupling, link prediction, expert search, influence analysis, and propagation phenomena.

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

UNIT BY UNIT SPECIFIC OBJECTIVES	
UNIT 1.- Introduction to Web Mining	
1.1	Characterize the origin, evolution and current status and structure of the World Wide Web
1.2	Know fundamental issues of Data Mining, Machine Learning, and Information Retrieval applied in the context of the Web
1.3	Characterize the main types of Web Mining, namely Web structure mining, Web content mining, and Web usage mining
UNIT 2.- Web data extraction and processing	
2.1	Know the fundamentals of Web crawling, and main design principles of Web crawlers and wrappers
2.2	Characterize the different types of Web crawlers, from all-purpose crawlers to topic specific crawlers
2.3	Know main tasks and techniques for processing text and Web documents
2.4	Know main mining techniques to analyze usage data from the Web
2.5	Know different techniques for identifying and extracting usage patterns in log data
UNIT 3.- Mining social network data	
3.1	Know main principles of social networks, and characterize the application of Web mining techniques to social networks
3.2	Know techniques, models, and metrics for analyzing social networks

3.3	Characterize the existing algorithms for graph node and link ranking, and know how to apply these algorithms in real world problems
3.4	Characterize the different types of information diffusion and know different techniques for analyzing information propagation in social networks.
3.5	Know real applications that make use of social network analysis
3.6	Use some existing resources and tools for social network analysis.
UNIT 4.- Mining user generated contents	
4.1	Know the existing main types of user-generated contents in the Web
4.2	Characterize the origin, evolution, and current status and structure of the so called Social Web (or Web 2.0)
4.3	Know existing techniques, resources, and tools for searching, extracting and processing information about opinions and sentiments in Web contents
4.4	Understand challenges and limitations of current opinion mining and sentiment analysis techniques
4.5	Understand the application of crowdsourcing as a mechanism of collective intelligence, and know specific applications, such as system evaluation
UNIT 5.- Mining web semantic data	
5.1	Know the existing main types of semi-structured data in the Web
5.2	Characterize the origin, evolution, and current status and structure of the so called Semantic Web (or Web of Data)
5.3	Explain the benefits of using semantic-based approaches in information access and retrieval applications in the Web
5.4	Know existing techniques, resources, and tools for automatically extracting structured data from the Web
5.5	Use some existing resources and tools to exploit structured knowledge sources in the (Semantic) Web

1.11. Course contents

1. **Introduction**
   1.1. The World Wide Web
   1.2. Web Data Mining
   1.3. Data Mining foundations
2. **Web data extraction and processing**
   2.1. Web crawling
   2.2. Web scraping
   2.3. Web text processing and document representation
   2.4. Web log extraction and mining
3. **Mining social network data**
   3.1. Social Network Analysis
   3.2. Information propagation in social networks
   3.3. Community discovery in social networks
   3.4. Expert finding in social networks
   3.5. Link prediction in social networks

4. **Mining user generated contents**
   - 4.1. The Social Web
   - 4.2. Opinion mining
   - 4.3. Mining microblogging data
   - 4.4. Mining social tagging data
   - 4.5. Crowdsourcing
5. **Mining Web semantic data**
   - 5.1. The Semantic Web
   - 5.2. Semi-structured data in the Web
   - 5.3. Structured data extraction
   - 5.4. Ontologies and Linked Data
   - 5.5. Semantic Web applications

1.12. Course bibliography

Bibliography available at the library's catalogue (click here)

1. Liu, B. **Web Data Mining: Exploring Hyperlinks, Contents and Usage Data**. Springer-Verlag (2009)
2. Chakrabarti, S. **Mining the Web: Discovering Knowledge from Hypertext Data**. Morgan Kaufmann (2002)
3. Chang, G., Healey, M. J., McHugh, J. A. M., Wang, J. T. L. **Mining the World Wide Web: An Information Search Approach**. Kluwer Academic Publishers (2001)
4. Feldman, R., Sanger, J. **The Text Mining Handbook**. Cambridge University Press (2006)
5. Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.). **The Adaptive Web: Methods and Strategies of Web Personalization**. Springer-Verlag (2007)
6. Watts, D. J. **Small Worlds: The Dynamics of Networks between Order and Randomness**. Princeton University Press (1999)
7. Kosala, R., Blockeel, H. **Web Mining Research: A Survey**. ACM SIGKDD Explorations Newsletter 2(1), pp. 1-15 (2000)
8. Arasu, A., Garcia-Molina, H. **Extracting Structured Data from Web Pages**. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 337-348 (2003)
9. Kaplan, A. M., Haenlein, M. **Users of the World, Unite! The Challenges and Opportunities of Social Media**. Business Horizons 53, pp. 59-68 (2010)
10. Srivastava, J., Cooley, R., Deshpande, M., Tan, P. **Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data**. SIGKDD Explorations 1(2), pp. 12–23 (2000)

11. Wasserman, S., Faust, K. **Social network analysis: Methods and applications.** Cambridge university press (1994).
12. Zafarani, R., Abbasi, M. A., Liu, H.. **Social media mining: an introduction.** Cambridge University Press (2014).
13. Liben-Nowell, D., Kleinberg, J.. **The link-prediction problem for social networks**. Journal of the American society for information science and technology 58.7: 1019-1031 (2007).
14. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.. **Graph structure in the Web**. Computer Networks 33(1-6), pp. 309-320 (2000)
15. Silvestri, F. **Mining Query Logs: Turning Search Usage Data into Knowledge**. Foundations and Trends in Information Retrieval 4(1–2), pp. 1–174 (2010)

1.13. Coursework and evaluation

The course involves lectures, theory and lab assignments, a small research project, and a written exam.

The project will take some of the lab sessions, and should address research topics related to those of the course. It may be proposed by the students under acceptance of the lecturers, or may be proposed by the lecturers. It will consist of two main stages:

- Developing a software implementation, attempting to reproduce and/or improve state of the art approaches, or to propose and evaluate novel approaches.
- Making an oral presentation in the classroom at the end of the course, explaining the work done and the results obtained in the project.

In both the ordinary and the extraordinary exam period it is necessary to have a pass grade (≥ 5) in the final exam, a pass grade in the research project, 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% Theory assignments
  - 30 % Lab assignments
  - 20 % Project
  - 30 % 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 theory assignments with corrections
- Turn in all the lab assignments with corrections
- Turn in all the project with corrections
- Do an extraordinary exam, in case the ordinary was failed

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

In the extraordinary exam period, the grade will be determined by:

- 20 % Theory assignments [with or without corrections]
- 30 % Lab assignments [with or without corrections]
- 20 % Project [with or without corrections]
- 30 % Written exam [ordinary or extraordinary]