Laboratorio 6: Recuperación por Ranking con PostgreSQL (GIN, GiST)

Prof. Heider Sanchez
ACLs: Ana María Accilio, Sebastián Loza

Introducción

En este laboratorio exploraremos índices especializados como GIN y GiST para mejorar el rendimiento de búsquedas textuales. A través de varios experimentos prácticos, analizaremos cómo estos índices impactan en la eficiencia de consultas sobre grandes volúmenes de datos, utilizando estructuras representativas como trigramas y vectores de pesos. El objetivo es comprender las ventajas y limitaciones de cada técnica en escenarios reales de recuperación y ranking de información.

P1. (2 puntos) Escaneo Secuencial vs Índice GIN

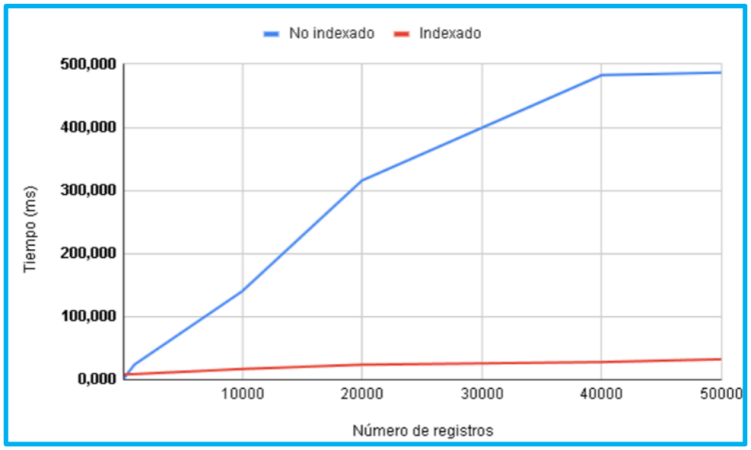
Este primer experimento tiene como objetivo comparar el rendimiento entre un búsqueda secuencial y el uso del índice invertido GIN, utilizando representación de texto mediante `trigramas`.

> Un trigrama es una secuencia de tres caracteres consecutivos dentro de una cadena. Por ejemplo, los trigramas de la palabra "amor" son "amo" y "mor". Indexar atributos textuales con trigramas mejora significativamente las búsquedas textuales en corpus.

Documentación oficial de pg_trgm en PostgreSQL

Actividades:

- Crear una tabla con dos columnas de tipo texto: una sin índice y otra con índice basado en trigramas.
- Insertar datos aleatorios con distintos volúmenes (10^2, 10^3, ..., 10^7).
- Ejecutar consultas sobre ambas columnas y registrar los tiempos de ejecución.
- Analizar y presentar el plan de ejecución.
- Generar un gráfico comparativo como resultado del experimento (ver ejemplo de referencia).



P2. (6 puntos) Búsqueda de Texto Completo en Películas

En este experimento se aplicará un índice GIN sobre atributos textuales de la tabla `film` de la base de datos `dvdrental`.

Actividades:

- Restaurar la base de datos `dvdrental` en el servidor PostgreSQL.
- Crear un nuevo atributo `full_text_idx` que concatene el título y la descripción de cada película, y convertirlo en un vector de pesos.
- Ejecutar consultas de texto completo sobre el atributo sin índice y sobre el atributo indexado, asegurándose de obtener resultados.

Aspectos a evaluar:

- Medir los tiempos de respuesta para distintas versiones de la tabla (replicando registros).
- Evaluar el impacto del ranking (top-k) en el tiempo de ejecución. ¿Se observan diferencias significativas?
- Confirmar que el índice está siendo utilizado en las consultas.

Database "dvdrental"

P3. (6 puntos) Búsqueda de Texto Completo en Noticias

Este experimento evalúa el rendimiento del índice GIN sobre atributos textuales en la tabla `articles`, basada en el dataset "All the News".

Actividades:

- Crear la tabla `articles` e importar los datos desde archivos CSV.
- Generar un nuevo atributo indexado que combine el título y el contenido de cada noticia, transformado en un vector de pesos (por ejemplo, `to_tsvector(title || ' ' || content)`).
- Ejecutar consultas sobre el atributo sin índice y el indexado, verificando que ambos arrojen resultados.

Aspectos a evaluar:

- Medir los tiempos de ejecución para diferentes tamaños de tabla (generar subconjuntos de datos aleatorios).
- Evaluar el impacto de diferentes valores de `top-k` en los tiempos. ¿Se aprecian diferencias significativas?
- Verificar que el índice se esté utilizando correctamente.
- Presentar el plan de ejecución y un gráfico comparativo como parte del análisis.

Dataset "All the News"

P4. (6 puntos) Comparación de Índices GIN vs GIST en Noticias

En este cuarto experimento se comparará el rendimiento de los índices GIN y GIST aplicados a textos, utilizando el mismo dataset de noticias (`articles`).

Actividades:

- Crear dos versiones de la tabla `articles`, una con un índice GIN y otra con un índice GIST sobre el mismo atributo de texto completo.
- Asegurarse de usar el mismo conjunto de datos para ambas tablas.

Aspectos a evaluar:

- Medir y comparar el **tiempo de creación** de los índices GIN y GIST.

- Medir y comparar el **tiempo de inserción** de nuevas filas en ambas versiones.
- Ejecutar consultas de búsqueda textual y comparar el **tiempo de respuesta** en cada caso.
- Comparar el **tamaño en disco** ocupado por ambos tipos de índice.
- ¿Cuál técnica ofrece mejor rendimiento general?

Entregables:

- Informe en PDF y los scripts SQL como evidenia
- Imágenes de los planes de ejecución obtenidos.
- Resultados experimentales en forma de tabla y gráfico.
- Análisis de resultados