Árboles Genéricos

Casi todos los SO almacenan ficheros en árboles, los directorios se almacenan en nodos internos, mientras que el resto de los archivos en hojas.

Los hijos de un nodo pueden variar por lo que cada uno en vez de tener hijoDer y hijoIzq tiene una lista **enlazada de hijos**. Para recorrer se usa el método **primer hijo, siguiente hermano**: el nodo guarda referencia a su hijo más a la izquierda (en caso de no ser hoja) y otra a su hermano a la derecha.

Grado: número de subárboles de un nodo

Nodo terminal: nodo de grado cero “hoja”

Nodo interno: aquel que no es terminal

Diagram

Description automatically generated

En un sistema de directorios cada / indica una arista del árbol. Si el camino comienza en la raíz del sistema completo de ficheros entonces es un nombre de camino completo. En caso contrario es un nombre de camino relativo.

Tries

Un trie es una estructura de datos basada en árboles para almacenar cadenas con el fin de respaldar la coincidencia rápida de patrones. La principal aplicación es la recuperación de información (coincidencia de patrones y coincidencias de prefijos).

Trie estándar (ningún string es prefijo de otro)

Sea **S** un conjunto de **s** cadenas del alfabeto o, tal que ninguna cadena es prefijo de otra. Un trie estándar para S es un árbol ordenado con las siguientes propiedades:

* El ordenamiento de los hijos de un nodo interno está determinado por un ordenamiento de o.
* Hay **s** nodos externos, cada uno asociado con una cadena de S, tal que el camino de la raíz al nodo externo produce la cadena **s**.
* Cada nodo excepto la raíz esta etiquetado con un carácter de o.
* Un nodo interno puede tener entre 1 y **d** hijos, donde **d** la longitud del alfabeto.
* Si hay dos caracteres en o, entonces será un árbol binario.
* La altura del trie es igual a la longitud de la cadena más larga en S.
* El numero de nodos es O(n).

El orden es O(d.m) donde **d** es el tamaño del alfabeto y **m** el de la cadena.

Mediante el uso de un diccionario se puede pasar a O(log) o O(1).

Operaciones primarias:

* Comparación de patrones
* Comparación de prefijos
* Búsqueda de palabras completas
* Búsqueda de prefijos
* Inserción

Trie comprimido

Es un trie estándar donde los nodos internos tienen al menos 2 hijos.

Los nodos se etiquetan con sub-cadenas en lugar de con caracteres.

La cantidad de nodos internos de T con L hojas será como máximo L-1

Si s es la cantidad de strings en S, entonces el tamaño de T será O(s)

La ventaja es que el número de nodos es proporcional al número de cadenas y no a su longitud total.

Un trie comprimido es ventajoso cuando se usa como estructura auxiliar y no se requiere almacenar los caracteres de la cadena en una colección.

Un nodo redundante es cuando es hoja.

Algoritmos de procesamiento de texto:

Método de fuerza bruta: es ineficiente pero muy aplicable

Para la compresión de texto se usa el método codicioso (greedy method)

Para la similitud de texto se usa la programación dinámica para convertir problemas de tiempo exponencial a lineal.

Cadena de caracteres: para permitir nociones generales de una cadena de caracteres, usamos el símbolo o para denotar el conjunto de caracteres. Asumimos que el tamaño del alfabeto o, denotado (o) es una constante fija.

Clase String 🡪 cadena inmutable

Clase StringBuffer 🡪 cadena mutable