Dynamic Order Statistics

Some of the slides are courtesy of Charles Leiserson and Carola Wenk

More Data structure ????
Isn’t it an Algorithm course ???

If you want to feel poetic
“Data Structures are Algorithms frozen in Time”

Anonymous

By now you are familiar with several data structures that supports the following operations on a dynamic set \( S \)

- **Insert \( (x, S) \):** inserts \( x \) into \( S \)
- **Delete \( (x, S) \):** deletes \( x \) from \( S \)
- **Find \( (x, S) \):** finds \( x \) in \( S \)
- **Succ\( (x, S) \):** find smallest element larger than \( x \) in \( S \)

Popular implementation uses any balanced search tree (not necessarily binary) or Skiplist. Each operation takes \( O(\log n) \) time.
Balanced search trees

*Balanced search tree:* A search-tree data structure for which a height of $O(\log n)$ is guaranteed when implementing a dynamic set of $n$ items.

**Examples:**
- AVL trees
- 2-3 trees
- 2-3-4 trees
- B-trees
- Red-black trees
- SkipList (only expected time bounds)
- Splay tress (Amortized time)

**Dynamic order statistics**

*Need a DS that supports the following operations in addition to Insert(x,S), Delete(x,S), Find(x), Succ(x,S):*

- **OS-SELECT(i, S):** returns the element with rank $i$ in the dynamic set $S$.
  - Smallest key has rank 1.
  - Largest has rank $n$.

- **OS-RANK(x, S):** returns the rank of $x \in S$ in the sorted order of $S$’s elements.
  - (many other problems could be solved in a similar techniques)

*First Try:* Each key stores its rank. So we only need to find the key (takes $O(\log n)$ in most data structures) and retrieve the rank.
So OS-Rank(x, S) takes $O(\log n)$

**Definition of the rank of a Key**

- **Assuming $S$ is a set of $n$ keys, all different from each other. Example $S$={10,20,30,40,50,60,70}**
- **The rank of a key $x$ in a set $S$ is 1+the number of keys in $S$ strictly smaller than $x$.

**Examples:**
- The rank of smallest key is 1
- The rank of largest key is $n$ (where $n=|S|$)
- The rank of the median (40 in the example) is $\left\lfloor n/2 \right\rfloor$
Dynamic order statistics-cont

- **Second Try:** (just for the protocol)
  - Store all keys in a sorted array.

- The index is the rank.
- So great for a static structure, less so for dynamic structure.

Dynamic order statistics-cont

**Third Idea:** (actually working) Use a balanced binary search tree for storing the set \( S \), but each node \( v \) has an extra field \( size[v] \) storing the number of keys in the subtree rooted at \( v \)

Notation for nodes:

Example of an OS-tree

Set of key = \{ A, B, .. Q \}

Key M  
Size=9

C  
D  
E  
F  
G  
H  
I  
J  
K  
L  
M  
N  
O  
P  
Q  
R  
S  
T  
U  
V  
W  
X  
Y  
Z

Agreement: \( size[leaf]=1 \). \( size[empty]=0 \)

Note that it is always true that \( size[x] = size[left[x]] + size[right[x]] + 1 \)

We will use it in the algorithm (wait for it)
Quick Reminder

Left subree
All keys \leq y

Right subree
All keys \geq y

How to answer OS-Rank(x,S)
Returns the number of keys in the tree which are strictly smaller than x.

• Assume we already performed find(x,S).
• Consider the search path from root to x.
  - It branches left and right.
  - All elements in yellow subtrees are < x.
  - All elements in blue trees are \geq x.

  e.g. in A_1, all keys < y < x.
  in B_1, all keys > u > x.

Need to sum numbers of all keys in yellow trees + # yellow nodes along the path
  - They are the left subtrees of every node where the search path branches right
  plus # these nodes.

OS_Rank(x,v,S).
If v==NULL return 0 \triangleright empty subtree
k \leftarrow size(left[v])
if key[v]==x then return k+1 else
if key[v] > x \triangleright Path branches LEFT
then return OS_Rank(x, left[v], S)
else \triangleright Path branches RIGHT
return k+1+OS_Rank(x, right[v], S)
How to answer OS-Select($i$, $S$)

Returns the $i$'th smallest key (e.g. OS-Select(0, $S$) returns the first. OS-Select($n-1$, $S$) return last)

$\text{OS-SELECT}(v, i) \\
\triangleright return$ $i$'th smallest element in the $\triangleright$ subtree rooted at $v$

$k \leftarrow size[left[v]] + 1$

if $i = k+1$ then return $v$ else

if $i \leq k$

then return OS-SELECT(left[v], $i$)
else return

OS-SELECT(right[v], $i-(k+1)$)

Example

OS-SELECT(root, 4)

Running time = $O(h) = O(\log n)$ for BSTs.

Data Structure Maintenance

Modifying operations: After an insertion or deletions took place, the size fields must be updated.

Assume a new node $x$ is inserted. (Deletion is analogous)

Two approaches (both legit): Top-down or Bottom-up:

1. The 'size' fields are updated on the way down to $x$.
2. They are updated once tracing up from the $x$ to the root, (bottom up by applying the formula

$Size[v] = size[left[v]] + size[right[v]] + 1$

If rotations are needed, the second approach must be used.
Example of insertion

\[
\text{\textbf{INSERT("K")}}
\]

Handling rebalancing

Don’t forget that \texttt{BST-INSERT} and \texttt{BST-DELETE} may also need to modify the binary search tree in order to maintain balance.

- Rotations: fix up subtree sizes in \(O(1)\) time.

Example:

\[
\begin{array}{c}
\text{C} & 11 \\
\text{E} & 16 \\
\end{array}
\quad
\begin{array}{c}
\text{C} & 16 \\
\text{E} & 8 \\
\end{array}
\]

\(\therefore\) \texttt{BST-INSERT} and \texttt{BST-DELETE} still run in \(O(\log n)\) time.

Data-structure augmentation

\textbf{Methodology:} (e.g., order-statistics trees)

1. Choose an underlying data structure (binary search trees, e.g. AVL or red-black trees).
2. Determine additional information to be stored in the data structure (subtree sizes).
3. Verify that this information can be maintained for modifying operations (\texttt{BST-INSERT}, \texttt{BST-DELETE} — don’t forget rotations).
4. Develop new dynamic-set operations that use the information (\texttt{OS-SELECT} and \texttt{OS-RANK}).

These steps are guidelines, not rigid rules.