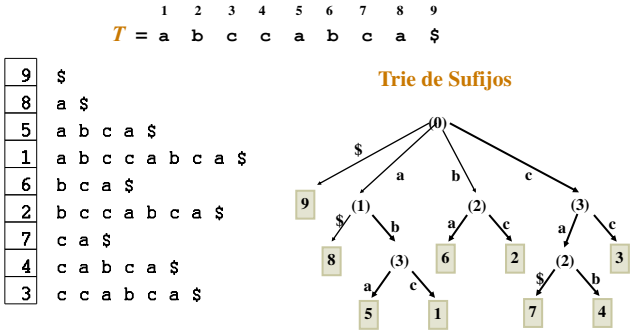
Técnicas de Indexación para Base de Datos Avanzadas

N. Herrera, D. Ruano, P. Azar, D. Welch, L. Speranza, A. De la Torre
Universidad Nacional de San Luis
A. De Battista, A. Pascal
Universidad Tecnológica Nacional Entre Ríos

Este trabajo se desarrolla en el ámbito de la línea Técnicas de Indexación para Datos no Estructurados del Proyecto Tecnologías Avanzadas de Bases de Datos, cuyo objetivo principal es realizar investigación básica en problemas relacionados al manejo y recuperación eficiente de información no tradicional. Forma parte del desarrollo de una Tesis Doctoral, dos Tesis de Maestría y un Trabajo Final de Licenciatura

Bases de Datos no Estructurados

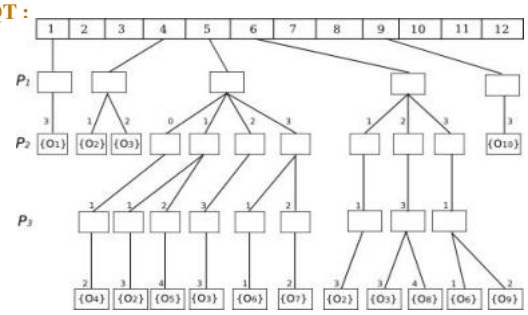
Una **Base de Datos de Texto** es un sistema que mantiene una colección grande de texto y provee acceso rápido y seguro al mismo.



El **modelo de espacios métricos** permite formalizar el concepto de búsqueda por similitud en bases de datos no tradicionales.

Una **Base de datos métrico-temporal** permite almacenar objetos no estructurados con tiempos de vigencia asociados y realizar consultas por similitud y por tiempo en forma simultánea.

H-FHQT :



Índices Comprimidos en Memoria Secundaria para BDT

Técnica de Páginación: Basadas en representación compacta del índice.

- Particionamos el trie en componentes conexas, denominadas partes, cada una de las cuales se almacena en una página de disco .

Casos de particionado: Sea ***x*** el nodo corriente a procesar.

- **Caso 1**: ***x*** y su primer hijo de mayor profundidad ***d*** entran en una pagina de disco.
- **Caso 2**: ***x*** y su primer hijo de mayor profundidad ***d*** no entran en una pagina de disco.

Índices en Memoria Secundaria para BDMT

Técnica de Paginación: Consideramos los siguientes casos:

- **Caso 1**: La lista de instantes de tiempos válidos entra en memoria primaria pero cada árbol correspondiente a cada instante de tiempo reside en memoria secundaria. Hay dos situaciones a tener en cuenta:
  - 1a: Cada árbol FHQT entra en una página de disco.
  - 1b: Cada árbol FHQT no entra en una página de disco.
- **Caso 2**: Ni la lista de instantes de tiempos válidos ni cada uno de los árboles FHQT correspondientes a cada instante de tiempo entran memoria primaria.

Aplicación de índices métricos al comercio electrónico

• El objetivo es agilizar las búsquedas por similitud sobre un conjunto de productos disponible en la plataforma Mercado Libre.

• Utilizamos el enfoque de indexación basado en pivotes para resolver la búsqueda de productos similares.

•La función de distancia utilizada es la distancia de edición (o Levenshtein).

Resultados obtenidos:

- A medida que aumentamos la cantidad de pivotes mejora el comportamiento del índice
- La selección incremental no siempre mejora a la selección aleatoria.
- El rango de búsqueda apropiado fue r= 23.
- Los espacios métricos obtenidos en este ámbito son de alta dimensionalidad.

Líneas de Investigación

Objetivo: Obtener índices eficientes, tanto en espacio como en tiempo, para el procesamiento de consultas en bases de datos textuales, espacios métricos temporales y espacios métricos.

Índices Comprimidos en Memoria Secundaria para BDT: Árbol de sufijos, representación secuencial, más una técnica de paginación para memoria secundaria. Diseño de mejoras a la técnica de páginación. Técnicas de compresión.
Índices en Memoria Secundaria para BDMT: Adecuar el índice H-FHQT para que el mismo resulten eficientes en memoria secundaria.
Índices en Memoria Secundaria para BDM: Diseñar un índice en memoria secundaria, para luego implementar un sistema de recomendación basado en espacios métricos.

 RedUNCI - UNdeC. ISBN 978-987-24611-4-0