ASIGNATURA DE MÁSTER:

MOTORES DE BÚSQUEDA WEB

Curso 2016/2017

(Código:31101042)

1.PRESENTACIÓN

Tipo	Optativa
Cuatrimestre	Primero
Créditos/horas totales	6/150
Horas de estudio teórico	100
Horas de prácticas	50
Horas complementarias	0

Esta es la guía de la asignatura "Motores de búsqueda Web" que se imparte dentro del máster en Lenguajes y Sistemas Informáticos de la UNED. En esta guía se contextualiza la asignatura dentro del máster, se especifican los conocimientos previos necesarios para cursarla con éxito, sus objetivos de aprendizaje y contenidos, y la metodología con la que se estudiará.

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". Dentro de esta especialidad, "Motores de búsqueda Web" aporta los fundamentos sobre los que aplicar tecnologías de procesamiento de textos más sofisticadas a gran escala.

3.REQUISITOS PREVIOS RECOMENDABLES

Lectura fluida del inglés y conexión a Internet, además de los requisitos propios del máster.

4.RESULTADOS DE APRENDIZAJE

Objetivos generales de la materia

En este curso se estudian los aspectos esenciales para la recuperación de información en la Web: desde la naturaleza del problema (topología de la Web y características de los usuarios) hasta los retos tecnológicos planteados en la nueva generación de buscadores, pasando por los sistemas clásicos de recuperación de información, la arquitectura básica de un buscador Web, y los sistemas de recuperación basados en notoriedad, de los que Google es el ejemplo canónico.
Al finalizar el curso, el alumno debe ser capaz de plantear la arquitectura completa de un buscador Web, y debe ser capaz de diagnosticar las limitaciones de los sistemas actuales y proponer soluciones novedosas para superarlas.

Destrezas y competencias

El alumno adquirirá las siguientes destrezas y competencias:
Debe tener una visión de conjunto de las tecnologías relacionadas con la búsqueda Web, comprendiendo su evolución temporal y los retos de investigación que se plantean en la actualidad.
Debe ser capaz de realizar una lectura crítica de artículos científicos sobre el tema, de localizar y discriminar información bibliográfica relevante, y de sintetizar información de distintas fuentes.
Debe ser capaz de redactar con rigor científico y de comunicar y debatir con pares (en este caso, sus compañeros) sus análisis y opiniones en torno a los temas de la asignatura.
Debe ser capaz de diagnosticar las limitaciones del campo de investigación en motores de búsqueda Web y apuntar caminos para superarlas.

5.CONTENIDOS DE LA ASIGNATURA

Estructura y contenido teórico

1. Características de la búsqueda de información en la WWW
- Topología de la WWW: Hubs, autoridades, islas, Internet Invisible, etc.
- Necesidades de información y búsquedas web: perfil de usuarios.
- Formas básicas de búsqueda: navegación y consulta. Directorios web versus motores de búsqueda.

2. Arquitectura básica de un motor de búsqueda.
- Crawling, Indexación, Procesado de la consulta, Recuperación, Presentación de resultados.
- Arquitectura hardware/software.

3. Motores de búsqueda pre-Google: recuperación basada en contenidos.
- Modelos tradicionales de recuperación de información (modelo booleano, modelo de espacio vectorial, modelos probabilísticos).
- Limitaciones de los modelos RI en la web: pertinencia versus autoridad, vulnerabilidad a la manipulación externa (spamdexing).

4. Motores de búsqueda actuales (generalistas): recuperación basada en autoridad.
- Autoridad absoluta: Algoritmos PageRank y HITS.
- Autoridad relativa a un tema/consulta: Hilltop, Topic Distillation.
- El motor de búsqueda Google: evolución de Pagerank (historia de URLs y enlaces, análisis de patentes de Google, Local Rank, Google Sandbox, etc), sistemas de publicidad contextual (adwords, adsense), vulnerabilidad.
- Otros motores de búsqueda generalistas.

Objetivos por tema y orientaciones breves

1. Características de la búsqueda de información en la WWW

Objetivos:
El objetivo principal del tema es que el alumno comprenda cuál es la funcionalidad de un sistema de búsqueda en la Web. Se puede dividir en subobjetivos de esta manera:
O.1.1 Comprender la estructura y naturaleza de la Web, y la importancia de los sistemas de búsqueda de información en este medio.
O.1.2 Conocer las necesidades típicas que se resuelven mediante buscadores Web.
O.1.3 Conocer los mecanismos básicos que utilizan los usuarios al buscar información en la Web.

2. Arquitectura básica de un motor de búsqueda.

Objetivos:
En este tema, el alumno debe familiarizarse con los componentes básicos de cualquier motor de búsqueda, y comprender cuáles son las implicaciones de manejar un volumen de datos inmenso para obtener respuestas en fracciones de segundo. Este objetivo se puede dividir en:
O.2.1 Conocer y comprender la funcionalidad de los componentes básicos de un motor de búsqueda.
O.2.2. Conocer y comprender la arquitectura típica hardware/software que soporta esa funcionalidad, y los problemas derivados de la escala a la que trabaja un buscador Web.

3. Motores de búsqueda pre-Google: recuperación basada en contenidos.

Objetivos:
Conocer el corpus teórico conocido como "Information Retrieval" (recuperación de información), cómo se ha utilizado en los motores de búsqueda Web, y qué limitaciones tiene en un entorno Web. Se puede dividir en:
O.3.1. Conocer los modelos tradicionales de recuperación de información.

O.3.2. Saber cómo se han aplicado a la búsqueda web, qué limitaciones tienen, y qué otras aplicaciones de estos modelos son factibles en la Web (como, por ejemplo, la inserción de publicidad contextual).

4. Motores de búsqueda actuales (generalistas): recuperación basada en autoridad.

Objetivos:
Conocer los principios teóricos y prácticos sobre los que se fundamentan los motores de búsqueda Web actuales, en particular:
O.4.1 Conocer y ser capaz de comparar los algoritmos más relevantes para calcular la autoridad de una página Web a partir de la estructura de hipervínculos de la Web (PageRank, HITS).
O.4.2 Conocer sus limitaciones, las variantes propuestas, y ser capaz de realizar análisis críticos sobre esas propuestas alternativas.
O.4.3 Conocer cómo se aplica lo anterior a los principales buscadores (Google, Yahoo, MSN, Ask), y en particular sobre Google.

5. Temas avanzados.

Objetivos:
En este tema se estudian las tendencias de la nueva generación de motores de búsqueda, con el objetivo de que el alumno sea capaz de diagnosticar los retos técnicos por resolver y proponer soluciones relativamente novedosas:
O.5.1. Conocer las corrientes de investigación más recientes en el campo de los buscadores Web.
O.5.2. Tener una panorámica de los nuevos servicios relacionados con la búsqueda en la Web.
O.5.3. Ser capaz de proponer temas relevantes sobre los que realizar el trabajo individual de la asignatura.

Actividades y plan de trabajo

1. Actividades prácticas programadas

Las tareas que se asignan en esta asignatura tienen tanto que ver con la asimilación de los conocimientos propios de la materia, como con el desarrollo de la capacidad para investigar.
Algunos de los tipos de tareas que se proponen son:

-Lectura y análisis de un artículo de investigación, contestando a preguntas como: ¿Se trata de un artículo de teoría, metodología, experimentación o aplicación? ¿Cuáles son sus aportaciones originales? ¿Cuáles son los argumentos/resultados esenciales que conducen a sus conclusiones?
-Evaluación simulada de un artículo, calificando de forma razonada su originalidad, su impacto potencial en el área, la pertinencia y completitud de las referencias bibliográficas, la calidad del trabajo (argumentos, metodología, diseño experimental, etc., la calidad de la presentación (organización, claridad expositiva, etc.). Discusión en grupo (tres alumnos) para alcanzar una única evaluación consensuada, estableciendo una figura de meta-revisor encargado de coordinar la discusión y redactar la evaluación final.
-Estudio del impacto de un artículo: ¿Cuáles son los aspectos del artículo por los que es referenciado? ¿Coinciden con los aspectos sobre los que los autores habían hecho énfasis, o son aspectos inicialmente marginales? ¿Se ha hecho algún avance sustancial respecto a las conclusiones del artículo? ¿Se han refutado las conclusiones del artículo, se han corroborado, se ha profundizado en ellas, se han propuesto vías alternativas?
-Actualización de un artículo de revisión del estado del arte, sintetizando los avances más significativos posteriores a la publicación de la revisión inicial.
Propuesta de "lecturas recomendadas" para un tema, consensuando una lista razonada a partir del debate entre todos los alumnos de la asignatura.
-Evaluación comparada de servicios de búsqueda Web alternativos, utilizando tanto la revisión bibliográfica como la experimentación directa.
-Diseño e implementación de un servicio de búsqueda Web con algún componente novedoso, partiendo de herramientas de código abierto (como Lucene) o servicios Web (como las API de Google, Yahoo, etc).

2 Otras actividades prácticas programadas

Se irán anunciando de forma dinámica en el entorno virtual.

3 Plan de trabajo

-Tema 1 (15 horas) Semanas 1-3. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

-Tema 2 (15 horas) Semanas 4-5. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

-Tema 3 (20 horas) Semanas 6-8. Estudio de materiales de referencia y ejercicios relacionados con la

consulta bibliográfica.

-Tema 4 (25 horas) Semanas 9-12. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica.

-Tema 5 (25 horas) Semanas 13-16. Estudio de materiales de referencia y ejercicios relacionados con la consulta bibliográfica. Determinación del trabajo individual en coordinación con el equipo docente.

Trabajo individual (50 horas). Semanas 16-23.

6.EQUIPO DOCENTE

- JULIO ANTONIO GONZALO ARROYO
- JUAN MARTINEZ ROMO

7.METODOLOGÍA

La general del programa de posgrado. En particular, el alumno realiza dos tipos de actividades en esta asignatura: las relacionadas con la consulta bibliográfica y las de implementación y experimentación. Las primeras son comunes a todos los alumnos y están fijadas dentro del material de estudio correspondiente a cada tema. En una segunda parte de la asignatura, cada alumno realiza un trabajo individual sobre un tema acordado con el equipo docente. Todo el material de estudio está disponible en el entorno virtual del posgrado, y toda la interacción entre profesores y alumnos se puede llevar a cabo en este entorno.

8.BIBLIOGRAFÍA BÁSICA

Comentarios y anexos:
Arvind Arasu, Junghoo Cho, Hector García-Molina, Andreas Paepcke and Sriram Raghavan. Searching the Web. ACM Transactions on Internet Technology, vol. 1, n. 1, August 2001, pages 2-43.

9.BIBLIOGRAFÍA COMPLEMENTARIA

Comentarios y anexos:
Tema 1. Características de la búsqueda de información en la WWW
Sobre estructura de la WWW:
- Kleinberg, JM. Hubs, authorities, and communities, ACM computing surveys 1999.
http://www.cs.brown.edu/memex/ACMCSHT/10/10.html
- A Borodin, GO Roberts, JS Rosenthal, P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. Proc. WWW 2001.
http://www10.org/cdrom/papers/314/
Sobre tipología de búsquedas web:
- Rose, D. y Levinson, D. Understanding User Goals in Web Search. WWW 2004.
http://wwwconf.ecs.soton.ac.uk/archive/00000537/01/p13-rose.pdf
Sobre navegación versus consulta:
- Marti A. Hearst. Next Generation Web Search: Setting Our Sites In IEEE Data Engineering Bulletin, 2002.
http://www.sims.berkeley.edu/hearst/papers/data-engineering
- A. Peñas, F. Verdejo, J. Gonzalo, 2002. Terminology Retrieval: towards a synergy between thesaurus and free text searching. Advances in Artificial Intelligence - IBERAMIA 2002, LNAI 2527.
http://nlp.uned.es/pergamus/pubs/iberamia2002.pdf

Tema 2. Arquitectura básica de un motor de búsqueda.
Sobre crawling:
- J Cho, H Garcia-Molina, L Page. Efficient Crawling Through URL Ordering, WWW 1998.
- Allan Heydon and Marc Najork. Mercator: A Scalable, Extensible Web Crawler. In Proceedings of World Wide Web Conference, 1999, pages 219-229.
Sobre soporte hardware:
- L. A. Barroso, J. Dean, U. Hoelzle. Web search for a planet: the Google cluster architecture. IEEE 2003.

Tema 3. Motores de búsqueda pre-Google: recuperación basada en contenidos.

- D Hiemstra. Using Language Models for Information Retrieval. CTIT Ph.D. Thesis, 2001.
- G Salton, A Wong, CS Yang. A Vector Space Model for Automatic Indexing. Comm. ACM, 1975.
- N Fuhr. Probabilistic Models in Information Retrieval. The Computer Journal, 1992.

Tema 4. Motores de búsqueda actuales (generalistas): recuperación basada en autoridad.
Referencias:
- M Hollander. Google's PageRank Algorithm to Better Internet Searching. TR UMN.
- Brin, S. y Page, L. The Anatomy of a Large-Scale Hypertextual Web
Search Engine. WWW 1998.
- CHQ Ding, X He, P Husbands, H Zha, HD Simon. PageRank, HITS and a unified framework for link analysis. SIGIR 2002.
TH Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE T. on Knowledge and data engineering, 2003.

Tema 5. Temas avanzados.
- Guha, R. y Garg, A. Disambiguating People in Search. Proc. WWW 2004.
- S Lawrence, NJ Princeton. Context in Web Search, IEEE data engineering
bulletin, 2000.
J Sivic, A Zisserman. Video google: A text retrieval approach to object matching in videos, ICCV 2003.
- SK Bhavnani, CK Bichakjian, TM Johnson, RJ Little. Strategy Hubs: Next-Generation Domain Portals with Search Procedures. Proc. ACM Conference on Human Factors in Computing Systems, 2003, ACM Press NY, USA.
- T Berners-Lee, J Hendler, O Lassila. The semantic Web. Scientific American, 2001.
- J Heflin, J Hendler. A Portrait of the Semantic Web in Action. IEEE Intelligent Systems, 2001.
- S Eissen, B Stein. Analysis of Clustering Algorithms for Web-Based
Search. Springer-Verlag, 2002.
- J. Cigarrán, A. Peñas, J. Gonzalo, F. Verdejo, 2005. Automatic selection of noun phrases as document descriptors in an FCA-based Information Retrieval system. ICFCA 2005. Springer LNCS 3403.
Search Engines: Technology, Society, and Business. Materiales online del curso:
http://www.sims.berkeley.edu/courses/is141/f05/schedule.html

10.RECURSOS DE APOYO AL ESTUDIO

La plataforma de enseñanza virtual de posgrados de la UNED será la interfaz de interacción entre el alumno y sus profesores. Esta plataforma permite impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.

11.TUTORIZACIÓN Y SEGUIMIENTO

Se realizará mediante la plataforma de posgrados de la UNED.

12.EVALUACIÓN DE LOS APRENDIZAJES

La evaluación se realizará a partir de las actividades realizadas en cada tema y el trabajo individual de cada alumno.

13.COLABORADORES DOCENTES

Véase equipo docente.