CC52D: Recuperación de Información

Profs: Ricardo Baeza Yates y Carlos Castillo

URL: www.baeza.cl/cursos/cc52d.html

Objetivo

Estudiar cómo encontrar información en grandes bases de datos textuales, semi o no estructuradas, y sus aplicaciones, en particular la Web.

Evaluación

- 2 controles y 1 examen con igual ponderación (70% de la nota y debe ser ≥ 4).
- Tareas (30% de la nota y debe ser ≥ 4).
  Castigo por atraso: 1 punto por día hábil o fin de semana.

Programa

El contenido tiene tres partes: teoría de recuperación de información, recuperación de información en la Web y recuperación de información en datos semiestructurados.

Introducción

El problema de recuperación de información. Conceptos básicos. Historia. Recuperación vs. navegación. Aplicaciones. Datos estructurados vs. datos semi o no estructurados. Nociones de XML.

Teoría de Recuperación de Información

Modelos de jerarquización de relevancia: booleano, vectorial, etc. Modelos de navegación. Precisión vs. recuperación. Evaluación de calidad: colecciones de referencia.

Operaciones booleanas. Otros tipos de consultas. Búsqueda aproximada. Expansión de la consulta. Operaciones sobre el texto. Agrupación de documentos.

Indices: Archivos invertidos y arreglos de sufijos. Búsqueda y resolución de consultas para cada caso. Uso de compresión.

Interfaces de sistemas de recuperación de información: funcionalidades e interacción requeridas. Visualización de documentos y conjuntos de documentos.

Recuperación de Información en la Web

¿Cómo es la Web? Arquitectura de un buscador Web. El recolector de páginas. Sistema de indexación, consultas y ranking. ¿Cómo pueden explotarse los enlaces? Búsqueda multimedia, metabuscadores, extensiones.

Recuperación de información en datos semiestructurados

Documentos semiestructurados y su uso. Lenguajes de marcas: SGML y XML. XML desde el punto de vista del programador. Referencias: XLink. Transformaciones: XSLT. Esquemas en XML. Consultas sobre XML: Xquery. Almacenamiento de documentos en XML.

Bibliografía

Todos los libros están en biblioteca.

- Agosti, M. y Smeaton, A. (editores) Information Retrieval and Hypertext, Kluwer, 1996.
- Baeza-Yates, R. y Ribeiro-Neto, B. Modern Information Retrieval, Addison-Wesley 1999. Ver en `sunsite.dcc.uchile.cl/irbook/`
- Abiteboul, S., Buneman, P. y Suciu, D. Data on the Web: from Relations to Semistructured Data and XML, Morgan Kauffman, 2000.
- Search Engine Watch, www.searchenginewatch.com, 2002.
- Witten, I., Moffat, A. y Bell, T. Managing Gigabytes, Morgan Kauffman, 1999 (segunda edición).
- World Wide Web Consortium, w3c.org, 2002.