CC4102 - Tarea 1

Prof. Gonzalo Navarro

Entrega: 5 de Noviembre de 2021

La tarea consiste en implementar y comparar algunas variantes de los árboles binarios de búsqueda (ABBs) vistas en clases, así como algunas clásicas, de manera de concluir sobre sus desempeños en distintas distribuciones de los datos.

1. Implementación

Deben incluirse las siguientes implementaciones:

1. El ABB clásico.
2. Una variante de ABB que garantice peor caso, ya sea AVLs o árboles rojo-negros. *Esta es la única variante que puede descargarse de internet sin implementarla usted mismo.*
3. Árboles B en memoria, permitiendo recibir el valor de B como parámetro al momento de la creación. Considere tres variantes, $B = 16$, $B = 256$, y $B = 4096$.
4. Árboles splay, los cuales se pueden consultar en la sección 3.7 del apunte.

Sobre esas estructuras, debe implementar la inserción y la búsqueda. No es necesario incluir el borrado.

2. Experimentación

Genere secuencias de n operaciones, mezclando inserciones, búsquedas exitosas y búsquedas infructuosas. Para ello, siempre se parte del árbol vacío con una inserción y luego, para cada operación, se escoge insertar con probabilidad p_i, buscar un elemento ya insertado con probabilidad p_{be}, y buscar uno no insertado con probabilidad p_{bi}, de modo que $p_i + p_{be} + p_{bi} = 1$. Debe generar primero la secuencia precisa de operaciones a realizar en un arreglo y luego aplicar la misma secuencia a cada una de las variantes de ABBs. Usaremos enteros positivos de 32 bits como los elementos a manejar en los árboles. Considere los siguientes esquemas de operaciones:

1. Aleatoria: El elemento a insertar o sobre el que hacer una búsqueda infructuosa se genera aleatoriamente, cuidando que no exista ya en el árbol. Para las búsquedas exitosas, elija un elemento al azar entre los ya insertados.
2. Creciente: Similar, pero los elementos que se van insertando tienden a ser crecientes. Para ello, al insertar, genere elementos al azar entre 0 y k y súmeles m, donde m es el número actual de elementos en el árbol. Considere los dos casos siguientes: $k = 0{,}1\,m$ y $k = 0{,}5\,m$.
3. Sesgada: Similar a aleatoria, pero las búsquedas exitosas se eligen de forma no uniforme entre los elementos ya insertados. Para ello, al momento de crear las consultas asígnele un peso $p(x)$ a cada elemento insertado con valor x, y al buscarlo, elíjalo con probabilidad $p(x)/P$, donde $P = \sum_x p(x)$. Use punto flotante para evitar overflows en este cálculo. Considere pesos de la forma $p(x) = x$, $p(x) = \sqrt{x}$, y $p(x) = \ln x$.

En total son 6 tipos de secuencias, y para cada tipo de secuencia se comparan las 6 variantes de ABBs (note que hay 3 de B-trees). Genere 100 secuencias de cada tipo, cada una de largo $n = 10^6$, $p_i = 1/2$, $p_{be} = 1/3$ y $p_{bi} = 1/6$, y promedie sobre ellas de modo de tener 36 curvas, una por cada tipo de secuencia y por cada implementación. Mida tiempos de CPU (user time), no system ni elapsed time. No incluya en la medición el tiempo de generar las secuencias, sólo el tiempo que toma el ciclo completo de recorrer las operaciones e irlas aplicando a un árbol inicialmente vacío.

Calcule la desviación estándar $\hat{\sigma}$ de su estimador (que es $\sigma/\sqrt{k}$, donde σ es la desviación estándar de la variable que estima y $k = 100$ el número de repeticiones que se promedian). Tomando la ley de los grandes números, el valor real de la variable que mide, si μ es el promedio obtenido, está en $[\mu - 2\hat{\sigma}, \mu + 2\hat{\sigma}]$ con 95 % de confianza. Si considera que este intervalo es demasiado ancho para sacar conclusiones, use un k mayor (genere más secuencias para promediar).

3. Presentación de Resultados

Realice un gráfico por cada comparación (uno por secuencia), cada uno con 6 curvas (una por implementación), donde el eje x sea el número de la operación y el eje y sea el tiempo promedio (subsamplee los valores de x cada 1000, por ejemplo, para no tener tantos puntos).

Con esta información básica extraiga resultados de más alto nivel, como

1. Tiempos promedio, medianas, varianzas y otros estadísticos de interés para $m = n$.
2. Ajustes de curvas según el modelo teórico de los tiempos esperados, en función de m.
3. Tendencias: qué estructuras se comportan mejor que otras en qué tipos de secuencias.

4. Discusión

Relacione los resultados de alto nivel obtenidos en los experimentos con las propiedades teóricas de las estructuras comparadas. Explique por qué son esperables, o no, los resultados obtenidos y concluya con recomendaciones prácticas de cuándo usar cada implementación (algunas pueden no ser nunca recomendables).

5. Informe

Su informe debe incluir los puntos anteriores, ser claro y estar bien redactado. Incluya en su entrega (no en su informe!) el código utilizado tanto para implementar las estructuras como para realizar los experimentos, de modo que éstos sean fácilmente replicables.

Se subieron al material del curso unas notas sobre cómo realizar y presentar experimentos de este tipo, para ampliar los consejos aquí dados.