Práctica Nº 2: 2 SESIONES

Estructura de la información y recuperación de documentos en la Web

0. OBJETIVOS

El objetivo de la práctica es ver la utilidad práctica de las estructuras de datos de grafos y árboles, en concreto los árboles binarios de búsqueda, y ver sus ventajas respecto a otros tipos de estructuras. También se pretende que los alumnos se familiaricen con algunos conceptos de los ficheros, tales como registros y campos.

1. INTRODUCCIÓN

En la Web podemos encontrar una enorme cantidad de documentos. Para realizar búsquedas de documentos que tengan unas determinadas características, los sistemas de recuperación de la información realizan previamente algún tipo de descripción de la forma y del contenido de los documentos, además de almacenar información sobre su ubicación.

En esta práctica vamos a analizar una serie de documentos Web obteniendo alguna información relevante sobre ellos que nos permita clasificarlos y recuperarlos posteriormente.

2. REALIZACIÓN DE LA PRÁCTICA.

La práctica se realizará en dos partes. La primera consistirá en analizar varias páginas web para clasificarlas y describir su contenido, y la segunda en simular búsquedas a partir de palabras clave.

2.1. Análisis de documentos y sus interrelaciones.

Se proporciona al alumno una lista de 15 páginas web, todas ellas relacionadas con temas similares. De cada uno de los documentos proporcionados debemos crear una ficha con los datos que nos servirán para identificar y describir el documento. Con ellos crearemos una tabla, con la información que sería almacenada dentro lo que llamaremos 'fichero de documentos'. Este fichero contendrá las fichas o *registros* de cada uno de los documentos. Cada registro contendrá una serie de campos, que serán:

- Identificador del documento (número entero empezando por 1).
- URL, o dirección de la página Web.
- Título de la página que ha puesto el autor.

ID	URL	Título
1	http://www.uv.es	….
2	http://….	….

Tabla 1. Fichero de documentos.

Además para cada página realizaremos dos tablas más. Una lista con las palabras clave de la página, que más adelante nos servirá para localizar documentos. Estas palabras clave hay que buscarlas en primer lugar en las etiquetas META de la cabecera de la página que ha creado el autor, y caso de no haberlas, habrá que poner las palabras clave que se consideren más apropiadas. Para simplificarlo, como máximo habrá que poner 5 palabras clave para cada página. Si en las etiquetas META aparecen más de cinco se seleccionarán las que se consideren más adecuadas.

ID	Palabra clave
1	…
1	….
…	..
2	…
2	….

Tabla 2. Palabras clave de documentos.

Por último por cada página una tabla con las categorías y subcategorías en las que lo clasificamos:

ID	Categoría	Subcategoría
1	…	….
1	…	….

Tabla 3. Categorías.

Parte 2. Localización de documentos a partir de palabras clave.

Es habitual tratar de recuperar documentos en Internet a partir de una o varias palabras que identifican la información que buscamos. Vamos a ver cuál sería el proceso de recuperación de alguno de nuestros documentos.

En primer lugar, como la búsqueda se realizará a partir de las palabras clave, sería costoso recuperar los documentos que contienen una palabra clave a partir del fichero de documentos que tenemos. Para facilitar el proceso vamos a crear un nuevo fichero en el que únicamente tendremos las palabras clave y los documentos en los que aparece. Par ello vamos a crear una nueva tabla, con dos columnas, la primera con las palabras ordenadas alfabéticamente, y la segunda con los identificadores de los documentos en los que aparece dicha palabras. De esta forma, a partir de una palabra clave podemos localizar en el fichero de documentos los documentos a los que pertenece, su URL, etc.

Palabra Clave	**ID Documentos**
viajes	3,6,9

Tabla 4. Tabla inversa de palabras clave.

La primera columna creada, vemos que puede corresponder a una ***lista ordenada*** de palabras, donde cada palabra clave está en un nodo y el nodo siguiente será el siguiente elemento de la columna.

Supongamos que alguien nos pide que busquemos las páginas en las que aparece información de cada una de las siguientes palabras clave:

- Medulas.
- Arqueología.
- Bierzo.
- Rutas.

→ Para cada una de ellas, ¿cuántos nodos de la lista hay que recorrer hasta encontrarla? ¿Cuáles son esos nodos?

A continuación vamos a mejorar el modo de búsqueda de palabras clave. Para ello hay que crear un *árbol binario de búsqueda* con las palabras. Cada uno de los nodos del árbol contendrá como datos, la palabra clave, y los ID de los documentos en los que aparece.

→ Dibujar el árbol binario de búsqueda resultante.

→ Para cada una de las palabras claves buscadas previamente, decir cuál será el recorrido en el árbol hasta encontrarla (por cuántos y cuáles nodos se pasará).

→ Comparar el resultado de esta búsqueda con el de la búsqueda en la lista. Para cada una de las palabra, ¿qué búsqueda ha sido más eficiente?.

→ Por último, describir el proceso completo de búsqueda de documentos de un tema. A partir de una palabra clave seleccionada entre las existentes, decir cuáles serán los pasos que seguirá nuestro sistema de recuperación de información y qué documentos nos devolverá.

Listado de páginas web:

1. http://www.info-red.net/base_datos/1.4_medio_ambiente/bosques.html
2. http://www.la-muela.com/
3. http://www.info-red.net/base_datos/2.16_turismo/index.html
4. http://www.turismobierzo.com/lugares/medulas.htm
5. http://www.tecnociencia.es/ventana/salas/paisajes/medulas/m_ini.htm
6. http://www.guiadelbierzo.com/las.htm

7. http://www.arrakis.es/~seha/medulas.htm
8. http://www.cerespain.com/peraconferenciadelbierzo-vinculos.html
9. http://club.telepolis.com/geografo/geomorfologia/acctantro.htm
10. http://www.fundacionlasmedulas.com/index.jsp?idioma=es
11. http://www.unav.es/dpp/tecnologia/proyectos2004/18/index.htm
12. http://www.ruralagoga.com/entorno.htm
13. http://www.caminosantiago.com/web/ecosistemas4.htm
14. http://www.infraestructuras-ferroviarias.com/empresa/medio_ambiente/ponferrada.html
15. http://club.telepolis.com/damalago/bierzo.htm

3. ENTREGA DE MEMORIAS

Las memorias de las prácticas se entregarán por correo electrónico **desde la cuenta de correo electrónico de la Universidad de Valencia** al profesor asignado antes de la sesión de comienzo de la práctica 3. En el mensaje deberán incluirse los nombres, apellidos y DNI de los integrantes del equipo de prácticas.

El profesor, tras comprobar el documento, enviará un mensaje de confirmación en las cuarenta y ocho horas siguientes. Si el alumno no recibiese dicho mensaje de confirmación en el plazo citado deberá ponerse en contacto con el profesor.