ASIGNATURA DE MÁSTER:

# MINERÍA DE LA WEB

Curso 2013/2014

(Código:31101023)

## 1.PRESENTACIÓN

### Objetivos generales de la materia

El curso se dirige a conocer las tecnologías existentes para extraer información de la web, tanto a partir de sus contenidos (recuperación de información, extracción de información, creación de recursos lingüísticos, etc.) como de su estructura y su uso. Al finalizar el curso, el alumno deberá ser capaz de plantear la arquitectura completa de un sistema automático de acceso y extracción de información en la web.

### Ficha técnica

*Tipo:* Optativa

*Cuatrimestre:* Primero

*Créditos y horas de trabajo*

*Créditos Totales y Horas:* 6 / 150

*Horas de estudio teórico:* 105

*Horas de trabajo práctico:* 45

*Horas de actividades complementarias:* 0

## 2.CONTEXTUALIZACIÓN

Esta asignatura se encuadra en el módulo “ESP-LSI-1 Tecnologías del Lenguaje en la Web” dentro de la especialidad con el mismo nombre de la titulación de posgrado “Master en Lenguajes y Sistemas Informáticos”.

## 3.REQUISITOS PREVIOS RECOMENDABLES

Diseño e implementación de Sistemas Informáticos.

Lectura fluida del inglés.

## 4.RESULTADOS DE APRENDIZAJE

1. Tener una visión amplia de las áreas relacionadas con la extracción de información en la web.
2. Hábito de lectura de artículos científicos.
3. Capacidad para buscar información que complete el material propuesto inicialmente. Esta búsqueda es un entrenamiento necesario en la formación del alumno como investigador. Con cada trabajo tendrá mayor capacidad para encontrar y discriminar fuentes de información relevantes, requisito para desarrollar cualquier trabajo de investigación posterior.

4. Capacidad de reflexión sobre el material estudiado, necesaria para poder realizar una síntesis de calidad.
5. Capacidad para escribir textos con un formato de artículo científico, tanto en lo referente a la estructuración de contenidos, como de formato del propio artículo.
6. Compartir el conocimiento adquirido, aprovechando el trabajo y el esfuerzo realizado por cada alumno.
7. Autoevaluar los conocimientos adquiridos por comparación del trabajo propio con el trabajo de sus compañeros, tanto en lo relativo a contenidos, como a estructura y redacción de los trabajos.

Objetivos por tema y orientaciones breves

Tema 1. Introducción

*Objetivos:*

O.1.1 Determinar los problemas que surgen al interactuar con la web.

O.1.2 Definir Minería de la web

O.1.3 Definir Crawling

O.1.4 Definir Búsqueda en web

O.1.5 Definir Minería de contenido de la web (minería de texto)

O.1.6 Definir Minería de uso de la web

O.1.7 Definir Minería de estructura de la web

O.1.8 Definir Dinámica de la web.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 2. Crawling, filtrado e indexación

## Objetivos:

O.2.1 Comprender la funcionalidad de un crawler.

O.2.2 Acotar los problemas que intenta resolver un crawler.

O.2.3 Determinar los problemas con los que se encuentra un crawler (técnicos, legales, etc.)

O.2.4 Establecer las etapas del crawling.

O.2.5 Identificar otras áreas de investigación relacionadas con el Crawling.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 3. Consulta y búsqueda en web

*Objetivos:*

O.3.1 Determinar las características propias de la web que afectan a la búsqueda.

O.3.2 Caracterizar los tipos de información a considerar en la búsqueda en web (Contenido textual, Información en los enlaces, Estructura de enlace entre páginas, etc.).

O.3.3 Comprender el proceso de indexación de la información en web.

O.3.4 Estudiar interfaces de exploración y visualización de la búsqueda.

O.3.5 Definir Metabúsqueda y Agentes web.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 4. Minería de textos

*Objetivos:*

O.4.1 Definir corpus.

O.4.2 Comprender cómo se puede crear y usar un corpus a partir de la web.

O.4.3 Definir Extracción de Información textual.

O.4.4 Conocer la arquitectura de un sistema de Extracción de Información.

O.4.5 Definir Extracción de terminología.

O.4.6 Conocer alguna metodología de extracción de terminología a partir de la web.

O.4.7 Identificar la problemática asociada al lenguaje natural.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 5. Minería de uso de la web

*Objetivos*:

O.5.1 Definir y establecer los objetivos de minería de uso de la web.

O.5.2 Determinar las etapas de procesamiento (Preprocesamiento, Inferencia de patrones, Análisis de patrones).

O.5.3 Conocer algunas herramientas existentes.

O.5.4 Identificar técnicas de aprendizaje aplicadas a minería de uso.

O.5.5 Saber qué son los sitios web adaptativos.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 6. Minería de estructura de la web

*Objetivos*:

O.6.1 Definir y establecer los objetivos de la minería de estructura de la web.

O.6.2 Definir y modelar las nociones de Autoridad (authoritative page), prestigio, Centralidad y Co-cita.

O.6.3 Conocer cómo se realiza el ranking de páginas web basado en enlaces: PageRank y HITS.

O.6.4 Estudiar cómo se realiza el análisis de comunidades en la web.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

Tema 7. Dinámica de la web

*Objetivos*:

O.7.1 Definir y establecer los objetivos del estudio de la dinámica de la web.

O.7.2 Determinar las características de la web susceptibles de estudio.

O.7.3 Estudiar la Ley de Zipf, "power laws" en la web así como sus aplicaciones.

O.7.4 Comprender cómo se determina el tamaño y tendencia de crecimiento de la web.

O.7.5 Comparar las web pública y web oculta.

O.7.6 Comprender cómo se determina la presencia de un idioma en la web.

O.7.7 Conocer estudios sobre la web española.

*Orientaciones:*

El alumno deberá leer una serie de artículos y completar un resumen.

## 5.CONTENIDOS DE LA ASIGNATURA

### Tema 1. Introducción

Problemas que surgen al interactuar con la web. Breve definición de Minería de la web y de Crawling, Búsqueda en web, Minería de contenido de la web (minería de texto), Minería de uso de la web, Minería de estructura de la web, Dinámica de la web.

### Tema 2. Crawling, filtrado e indexación

Qué es un crawler.  Problemas que intenta resolver un crawler. Problemas con los que se encuentra un crawler (técnicos, legales, etc.) Etapas del crawling. Otras áreas de investigación relacionadas con el Crawling.

### Tema 3. Consulta y búsqueda en web

Características propias de la web que afectan a la búsqueda. Tipos de información a considerar en la búsqueda en web (Contenido textual, Información en los enlaces, Estructura de enlace entre páginas, etc.). Proceso de indexación de la información en web.  Interfaces, browsing y visualización de la búsqueda. Metabúsqueda. Agentes web.

### Tema 4. Minería de textos

Qué es un corpus. Creación de corpus. Posibles usos y utilidad de un corpus. Creación de corpus a partir de la web. Ejemplos de algunos corpus y su finalidad. Extracción de Información textual (Automatic Information Extraction). Arquitectura de un sistema de EI. Extracción de terminología (Automatic Terminology Extraction). Extracción de terminología a partir de la web. Problemática asociada al lenguaje natural. Similitud, clasificación, clustering.

### Tema 5. Minería de uso de la web

Definición y objetivos de minería de uso de la web. Etapas de procesamiento (Preprocesamiento, Inferencia de patrones, Análisis de patrones). Herramientas existentes. Técnicas de aprendizaje aplicadas a minería de uso. Sitios web adaptativos.

### Tema 6. Minería de estructura de la web

Definición y objetivos de la minería de estructura de la web. Definición, modelado y uso de las nociones de Autoridad (authoritative page), prestigio, Centralidad y Co-cita. Ranking de páginas web basado en enlaces: PageRank y HITS. Análisis de comunidades en la web. Otras aplicaciones de la minería de estructura.

### Tema 7. Dinámica de la web

Definición y objetivos del estudio de la dinámica de la  web. Características de la web susceptibles de estudio. Ley de Zipf, "power laws" en la web. Tamaño y tendencia de crecimiento de la web. Web pública y web oculta. Idiomas en la web. Dominios en la web. Estudios sobre la web española.

## 6.EQUIPO DOCENTE

- ANSELMO PEÑAS PADILLA
- LAURA PLAZA MORALES

## 7.METODOLOGÍA

El curso de doctorado consta de siete temas cuyo estudio se realiza con la siguiente metodología dentro de un paradigma de construcción de conocimiento:

Para cada tema, el alumno debe acceder al material propuesto por el equipo docente. Este material consta de:

- Bibliografía básica común a todos los temas. Se trata de libros con un conocimiento ya estructurado facilitando la introducción del alumno en la materia.
- Artículos científicos. Se propone la lectura de algunos artículos de carácter científico. Su contenido es más específico y de más difícil lectura. A partir de ellos, el alumno conocerá la estructura y formato que deben seguir los textos de estas características y que el tendrá que escribir más adelante.
- Enlaces web: enlaces que apuntan a sitios web donde encontrar nuevas referencias bibliográficas, enlaces a sitios web con recursos y herramientas relacionados con el tema, enlaces a otros cursos o tutoriales, etc.

A partir de este material y con la guía de un cuestionario, el alumno debe realizar un resumen sintetizando el conocimiento que ha adquirido. La elaboración del resumen se dirige a:

- Estimular la lectura detenida del material propuesto.
- Provocar la necesidad de buscar información que complete el material propuesto inicialmente. Esta búsqueda es un entrenamiento necesario en la formación del alumno como investigador. Con cada trabajo tendrá mayor capacidad para encontrar y discriminar fuentes de información relevantes, requisito para desarrollar cualquier trabajo de investigación posterior.
- Estimular una reflexión sobre el material estudiado, necesaria para poder realizar una síntesis de calidad.
- Aprender a escribir textos con un formato de artículo científico, tanto en lo referente a la estructuración de contenidos, como de formato del propio artículo. En especial, contextualizar la síntesis referenciando correctamente las fuentes utilizadas.

Tras la elaboración del resumen, el alumno debe realizar una entrega electrónica de su resumen y de los nuevos enlaces y referencias más importantes que ha encontrado a lo largo de su trabajo. Esto servirá de material de evaluación para el equipo docente, que podrá valorar no sólo los conocimientos adquiridos, sino también la evolución y el progreso del alumno en la adquisición de la metodología y actitud necesaria para un investigador.

Los últimos meses del curso se dirigen a afianzar los conocimientos adquiridos mediante la elaboración de un trabajo final de carácter personal. El trabajo puede ser propuesto por el propio alumno y preferiblemente deberá tener un carácter de aplicación de los conocimientos adquiridos.

Plan de trabajo

Tema 1. Introducción. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Primera quincena de diciembre.

Tema 2. Crawling, filtrado e indexación. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Segunda quincena de diciembre.

Tema 3. Consulta y búsqueda en web. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Primera quincena de enero.

Tema 4. Minería de textos. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Segunda quincena de enero.

Tema 5. Minería de uso de la web. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Primera quincena de febrero.

Tema 6. Minería de estructura de la web. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Segunda quincena de febrero.

Tema 7. Dinámica de la web. 15 horas. Plazo para realizar las lecturas y entregar el resúmen: Primera quincena de marzo.

Trabajo Práctico. 45 horas. Plazo para realizar el trabajo y entregar la memoria: Segunda quincena de marzo y mes de abril.

Actividades prácticas programadas:

Definición, implementación y redacción de una memoria correspondiente a un trabajo práctico de carácter individual que se definirá con los porfesores de la asignatura de acuerdo a los intereses del alumno y sobre la base de los conocimientos adquiridos en los temas teóricos.

Otras actividades programadas:

Se irán anunciando de forma dinámica en el entorno virtual.

## 8.BIBLIOGRAFÍA BÁSICA

LIBRO ACTUALMENTE NO PUBLICADO
ISBN(13):
Título: MINING THE WEB: DISCOVERING KNOWLEDGE FROM HYPERTEXT DATA (2002)
Autor/es: Soumen Chakrabarti ;
Editorial: MORGAN KAUFMANN

LIBRO ACTUALMENTE NO PUBLICADO
ISBN(13):
Título: WEB DATA MINING: EXPLORING HYPERLINKS, CONTENTS, AND USAGE DATA (2007)
Autor/es: Bing Liu ;
Editorial: Springer

## 9.BIBLIOGRAFÍA COMPLEMENTARIA

## 10.RECURSOS DE APOYO AL ESTUDIO

Material de estudio

Artículos (generalmente en inglés) disponibles en el sitio web de la asignatura:

Tema 1.

· Kosala, R. and Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining. 2000.

· Chakrabarti, S. Data Mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining, 2000.

· Ricardo baeza-Yates. Excavando la web. El profesional de la información. v13, n1, 2004

· Bibliografía básica

Tema 2.

· Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, vol. 30, 1998.

· Junghoo Cho, Hector Garcia-Molina, Lawrence Page. Efficient Crawling Through URL Ordering. Computer Networks and ISDN Systems, vol. 30, 1998.

· Mercator:

· Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler. In Proceedings of World Wide Web, 1999, pages 219-229.

· Marc Najork, Allan Heydon. High-Performance Web Crawling. SRC Research Report, Compaq Systems Research Center, 2001 (versión ampliada del anterior)

· Brian Pinkerton. Finding what people want: Experiences with the WebCrawler. In Proc. 1st International World Wide Web Conference, 1994.

Tema 3.

· Steve Lawrence and C. Lee Giles. Searching the World Wide Web. Science vol. 280, 1998.

· Nick Craswell, David Hawking and Stephen Robertson. Effective Site Finding using Link Anchor Information. Research and Development in Information Retrieval, SIGIR 2001.

· Dunja Mladenic. Text-Learning and Related Intelligent Agents: A survey. IEEE Intelligent Systems, 1999

· Bibliografía básica.

· Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information retrieval. ACM Press. Addison Wesley, 1999

Tema 4.

· Marti A. Hearst. Untangling Text Data Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper).

· Turmo, Jordi. Information Extraction, Multilinguality and Portability. Revista Iberoamericana de Inteligencia Artificial, N.22, vol. 5, Invierno 2003.

· Peñas, A., Verdejo, F. and Gonzalo, J. Terminology Retrieval: towards a synergy between thesaurus and free-text searching. In F.J. Garijo, J.C. Riquelme and M. Toro editors, Advances in Artificial Intelligence - IBERAMIA 2002, LNAI 2527,

Lecture Notes in Computer Science. Springer-Verlag, 2002.

Tema 5.

· R. Cooley, B. Mobasher, and J. Srivastava. Web mining and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, 1997

· J. Srivastava, R. Cooley, M. Deshpande, P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 2000.

· B. Mobasher. Web Usage Mining and Personalization. Chapter in Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2004

· Mike Perkowitz and Oren Etzioni. Adaptive Web Sites: an AI Challenge, IJCAI, 1997

· Thorsten Joachims, Dayne Freitag and Tom M. Mitchell.Web Watcher: A Tour Guide for the World Wide Web, IJCAI, 1997

Tema 6.

· Soumen Chakrabarti et al. Mining the Link Structure of the World Wide Web. Computer, volume 32, n.8, pp. 60-67, 1999.

· Ravi Kumar et al. The Web as a Graph. Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS, ACM Press, 2000.

· Broder et al. Graph Structure in the web. Proc.WWW9, 2000

Tema 7.

· M. Levene and A. Poulovassilis. Web Dynamics. Software Focus, 2, (2001), 60-67.

· Ricardo Baeza-Yates, Bárbara J. Poblete y Felipe Saint-Jean. Evolución de la Web Chilena. Centro de Investigación de la Web, 2003.

· Edward T. O'Neill, Brian F. Lavoie, Rick Bennett. Trends in the Evolution of the Public Web (1998 - 2002). D-Lib Magazine,Volume 9 Number 4, April 2003.

· Broder et al. Graph Structure in the web. Proc.WWW9, 2000.

## 11.TUTORIZACIÓN Y SEGUIMIENTO

La tutorización de los alumnos se llevará a cabo a través de la plataforma de e-Learning Alf y del correo electrónico anselmo@lsi.uned.es

## Información de contacto

Anselmo Peñas Padilla (Coordinador). Dpto. Lenguajes y Sistemas Informáticos (U.N.E.D)

e-mail: anselmo@lsi.uned.es

web personal: http://nlp.uned.es/~anselmo

Fernando López Ostenero. Dpto. Lenguajes y Sistemas Informáticos (U.N.E.D)

e-mail: flopez@lsi.uned.es

web personal: http://nlp.uned.es/~ostenero

## 12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación se realizará a partir de los resúmenes entregados en cada tema teórico y del proyecto de final de asignatura.

## 13.COLABORADORES DOCENTES

Véase equipo docente.