



PRÁCTICA Nº 7: 1 sesión

(del 26 al 30 de Mayo de 2003)

BÚSQUEDA DE PALABRAS EN DOCUMENTOS DE TEXTO: UTILIZACIÓN DE ÁRBOLES BINARIOS DE BÚSQUEDA

0. OBJETIVOS

El objetivo de esta práctica es la implementación de la clase Arbol binario de búsqueda y su uso para localizar palabras en diferentes documentos de texto.

1. INTRODUCCIÓN

Cada vez tenemos mayor necesidad de localizar documentos que contengan una información específica (p.e. páginas web). En ese caso deberemos buscar dentro de cada documento aquellas palabras que identifiquen la información buscada y mostrar al usuario la relación de documentos en dónde hayamos encontrado dichas palabras.

Esta técnica es la que se suele utilizar en los buscadores de internet.

En esta práctica partiremos de un conjunto de documentos y prepararemos una estructura para poder buscar fácilmente las palabras incluidas en los mismos.

Una vez preparada la estructura realizaremos un buscador que nos diga en que documentos aparecen las palabras que indiquemos.

2. ESPECIFICACIÓN DE LAS ESTRUCTURAS DE DATOS

La estructura que utilizaremos para guardar las palabras contenidas en cada documento será un **árbol binario de búsqueda** con las operaciones básicas descritas en teoría.

Como vamos a necesitar guardar diferentes árboles binarios de búsqueda (uno por cada uno de los documentos que tengamos) necesitaremos poder guardar todos estos árboles en una estructura dónde la información que guardemos será por un lado el árbol binario con las palabras y por otro el nombre del fichero al que corresponde. Esta estructura puede ser una **lista**.

De esta forma el tipo Valor contenido en la lista quedará como sigue:

```
struct Valor
{
    string nombre_documento;
    Arbol arb;
};
```

3. REALIZACIÓN DE LA PRÁCTICA

En esta práctica existen claramente dos partes bien diferenciadas: Una primera parte de generación de la estructura (lista) y una segunda parte de búsqueda en la estructura.

a.- Generación de la estructura.

El buscador que vamos a programar buscará palabras entre un conjunto de documentos que especificaremos mediante un fichero de texto que dirá qué ficheros vamos a indexar.

El programa al arrancar, pedirá el fichero dónde se encuentra este listado y generará para cada uno de los ficheros contenidos en él un árbol binario de búsqueda que se insertará en la lista.

El formato del fichero de listado se limitará a contener en diferentes líneas los nombres de los ficheros que queremos indexar. Por ejemplo:

```
c:\tmp\documento.txt
c:\tmp\mates\La_aritmetica_cardinal.txt
c:\tmp\mates\Las_paradojas.txt
c:\tmp\practica4\pr_01_2003.txt
c:\tmp\practica4\pr_02_2003.txt
c:\tmp\practica4\pr_03_2003.txt
```

Consideraremos palabras a insertar en el árbol como todo conjunto de caracteres que contengan exclusivamente letras ('a'..'z', 'A'..'Z', 'ñ' y 'Ñ'). En el proceso de inserción se pasarán todas las letras a mayúsculas. No insertaremos palabras compuestas (con espacios en blanco ni '-'), ni palabras que contengan símbolos que no sean letras (números, comas, puntos, etc.). Cualquier símbolo distinto de las letras, lo consideraremos separador de palabras.

b.- Búsqueda de palabras.

Una vez generada la lista, el programa pedirá una palabra y nos dirá en que documento o documentos aparece la palabra indicada y cuantas veces en cada uno de ellos.

Para ello recorrerá la lista donde hemos guardado el bosque de árboles e irá buscando en cada árbol la palabra. Si encuentra la palabra, nos mostrará por pantalla el nombre del fichero en dónde se encuentra esa palabra.

c.- Tarea adicional. Búsqueda de varias palabras.

Se trata de mejorar el buscador, pidiendo varias palabras y mostrando los nombres de los documentos en los que se encuentren todas las palabras indicadas y el número de veces que aparece cada una de ellas.

4. ENTREGA DE PROGRAMAS

Al comienzo de la siguiente sesión de prácticas se entregarán al profesor tres ficheros:

- 1) Ficheros de la clase `ArbolBB` (`##ArbolBB.cpp` y `##ArbolBB.h`)
- 2) Ficheros de la clase `Lista` (`##Lista.cpp` y `##Lista.h`). Deben utilizarse los ficheros de la práctica 5, debidamente modificados.
- 3) Programa principal donde se genera el bosque y se realizan las búsquedas (`##Buscador.cpp`).

Nota Muy Importante

Antes de poder empezar a realizar cualquiera de las prácticas **es necesario** presentar las hojas de especificación de programas (documentación de programas) con las tareas que se van a realizar en la práctica, explicando brevemente como se van a solucionarse los problemas que se plantean.

ENTREGA DE PROGRAMAS: Al comenzar la sesión de prácticas del 2 al 6 de Junio.