

SISTEMAS DE INFORMACIÓN NO ESTRUCTURADA

Curso 2017/2018

(Código: 31106116)

1. PRESENTACIÓN

En la actualidad, la cantidad de información online es prácticamente ilimitada, lo que ha supuesto una de las transformaciones sociales más importantes en este siglo. Las aplicaciones prácticas son muy amplias, incluyendo estudios de mercado, buscadores especializados, sistemas de recomendación, análisis de redes sociales, etc.

El principal reto informático es el acceso a información no estructurada, es decir, a información textual o multimedial, debido a la variabilidad del lenguaje y a la dificultad de interpretar y organizar contenidos. Es decir, cómo enlazar, agrupar o clasificar toda esta información según las necesidades de un sistema. Dentro del acceso a información no estructurada se incluyen las tecnologías de búsqueda, los clasificadores automáticos, las técnicas de agrupación (clustering), la extracción de información y la minería de datos textuales, etc.

Un sistema de acceso a la información cubre aspectos tan diversos como la eficiencia de acceso y escalabilidad de bases de datos, la compresión, el diseño de interfaces de usuario, reconocimiento de patrones en imágenes, evaluación de componentes o la interacción con el usuario. No es posible cubrir todos estos aspectos en un curso. Sin embargo, a través de esta asignatura el alumno adquirirá las capacidades básicas para el desarrollo de sistemas de acceso a información, centrándose en los desafíos específicos dados por el uso de información no estructurada.

2. CONTEXTUALIZACIÓN

"Esta asignatura de carácter optativo pertenece al Máster Universitario en "Inteligencia Artificial Avanzada: Fundamentos, Métodos y Aplicaciones" que se imparte por profesores del Departamento de "Lenguajes y Sistemas Informáticos" de la ETSI Informática de la UNED. Esta asignatura es de carácter anual con una carga de 6 ECTS. El objetivo fundamental es dar una visión integradora de técnicas relacionadas con los sistemas de acceso a la información y en concreto en el caso de información no estructurada.

Las competencias de esta asignatura se pueden consultar en la guía del máster.

3. REQUISITOS PREVIOS RECOMENDABLES

Los propios del máster. Además son necesarios conocimientos de inglés técnico (lectura y escritura) para manejar las fuentes bibliográficas.

4. RESULTADOS DE APRENDIZAJE

Los resultados del aprendizaje esperados en este curso son:

- Entender los principales retos a los que se enfrenta un sistema de acceso a información no estructurada, así como las posibilidades que ofrece el desarrollo de este tipo de sistemas en un contexto industrial.

- Ser capaz de modularizar el desarrollo de un sistema de acceso a la información y abordar cada uno de los aspectos implicados de forma independiente.
- Conocer las técnicas básicas más empleadas en las distintas fases que componen un sistema de acceso a la información, desde el procesamiento de texto, tratamiento de la estructura de la Web y de las redes sociales, hasta los mecanismos de clasificación, ordenación (ranking) o agrupación (clustering) de contenidos.
- Capacidad de implementar algunos mecanismos básicos de tratamiento de la información no estructurada.
- Capacidad de recopilar, analizar y sintetizar conocimiento de aplicación industrial a partir de publicaciones académicas.

5. CONTENIDOS DE LA ASIGNATURA

El temario de la asignatura es el siguiente:

1. Introducción: Fundamentos y arquitectura de un sistema de acceso a la información no estructurada. Adquisición, transformación e indexación.
2. Adquisición y procesamiento de contenidos en internet: *Crawling* y tratamiento de texto.
3. Técnicas de organización de la información: ordenación (ranking), agrupación (clustering) y clasificación de documentos.
4. Redes sociales y otros objetos de búsqueda.

Muchos de los sistemas de acceso a información no estructurada se aplican sobre conjuntos de datos locales o centrados en un dominio. Por ello, con el fin de evitar exceso de contenidos, no se abordarán técnicas específicas de indexación y compresión para búsquedas sobre cantidades masivas de datos (como es el caso de los buscadores Web). Por el mismo motivo no se discuten las metodologías de evaluación de sistemas de acceso a la información. La evaluación es particular de cada problema concreto independientemente del tipo de información que se accede, mientras que este curso se centra en los aspectos característicos de la información no estructurada. Por ello, más que en la evaluación global de un sistema de información, el curso cubrirá evaluación de componentes específicos del tratamiento de información no estructurada, como son el ranking, agrupación o clasificación de documentos.

6. EQUIPO DOCENTE

- [ENRIQUE AMIGO CABRERA](#)
- [JULIO ANTONIO GONZALO ARROYO](#)
- [JORGE AMANDO CARRILLO DE ALBORNOZ CUADRADO](#)

7. METODOLOGÍA

Las diferentes asignaturas que integran este Master se imparten conforme a la metodología no presencial que caracteriza a la UNED, en la cual prima el autoaprendizaje del alumno, pero asistido por el profesor y articulado a través de diversos sistemas de comunicación docente-discente. El Máster en Ingeniería Informática se imparte con apoyo en una plataforma virtual interactiva de la UNED, donde el alumno encuentra tanto materiales didácticos básicos como complementarios, informaciones, noticias, ejercicios y también permite la evaluación correspondiente a las diferentes materias.

Las diferentes actividades formativas se organizan como sigue:

- Estudio de contenidos: 50 horas.
- Tutorías: 10h

- Actividades en la plataforma virtual: 10 horas.
- Trabajos individuales: 40 horas.
- Trabajos en equipo: 10 horas.
- Práctica 30 horas.

En cada caso se refieren a:

1. Estudios de contenidos teóricos: Lectura de las orientaciones, material multimedia y estudio del libro base.
2. Tutorías: Interacción presencial o virtual con los tutores y aclaración de cuestiones planteadas por los estudiantes.
3. Actividades en la plataforma virtual: Lectura de las orientaciones para la realización de las actividades prácticas planteadas, problemas y ejercicios, e interacción con los compañeros en el foro.
4. Trabajos individuales: Realización de trabajos y resúmenes guiados por cuestiones planteadas por el equipo docente. Todos los trabajos realizados se apoyarán en un escenario hipotético de acceso a información no estructurada definido al comienzo del curso.
5. Trabajos en equipo: Discusión y evaluación conjunta de trabajos.
6. Práctica: Uso de herramientas básicas de recuperación de información sobre conjuntos reducidos de documentos, o bien la implementación de un algoritmo en tareas más específicas como problemas de recomendación o agrupación de documentos. Los ejercicios prácticos se realizarán mediante software libre disponible y recursos de libre disposición.

8. BIBLIOGRAFÍA BÁSICA

ISBN(13): 9780136072249

Título: SEARCH ENGINES: INFORMATION RETRIEVAL IN PRACTICE (Primera)

Autor/es: Bruce Croft ; Trevor Strohman ; Donald Metzler ;

Editorial: PEARSON

Buscarlo en librería virtual UNED

Buscarlo en bibliotecas UNED

Buscarlo en la Biblioteca de Educación

Buscarlo en Catálogo del Patrimonio Bibliográfico

9. BIBLIOGRAFÍA COMPLEMENTARIA

Comentarios y anexos:

El acceso a la información es una línea de investigación en constante desarrollo. Por ello, el texto base se complementará con artículos académicos publicados en los principales foros del área, como los congresos:

- SIGIR: Association for Computing Machinery's Special Interest Group on Information Retrieval. Since 1963, we have promoted research, development and education in the area of search and other information access technologies.
- CLEF: Conference and Labs of the Evaluation Forum. Information Access Evaluation meets Multilinguality, Multimodality and Interaction.
- CERI: Spanish Conference in Information Retrieval.
- SPIRE: rnational Symposium on String Processing and Information Retrieval.

Y revistas especializadas como:

- Information Processing and Management (IP&M), Elsevier
- Information Retrieval, Springer
- Transactions on the Web (TWEB), ACM

10. RECURSOS DE APOYO AL ESTUDIO

Para las actividades prácticas se utilizarán lenguajes de programación estándar y herramientas de libre distribución. Se proporcionarán a los estudiantes los dataset o conjuntos de prueba (colecciones de documentos y anotaciones manuales) necesarios. Para la realización de trabajos de carácter teórico se partirá de guías elaboradas por el equipo docente sobre un escenario hipotético de acceso a la información.

11. TUTORIZACIÓN Y SEGUIMIENTO

Los alumnos serán tutorizados vía los cursos virtuales o vía telefónica en horario de tutorías, atendiendo a las consultas de carácter teórico, dudas en cuanto a la realización de los trabajos prácticos. Asimismo, se establecerán foros de discusión para temas concretos.

12. EVALUACIÓN DE LOS APRENDIZAJES

La calificación de la asignatura se obtendrá sumando, hasta un máximo de 10 puntos, de la siguiente manera:

1. Examen final: se calificará sobre un máximo de siete puntos. El examen es de carácter obligatorio, y es necesario obtener un mínimo de dos puntos para poder aprobar la asignatura.
2. Práctica: se calificará sobre un máximo de cuatro puntos. La práctica es de carácter obligatorio, y es necesario obtener un mínimo de dos puntos para poder aprobar la asignatura.
3. Evaluación continua: la participación del estudiante en las tareas colaborativas propuestas a lo largo del curso se valorará con un máximo de tres puntos. Estas tareas son de carácter voluntario.

Dicho de otra forma, la calificación de la asignatura es la suma de las calificaciones de los apartados 1, 2, y 3, hasta el límite de 10 puntos, siempre que se haya obtenido al menos un 3 sobre 7 en el examen y un 2 sobre 4 en la práctica. Suspenso, en caso contrario.

La práctica consistirá en la utilización de una herramienta básica para el desarrollo de un motor de búsqueda sobre información textual, o bien la implementación de un algoritmo en tareas más específicas como problemas de recomendación o agrupación de documentos.

13. COLABORADORES DOCENTES

Véase equipo docente.