

# GUÍA DOCENTE

Minería Web

## I.- DATOS INICIALES DE IDENTIFICACIÓN

<b>Nombre de la asignatura:</b>	<i>Minería Web</i>
<b>Carácter:</b>	<i>Optativo</i>
<b>Titulación:</b>	<i>Master en Sistemas y servicios en la sociedad de la información.</i>
<b>Ciclo:</b>	<i>Postgrado</i>
<b>Departamento:</b>	<i>Informática</i>
<b>Profesores responsables:</b>	<i>Miguel Lozano</i>

## II.- INTRODUCCIÓN A LA ASIGNATURA

Esta asignatura cubre un amplio espectro de modelos computacionales orientados a la extracción de información o conocimiento potencialmente útil y previamente desconocido a partir de la World Wide Web, es decir, se centra en el contexto conocido como Minería Web. La asignatura contempla los tres aspectos fundamentales del proceso de minería web:

- Minería de la Estructura: Estudio del grafo que representa la WWW y de los sistemas capaces de inferir dicha estructura, es decir, los agentes sw conocidos como crawler/spiders. Además se revisan las técnicas empleadas para extraer información útil de dicho grafo (ej. La importancia o ranking de una página web: PageRank, HITS), así como las arquitecturas principales de los buscadores que emplean dichas técnicas.
- Minería del contenido: Revisión de las técnicas de aprendizaje más comunes empleadas en minería de texto/hipertexto (ej: Modelos Bayesianos, Árboles de inducción, técnicas de clustering, etc).
- Minería del uso: Revisión de las técnicas que puedan predecir el comportamiento del usuario cuando interacciona con la web. Extracción de patrones de navegación (HGPS), preprocesado y análisis de ficheros de uso web (ej: logs de servidores web).

## III.- VOLUMEN DE TRABAJO

La asignatura tiene asignados 3 ECTS. Considerando que cada ECTS debe corresponderse con un volumen de trabajo de entre 25 y 30 horas, supone un volumen total de entre 75 y 90 horas. Para el cálculo del volumen de trabajo se ha tomado como referencia un total de 18 horas presenciales que incluyen tanto las clases de teoría como las de prácticas. La distribución prevista del trabajo es la siguiente:

<b>ACTIVIDAD</b>	<b>Horas/curso</b>
ASISTENCIA A CLASES TEÓRICAS:	13,5
ASISTENCIA A CLASES PRÁCTICAS:	4,5
PREPARACIÓN DE TRABAJOS:	36
ESTUDIO PREPARACIÓN CLASES:	18
PREPARACIÓN PROBLEMAS:	6
ASISTENCIA A TUTORÍAS:	10
<b>TOTAL VOLUMEN DE TRABAJO:</b>	<b>88</b>

#### **IV.- OBJETIVOS GENERALES**

- **Objetivos de carácter general:**
  - Conocer los problemas derivados de la extracción de conocimiento en general y profundizar en aquellos aplicados a la web.
  - Conocer la estructura y funcionamiento de un buscador, como se realizan las búsquedas y se clasifican los documentos (crawler, indexer).
  - Conocer las principales técnicas estadísticas y lingüísticas aplicadas a la minería del contenido de la web (texto, hipertexto, XML).
  - Conocer el grafo de enlaces que representa la web, así como los principales modelos estructurales de la misma: Redes sociales y análisis de citas.
  - Conocer las principales técnicas empleadas para predecir el comportamiento del usuario cuando interactúa con la web. Técnicas para la búsqueda de patrones de navegación.
- **Objetivos metodológicos:**
  - Familiarización con el uso de herramientas informáticas.
  - Aplicación de métodos científicos en la resolución de los trabajos experimentales.
  - Familiarización con las fuentes de información tradicionales (artículos científicos) y las nuevas tecnologías.

## **V.- CONTENIDOS**

Parte I:

- Introducción a la minería web

Parte II: Minería de la Estructura de la web

- Motores de búsqueda
- Agentes de búsqueda e indexación
- Clasificación de documentos

Parte III: Minería del Contenido de la web

- Técnicas de recuperación de información semi-estructurada

Parte IV: Análisis de uso de la web

- Análisis de *'logs'* y técnicas de extracción de patrones de navegación.

## **VI.- DESTREZAS A ADQUIRIR**

Solidez en los conocimientos básicos relacionados con la minería de datos y la extracción de información relevante.

Familiarización con las herramientas computacionales empleadas en el área de minería web.

Capacidad de análisis y síntesis.

Manejo de bibliografía científica.

Enfrentamiento y resolución de problemas computacionales reales y complejos.

## **VII.- HABILIDADES SOCIALES**

Manejo de las nuevas fuentes de información (web).

Habilidad en la búsqueda, selección y valoración de información.

Resolución de problemas complejos y análisis de resultados.

Trabajos individuales y en grupos (coordinación).

Capacidad para realizar una exposición clara y elegante de los trabajos científicos desarrollados.

## VIII.- TEMARIO Y PLANIFICACIÓN TEMPORAL

1. **Introducción** (1'5 horas)
  - a. Problemas de la minería de datos y contexto (la web)
  - b. Características de la información no estructurada
2. **Minería de la estructura** (6 horas)
  - a. La web como un grafo de enlaces
  - b. Motores de búsqueda (arquitecturas centr. vs distr.)
    - i. Agentes de búsqueda (crawlers)
    - ii. Indexación y Ranking de documentos (Boolean/Vector models)
    - iii. Metabuscadores
  - c. Clase práctica (Boolean and Vector models)
3. **Minería del Contenido de la web** (6 horas)
  - a. Textmining: Introducción y problemas
    - i. Modelos de representación: "Bolsas de palabras", N-gramas.
    - ii. Clasificación de textos
  - b. Clase práctica (Clasificación de textos)
4. **Minería del uso de la web** (3 horas)
  - a. Análisis de logs
  - b. Extracción de patrones de navegación
    - i. Hypertext Probabilistic Grammar (HPG)
  - c. Clase práctica (Análisis de logs de la UV)
5. **Futuro** (1'5 horas)
  - a. Personalización y NLP.

## IX.- BIBLIOGRAFÍA DE REFERENCIA

### Bibliografía básica:

[Baeza&Neto99] Modern Information Retrieval. Baeza-Yates, Ribeiro-Neto. Addison Wesley 1999.

[Orallo99] Introducción a la Minería de Datos. Hdez. Orallo,

Ramirez Quintana, Ferri Ramirez. Prentice-Hall 2004.

[Chakarbarti03] Mining the Web, Discovering knowledge from hypertext data. S. Chakarbarti. Morgan-Kaufmann 2003

### Web:

Información sobre XML: <http://www.w3.org>

Información sobre web mining: <http://www.kdnuggets.com/>

## **Bibliografía complementaria:**

[SIGKDD00] Especial sobre Web Mining: Junio 2000, Vol2, nº 1 de la revista ACM SIGKDD Explorations, Newsletters of the ACM Special Interest Group on Knowledge Discovery and Data Mining.

[Kosala00] Web Mining Research: A Survey (Incluido en [SIGKDD00])

Web:

The center for web research: <http://www.cwr.cl/events/index.html>

## **X.- METODOLOGÍA**

El desarrollo de la asignatura se estructura en torno a cuatro ejes: las sesiones de teoría, las de problemas prácticos (resueltos en el aula con el ordenador), las tutorías y la preparación y posterior defensa de un trabajo individual por cada alumno. Por lo que respecta a las primeras, el alumno asistirá a una clase teórica por semana, donde el profesor desarrollará los puntos principales del temario comentado. El alumno debe atender al tiempo de preparación de las clases previsto para su aprovechamiento óptimo. Las clases prácticas (3 en total) servirán para que el alumno verifique el grado de conocimiento adquirido, ya que se enfrentará a problemas complejos (ej: Clasificación, ranking y clustering de páginas web), así como al análisis de los resultados obtenidos. Todas las sesiones prácticas llevarán asociado un formulario para que el alumno realice el reporte de los datos obtenidos y posteriormente envíe al profesor. Al igual que antes, el alumno deberá preparar dichas sesiones para poder realizar los experimentos en el tiempo previsto. Todo intercambio de información entre profesor y alumno se realizará a través del aula virtual de la Universitat de València ([aulavirtual.uv.es](http://aulavirtual.uv.es)).

Los trabajos finales pueden ser de carácter teórico (resumen o profundización de alguna parte del estado del arte) o práctico (colección de problemas a resolver con alguna de las herramientas utilizadas en el curso).

## **XI.- EVALUACIÓN DEL APRENDIZAJE**

Cada alumno deberá elaborar individualmente uno de los trabajos propuestos (70% de la nota final) y presentarlo en clase (30% de la nota final). La calificación final del alumno se obtiene sumando el 70% de la nota del trabajo más el 30% de la nota de la presentación. Los alumnos que no realicen un trabajo serán evaluados mediante un examen, que constará de una serie de 10 cuestiones cortas a contestar en 1 hora de tiempo sin apuntes ni material de consulta. En ambos casos los requerimientos mínimos para superar la asignatura pasan por obtener un 5 en la prueba realizada. Voluntariamente, el alumno puede aumentar el dominio o la complejidad de los ejercicios prácticos realizados en clase, lo cual puede aumentar hasta un 10% (máximo) la nota final obtenida.