CC52D: Recuperación de Información

Profs.: Ricardo Baeza Yates y Carlos Castillo

IMPORTANTE

Los anuncios del curso son publicados en `news:uch.ing.cursos.cc52d`.
Las notas y apuntes del curso son publicadas en `http://ucursos.ing.uchile.cl/`.

1. Objetivo

Estudiar cómo encontrar información en grandes bases de datos textuales, semi o no estructuradas, principalmente en la Web.

2. Evaluación

- 1 control a mediados de semestre (35 %) y 1 examen (35 %). El promedio entre ambas pruebas debe ser ≥ 4.
- Tareas (30 %). Debe ser ≥ 4. Descuento por atraso: 1 punto por día hábil, 1 punto por fin de semana.

3. Programa

3.1. Introducción

El problema de recuperación de información. Conceptos básicos. Historia. Recuperación vs. navegación. Aplicaciones.

3.2. Modelos y Evaluación

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

3.3. Lenguajes de Consulta y Procesamiento del Texto

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

3.4. Algoritmos y estructuras de datos

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

3.5. Búsqueda en la Web

Característsticas de la web global y de la web chilena. Componentes de un buscador Web. Recolectores de páginas Web. Análisis de enlaces. Extensiones: metabuscadores, multimedia.

3.6. Repositorios semiestructurados

Datos estructurados vs. datos semi o no estructurados. Lenguajes de marcas: SGML y XML. Espacios de nombre. Lenguajes de Schema. Expresiones de ruta XPath. Lenguajes de consulta en XML. Transformación de documentos XML usando XSLt. Bases de datos XML.

3.7. Visualización

Interfaces de sistemas de recuperación de información. Visualización de conjuntos de documentos.

4. Breve explicación de la tarea

La tarea consiste en implementar una aplicación para buscar en una colección de texto. La aplicación consiste en dos programas: uno para crear un índice del texto entregado y uno para realizar consultas.

Normalmente se realizan 1 o 2 entregas parciales durante el semestre más la entrega final de la tarea.

Los errores más comunes son:

- Comenzar a hacer la tarea demasiado tarde.
- No saber como separar un texto en palabras, o intentar utilizar expresiones regulares para hacerlo.
- Intentar implementar un parser XML, o utilizar un parser desconocido o mal documentado.
- No probar la tarea adecuadamente en la máquina de referencia, o enviarla por e-mail y esperar que el corrector la porte a linux.
- No entregar la tarea $\Rightarrow R$.

La tarea se hace normalmente en Java, Perl o C/C++, y se provee de un formato estándar de entrada y salida para poder comparar los programas. Además se entrega instalada en una máquina específica (un Linux) a la que se les dará acceso durante el semestre.

5. Bibliografía

En el sitio de u-cursos hay material docente para el curso. Los "apuntes parciales del curso" en formato Postscript son principalmente material de la primera mitad del curso, y las slides en formato OpenOffice.org son principalmente de segunda mitad.

- Baeza-Yates, R. y Ribeiro-Neto, B. Modern Information Retrieval, Addison-Wesley 1999. Ver en `sunsite.dcc.uchile.cl/irbook/`
- Witten, I., Moffat, A. y Bell, T. Managing Gigabytes, Morgan Kauffman, 1999 (segunda edición).