Pauta para la última entrega

Instrucciones

En esta última entrega se pide que, dada unas consultas de complejidad considerable, se comparen el rendimiento de la base de datos cuando hay índices árbol B+, hashing, y simplemente no hay índices. El rendimiento se debe estudiar para distintos tamaños de las tablas de la base de datos.

Detalle de los experimentos

Primero, es necesario elegir consultas adecuadas en cuanto a su complejidad. Cada consulta debe tener un *plan de índices* (cuáles serán los atributos indizados) que use árbol B+ y otro que use hashing. Estos planes deben ser adecuados, naturalmente.

Luego, es necesario hacer los experimentos. Se desea conocer cuánto demora una consulta con y sin índices, para tamaños diferentes de las tablas: para cada plan de índice (o sin índices) y tamaño de la base de datos, se debe ejecutar cada consulta varias veces y obtener el costo promedio por consulta. Se debe tener cuidado con el caché de la base de datos, pues puede reducir los costos de las consultas, prescindiendo de los algoritmos usados para las consultas.

Consultas

Las consultas deben leer al menos tres tablas de la base de datos. También, deben considerar consultas anidadas o agrupación. Estas últimas se pueden evitar si se leen más tablas en una consulta simple (5 ó 6 tablas).

Elección de los índices

Cada plan de consulta debe consistir sólo de un tipo de índices. Los atributos indizados deben ser los adecuados, para que se refleje el uso correcto de los algoritmos por parte del optimizador.

Tamaño de las tablas

Se deben parametrizar las tablas según tamaño: desde 100 hasta 2500 filas (tuplas) o más si se estima conveniente. Para cada tamaño de estas tablas, construidas con datos aleatorios, se debe ejecutar cada consulta varias veces, registrando el tiempo que tarda cada consulta.

Presentación de los resultados

Los datos se deben presentar con gráficos y comentarios. Se deben descubrir puntos de inflexión para determinar cuándo se usa un algoritmo u otro. También se pueden presentar tablas como complementos a los gráficos.

Análisis de los datos

Éste es uno de los puntos más importantes, pues determina casi toda la nota (junto a la calidad de los experimentos y su adecuada presentación). Se debe relacionar la materia vista en clases sobre almacenamiento de datos y optimización de consultas con los resultados obtenidos.

Dump de los datos

Es necesario adjuntar en el informe **sólo una porción** del *dump* de la base de datos. Esto para validar el trabajo realizado. Entregue sólo una porción, unas 4 ó 5 páginas, nada más.

PostgreSQL

La base de datos a usar es PostgreSQL y está en Anakena y Dichato. También se puede descargar por separado, pues es software libre.

Para montar una copia local de PostgreSQL en Unix/GNU Linux use:

```
export PGDATA=~/dirBD/
initdb       #se ejecuta una vez, por si no existe
postmaster &
createdb     #al igual que initdb, se ejecuta la primera vez
```

Lo anterior inicializa una copia local de PostgreSQL en la ruta `~/dirBD/` (está en la raíz de

una cuenta de usuario, ¡y debe existir!).

La segunda vez que se inicie localmente el demonio:

```
export PGDATA=~/dirBD/
postmaster &
```

Luego, para ejecutar un cliente:

```
psql
```

Para eliminar el demonio de PostgreSQL, luego de acabar una sesión:

```
killall postmaster
```

Para crear índices:

```
CREATE [ UNIQUE ] INDEX nombre ON tabla(atribs) [ USING método ];
método = btree | hash
```

Y para borrarlos:

```
DROP INDEX nombre;
```

También puede ser muy útil:

```
EXPLAIN ANALYZE consulta;
```

Para ver las opciones básicas de PostgreSQL escriba:

```
\h
```

Para la ejecución de *scripts*, necesarios en la realización de experimentos, puede usar:

```
\i archivo.sql
```

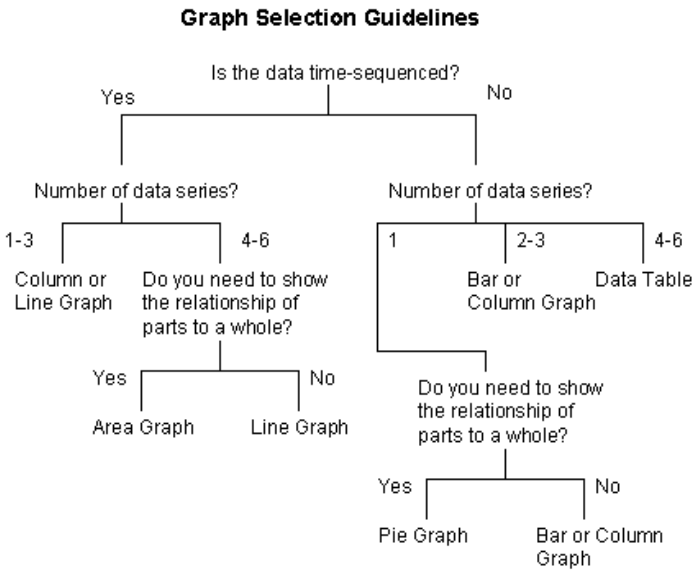
El archivo puede contener otras inclusiones de scrips con `\i`.

Para más información, visite http://www.postgresql.org/. Allí puede descargar PostgreSQL en su equipo, pues es software libre.

Generación de los gráficos

Es necesario generar gráficos que presenten adecuadamente los resultados. A continuación se

presentan algunos criterios para generar gráficos adecuados:



¿Qué se desea mostrar? Para presentar una comparación de las eficiencias, es necesario un gráfico de líneas. Ese gráfico debiera ir en el informe. Sin embargo, se pueden mostrar otros resultados con otros tipos de gráficos.

Para comparar la eficiencia de las consultas, se puede usar un mismo gráfico para todos los resultados de la misma respuesta (cada estrategia es una curva distinta). Asegúrese de hacer las líneas distinguibles. A continuación se presentan gráficos tipo:

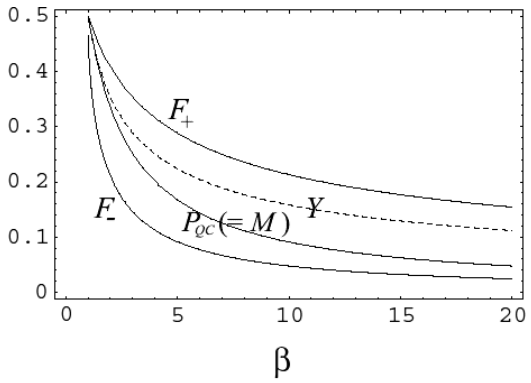
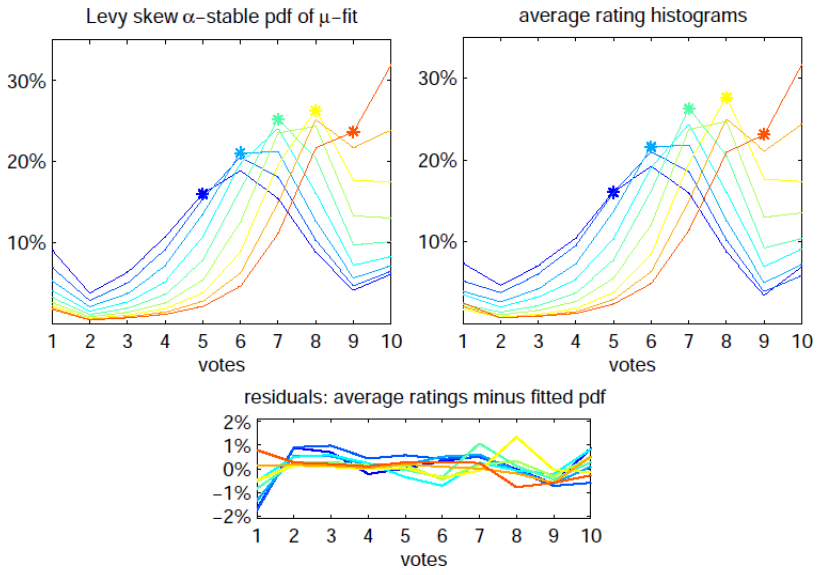


FIG. 1: Behavior of the various bounds $Y^{(1)}$, $M^{(1)}$, $P^{(1)}_{QC}$, F_+ and F_- versus the eigenvalue β in the discrimination of a thermal state $\sigma(\beta)$ from a vacuum state. Notice that $M^{(1)} = P^{(1)}_{QC}$ in this example.



También, para demostrar lo errático que son los resultados, y mostrar la densidad de los datos, se puede usar un gráfico de puntos dispersos, como el que sigue:

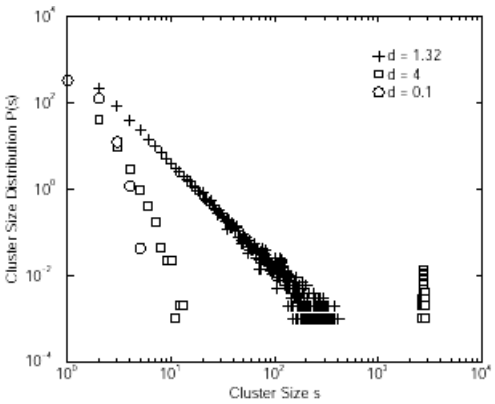


Fig. 4 Cluster size distribution. Far from the critical threshold ($d = 0.1$ and $d = 4$). $P(s)$ is well peaked. At $d_c = 1.32$, $P(s) \propto s^{-\alpha}$ with $\alpha = 2.45 \pm 0.05$. Here $N = 3200$. (After ref. [6]).

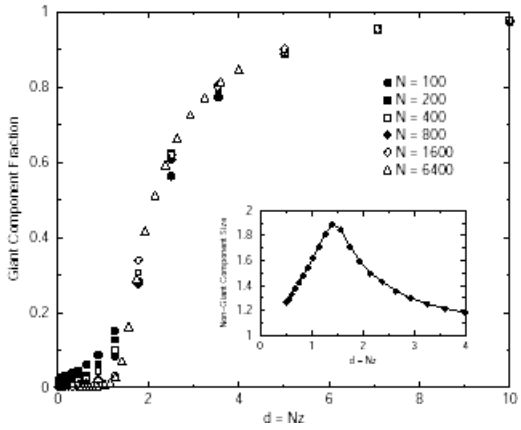


Fig. 5 Main panel: the fraction of nodes in the giant component for different network sizes as a function of d. Inset: the non-giant component average size as a function of d for $N = 6400$. (After ref. [6]).