Advanced Algorithmics (6EAP)

MTAT.03.238

Trees

Jaak Vilo

2015 Fall
Contents

• Tree as a data model
• Data structures
• Search trees
  – binary trees and balancing
  – (2,4)-trees, B-trees
  – k-d trees
• Heaps
• Union-find problem
• ...
Trees

• Some of the very basic essence of computer science and programming

• Chapter 5 – “The Tree Data Model” (pp 223-285) in
  • Foundations of Computer Science: C Edition
  • Alfred V. Aho, Jeffrey D. Ullman
  • W. H. Freeman (October 15, 1994)
Tree

• Acyclic graph
  – *root of a tree*
  – *children, parents, siblings, internal nodes, leaves*

• Binary tree – node has 0, 1, or 2 children
Data model

- Abstraction
- File directory system
- Hierarchical organisation structure
  - divide and conquer
- Hierarchical controlled vocabulary (simple ontology)
- Syntactic structure of a (sentence in a) language
- Syntax – e.g. paired parentheses
- ...

...
Example: XHTML and CSS

- This defines a single tree
Example: XHTML and CSS

• The nested tags define sub-trees

```html
<html>
  <head>
    <title>Hello World!</title>
  </head>
  <body>
    <h1>This is a <u>Heading</u></h1>
    <p>This is a paragraph with some <u>underlined</u> text.</p>
  </body>
</html>
```
Example: XHTML and CSS

- This may be rendered by a web browser
Terminology

root A
leaf {D,F,G,J,K,L,M}

degree B = 2
degree I = 3
Every node has 1 parent
except root has 0 parents

Depth = 3
Level C,E,I,M = 2

width = 6 /at level 3/

successor = children
siblings {F,G}, {J,K,L}

Path to K = A,H,I,K
• List is also a tree:
**Terminology**

**Descendants** (of B) = B,C,D,E,F,G

= a subtree with a root B

**Ancestors** of I = I,H,A

Every node is connected via a path to root
Terminology

Topologically equal to previous slide

Depends on application if order is important or not
Trie

From Wikipedia, the free encyclopedia

In computer science, a trie, or prefix tree, is an ordered tree data structure that is used to store an associative array where the keys are usually strings. Unlike a binary search tree, no node in the tree stores the key associated with that node; instead, its position in the tree shows what key it is associated with. All the descendants of any one node have a common prefix of the string associated with that node, and the root is associated with the empty string. Values are normally not associated with every node, only with leaves and some inner nodes that happen to correspond to keys of interest.

The term trie comes from "retrieval." Following the etymology, the inventor, Edward Fredkin, pronounces it [tri] ("tree") [1]. However, it is usually pronounced [træ] ("try"). [2]
Trie for $P=\{he, she, his, hers\}$
Binary trees
Binary tree

• This peach tree is not a binary tree...
Binary Trees
Definition: Any node can have 0, 1 or 2 children

- A \textit{full} node is a node where both the left and right sub-trees are non-empty trees

• Legend:
  - \textcolor{black}{full nodes}
  - \textcolor{blue}{neither}
  - \textcolor{green}{leaf nodes}
Basic node structure
Binary Trees

• An *empty node* or a *null sub-tree* is any location where a new leaf node could be inserted.
Binary Trees

Definition

• **A full binary tree** is where each node is:
  – a full node, or
  – a leaf node

• This has applications in
  – expression trees and Huffman encoding

Q: size of such a tree? $N$ leaves -> Total size?
Perfect Binary Trees: Definition

- A perfect binary tree of height $h$ is a binary tree where
  - All leaves have the same depth $h$
  - All other nodes are full
Perfect Binary Trees

Examples

- Perfect binary trees of height $h = 0, 1, 2, 3$ and $4$
Perfect Binary Trees

$2^h + 1 - 1$ Nodes

- Using the recursive definition, both sub-trees are perfect trees of height $h = k - 1$
- By assumption, each sub-tree has $2^{k+1} - 1$ nodes
- Therefore the total number of nodes is
  
  $$(2^{k+1} - 1) + 1 + (2^{k+1} - 1)$$
  
  $$= 2^{k+2} - 1$$
Complete Binary Trees: Definition

- A complete binary tree filled at each depth from left to right:
Complete Binary Trees: **Array Storage**

- Fill the array following a breadth-first traversal:

  \[
  \text{left}(i) = i \times 2 \\
  \text{right}(i) = i \times 2 + 1 \\
  \text{parent}(i) = \left\lfloor i/2 \right\rfloor
  \]
Complete Binary Trees: Array Storage

- To insert another node while maintaining the complete-binary-tree structure, we must insert into the next array location.
Complete Binary Trees

**Array Storage**

- To insert another node while maintaining the complete-binary-tree structure, we must insert into the next array location.
Traversal of a binary tree

Tree-Walk( x )

if x ≠ NULL then

   Tree-Walk( left(x) )

   Tree-Walk( right(x) )
Traversing a binary tree

Tree-Walk( x )

    if x ≠ NULL then

        // pre-order operations
        Tree-Walk( left(x) )

        // in-order operations
        Tree-Walk( right(x) )

        // post-order operations
Traversals of a binary tree - size

```c
int Tree-Size( x )

if x == NULL then return 0
return
    Tree-Size( left(x) ) +
    Tree-Size( right(x) ) + 1
```
Traversing a binary tree – parenthesisation

Tree-Walk( x )

if x ≠ NULL then
    print "(" , x.value ;
    Tree-Walk( left(x) ) ;
    Tree-Walk( right(x) ) ;
    print ")" ;
Parenthesisation

\(( 3 ( 9 ( 14 (17) (15) ) (10 (13) (23) ) ) ) \) ....
Binary Trees

Application: Expression Trees

• Expression trees

\[ 3(2a + c + a) + b/3 + (a – 2) \]
Binary Trees

Application: Expression Trees

– internal nodes store operators
– leaves store operands
– no node has just one sub tree
– the order is not relevant for addition and multiplication (commutative)
– the order is relevant for subtraction and division (non-commutative)
– to ignore order completely, represent subtraction and division as unary operators

\[
\frac{a}{b} = a \times b^{-1} \quad (a - b) = a + (-b)
\]
Binary Trees

Application: Expression Trees

• Performing appropriate tree traversals allows you to convert the representation

• Post-order results in reverse-Polish:

$$3 2 a * c a + + * b 3 / a 2 − + +$$
int Eval-Tree(x)
  int val1, val2;
  if x->op == ‘i’ return x->value; // x is a leaf, integer value
  else
    val1 = Eval-Tree(x->left);
    val2 = Eval-Tree(x->right);
    switch(n->op){
      case ‘+’: return val1 + val2;
      case ‘-’: return val1 - val2;
      case ‘*’: return val1 * val2;
      case ‘/’: return val1 / val2;
    }

General Trees: Design

- Children – in a linked list
Traversal of a general tree

Tree-Walk( x )

if x \neq \text{NULL} \text{ then }

\textbf{foreach } c \textbf{ in } \text{children}(x)

\hspace{1em} \text{Tree-Walk}( \ c \ )
Traversals of a general tree

Tree-Walk( x )

\[
\text{if } x \neq \text{NULL then}
\]

// pre-order operations

\[
\text{foreach } c \text{ in children}(x)
\]

Tree-Walk( c )

// post-order operations
Traversals of a general tree

Tree-Walk(x)

if x ≠ NULL then
  print "("; // pre-order operations
  foreach c in children(x)
    Tree-Walk(c)
  print ")"; // post-order operations
Depth-first Traversal

• We note that each node could be visited twice in such a scheme
  – the first time the node is approached, and
  – the last time it is approached.
Pre-order Depth-first Traversal

• Visiting each node first results in the sequence A, B, C, D, E, F, G, H, I, J, K, L, M
Post-order Depth-first Traversal

• Visiting the nodes with their last visit:
  
Parenthesised tree serialisation

• Passing such a visitor results in the output:

\[ A(B(C(D))(E(F)(G)))(H(I(J)(K)(L))(M)) \]
Breadth-First Traversal

• Breadth-first traversal would visit the nodes in the order:
Breadth-First Traversal

Breadth-First (x)

1 enqueue(Q, x)
2 while not empty(Q)
3 x = dequeue(Q)
4 print x->name // process node x
5 foreach c in next-child(x)
6 enqueue(Q, c)
Printing Directories

• Given the directory structure

```
/  
  usr/ 
    bin/ 
    local/ 
  var/ 
    adm/ 
    cron/ 
  log/ 
```
Exercise

• Print the following statistics for a given (e.g. current working) directory:
  – subdirectory size (# of all subdirectories and files)
  – depth (maximal height)
  – width at all levels of depth...
  – maximal depth
  – largest directory in nr of subdirs and files in that directory
  – ...

Binary Search Tree (BST)

- MIT [MIT](http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-046JFall-2005/VideoLectures/detail/embed09.htm)

- Binary tree where values of the keys have a special order:

  \[ \text{values(left subtree)} < \text{value(root)} \leq \text{values(right subtree)} \]
What is the **time complexity** of:

1. Quicksort
2. Merge Sort
3. Heap Sort
4. Radix sort
5. Theoretically best comparison based sort
6. Theoretically best sorting (any)

<table>
<thead>
<tr>
<th>Worst case</th>
<th>Average case</th>
<th>Could you implement it without consulting literature?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1.</td>
<td>1.</td>
</tr>
<tr>
<td>2.</td>
<td>2.</td>
<td>2.</td>
</tr>
<tr>
<td>3.</td>
<td>3.</td>
<td>3.</td>
</tr>
<tr>
<td>4.</td>
<td>4.</td>
<td>4.</td>
</tr>
<tr>
<td>5.</td>
<td>5.</td>
<td>5.</td>
</tr>
<tr>
<td>6.</td>
<td>6.</td>
<td>6.</td>
</tr>
</tbody>
</table>

**Draw a binary search tree (BST) after inserting: 2, 5, 6, 1, 3 into an initially empty tree**

What is the purpose of BST?

25 (of 52) did not draw any binary search tree …
6 did AVL-like?
Examples

• Here we see a complete binary search tree, and a binary search tree which is close to being complete -- balanced
Examples

• There are many different representations of the same ordered data:
Operations on dynamic sets

SEARCH(S, k)
A query that, given a set S and a key value k, returns a pointer x to an element in S such that key[x] = k, or NIL if no such element belongs to S.

INSERT(S, x)
A modifying operation that augments the set S with the element pointed to by x. We usually assume that any fields in element x needed by the set implementation have already been initialized.

DELETE(S, x)
A modifying operation that, given a pointer x to an element in the set S, removes x from S. (Note that this operation uses a pointer to an element x, not a key value.)

MINIMUM(S)
A query on a totally ordered set S that returns a pointer to the element of S with the smallest key.

MAXIMUM(S)
A query on a totally ordered set S that returns a pointer to the element of S with the largest key.

SUCCESSOR(S, x)
A query that, given an element x whose key is from a totally ordered set S, returns a pointer to the next larger element in S, or NIL if x is the maximum element.

PREDECESSOR(S, x)
A query that, given an element x whose key is from a totally ordered set S, returns a pointer to the next smaller element in S, or NIL if x is the minimum element.
Operations - search

TREE-SEARCH \((x, k)\)

1 if \(x = NIL\) or \(k = x.key\)
2 then return \(x\)
3 if \(k < x.key\)
4 then return TREE-SEARCH\((x.left, k)\)
5 else return TREE-SEARCH\((x.right, k)\)
Iterative search

ITERATIVE-TREE-SEARCH \((x, k)\)

1 while \(x \neq NIL\) and \(k \neq x.key\)
2   if \(k < x.key\)
3     then \(x = x.left\)
4   else \(x = x.right\)
5 return \(x\)

(Tail) Recursion “unrolling” – should be more efficient
Min and Max

Tree-Minimum (x)
1 while left[x] ≠ NIL
2 x = left[x]
3 return x

Tree-Maximum (x)
1 while right[x] ≠ NIL
2 x = right[x]
3 return x
Successor

Tree-Successor ( x )

1  if right[x] ≠ NIL
2    then return Tree-Minimum( right[x] )
3  y = parent[x]
4  while y ≠ NIL and x = right[y]
5    x = y; y = parent[y]
6  return y
Insert a node

• Find such a node where “next” position is missing...
Remove

• Suppose we wish to remove a node
• There are three situations: the node being removed
  – is a leaf node,
  – has exactly one child, or
  – is a full node (two children).
Remove

- If it is a leaf node, we can remove it:
Remove

• If the node has only one child, we can promote that child (with all the subtree underneath):
Remove

- If it is a full node, we copy the minimum element from the right sub-tree
- Recursively delete the value we copied
Example

• Consider the following tree
• We will twice remove the root
Example

- First, to remove 15, it is a full node
- We find the minimum element in the right sub-tree
Example

• We promote 42 to the root
• Proceed to remove 42 from the right sub-tree
Example

• This has one child, so we promote the entire sub-tree to replace 42
Example

• The root has been deleted, and the result is still a binary search tree
Example

- Next, let us remove 42
- Once again, it is a full node, so get the minimum element in the right sub-tree
Example

• We promote 45 to the root and proceed to delete 45 from the right sub-tree
Example

• The node 45 is a leaf node, so we may simply remove it
Example

• Thus, the final tree, having removed 15 and then 42 is
Reading

• CLRS: Binary Search Trees

• Visualisations:
Complexity...

• (Almost) all operations depend on the depth of the tree (or node affected)

• Binary search tree can get unbalanced, depth $O(n)$

• How to ensure this does not happen?
Balanced Binary Search Trees

- MIT
- AVL trees
- 2-3 trees
- 2-3-4 trees
- B-trees
- Red-black trees
Balance

- If elements are added in random, tree is “automatically balanced” on average
- Otherwise: we must re-balance it ourselves...
There are 4 cases in all, choosing which one is made by seeing the direction of the first 2 nodes from the unbalanced node to the newly inserted node and matching them to the top most row.

Root is the initial parent before a rotation and Pivot is the child to take the root's place.
AVL-trees

• Adelson-Velskii and Landis
• In an AVL tree, the heights of the two child subtrees of any node **differ by at most one**;

• In an AVL tree, the **heights** of the two child subtrees of any node **differ by at most one**;
• Difference: -1, 0, 1
• Re-balance using rotations when getting out of balance...
• O( lg n ) normal operations
• Up to O( lg n ) re-balancing operations of O(1)
• an AVL tree's height is limited to 1.44 lg n
Height of an AVL Tree

• If $n = 88$, the worst- and best-case scenarios differ in height by only 2:
Height of an AVL Tree

• If \( n = 10^6 \), the bounds on \( h \) are:
  
  – The minimum height: \( \log_2(10^6) - 1 \approx 19 \)
  
  – the maximum height: \( \log_2(10^6 / 1.8944) < 28 \)

An AVL tree's height is strictly less than:

\[
\log_\varphi(\sqrt{5}(n+2)) - 2 = \frac{\log_2(\sqrt{5}(n+2))}{\log_2(\varphi)} - 2 = \log_\varphi(2) \cdot \log_2(\sqrt{5}(n+2)) - 2 \approx 1.44 \log_2(n+2) - 0.328
\]

where \( \varphi \) is the golden ratio.
Red-Black trees
Red-Black Trees

1. A node is either red or black.
2. The root is black. (This rule is used in some definitions and not others. Since the root can always be changed from red to black but not necessarily vice-versa this rule has little effect on analysis.)
3. All leaves are black.
4. **Both children of every red node are black.**
5. Every simple path from a node to a descendant leaf contains the same number of black nodes.
Figure 1: The four cases for balancing a red-black tree.
Red-black trees

This data structure requires an extra one-bit color field in each node.

Red-black properties:
1. Every node is either red or black.
2. The root and leaves (NIL’s) are black.
3. If a node is red, then its parent is black.
4. All simple paths from any node $x$ to a descendant leaf have the same number of black nodes = black-height($x$).
Example of a red-black tree

4. All simple paths from any node $x$ to a descendant leaf have the same number of black nodes $= \text{black-height}(x)$. 
Height of a red-black tree

**Theorem.** A red-black tree with $n$ keys has height

\[ h \leq 2 \lg(n + 1). \]

**Proof.** (The book uses induction. Read carefully.)

**Intuition:**
- Merge red nodes into their black parents.
Height of a red-black tree

**Theorem.** A red-black tree with n keys has height $h \leq 2 \lg(n + 1)$.

**Proof.** (The book uses induction. Read carefully.)

**Intuition:**
- Merge red nodes into their black parents.
**Theorem.** A red-black tree with $n$ keys has height

$$h \leq 2 \log(n + 1).$$

**Proof.** (The book uses induction. Read carefully.)

**Intuition:**
- Merge red nodes into their black parents.
- This process produces a tree in which each node has 2, 3, or 4 children.
- The 2-3-4 tree has uniform depth $h'$ of leaves.
Proof (continued)

- We have $h' \geq h/2$, since at most half the leaves on any path are red.

- The number of leaves in each tree is $n + 1$
  \[ \Rightarrow n + 1 \geq 2^{h'} \]
  \[ \Rightarrow \lg(n + 1) \geq h' \geq h/2 \]
  \[ \Rightarrow h \leq 2 \lg(n + 1). \]
Query operations

**Corollary.** The queries SEARCH, MIN, MAX, SUCCESSOR, and PREDECESSOR all run in $O(\lg n)$ time on a red-black tree with $n$ nodes.
Modifying operations

The operations \textsc{Insert} and \textsc{Delete} cause modifications to the red-black tree:
\begin{itemize}
  \item the operation itself,
  \item color changes,
  \item restructuring the links of the tree via \textit{``rotations''}.
\end{itemize}
Rotations maintain the inorder ordering of keys:
• $a \in \alpha$, $b \in \beta$, $c \in \gamma \Rightarrow a \leq A \leq b \leq B \leq c$.

A rotation can be performed in $O(1)$ time.
**Insertion into a red-black tree**

**Idea:** Insert \( x \) in tree. Color \( x \) red. Only red-black property 3 might be violated. Move the violation up the tree by recoloring until it can be fixed with rotations and recoloring.

**Example:**

```
    7
   / \   \\
  3   10  18
   \      /   \\
    \    11   22
        /  \\
       8  26
```
Insertion into a red-black tree

**Idea:** Insert \( x \) in tree. Color \( x \) red. Only red-black property 3 might be violated. Move the violation up the tree by recoloring until it can be fixed with rotations and recoloring.

**Example:**
- Insert \( x = 15 \).
- Recolor, moving the violation up the tree.
Insertion into a red-black tree

**IDEA:** Insert $x$ in tree. Color $x$ red. Only red-black property 3 might be violated. Move the violation up the tree by recoloring until it can be fixed with rotations and recoloring.

**Example:**
- Insert $x = 15$.
- Recolor, moving the violation up the tree.
- **RIGHT-ROTATE(18).**
Insertion into a red-black tree

**Idea:** Insert $x$ in tree. Color $x$ red. Only red-black property 3 might be violated. Move the violation up the tree by recoloring until it can be fixed with rotations and recoloring.

**Example:**
- Insert $x = 15$.
- Recolor, moving the violation up the tree.
- **Right-Rotate**(18).
- **Left-Rotate**(7) and recolor.
**Insertion into a red-black tree**

**Idea:** Insert $x$ in tree. Color $x$ red. Only red-black property 3 might be violated. Move the violation up the tree by recoloring until it can be fixed with rotations and recoloring.

**Example:**
- Insert $x = 15$.
- Recolor, moving the violation up the tree.
- **Right-Rotate(18).**
- **Left-Rotate(7)** and recolor.
Pseudocode

RB-INSERT($T, x$)

TREE-INSERT($T, x$)

$color[x] \leftarrow$ RED \Comment{only RB property 3 can be violated}

while $x \neq root[T]$ and $color[p[x]] = RED$

do if $p[x] = left[p[p[x]]$

then $y \leftarrow right[p[p[x]]$ \Comment{$y =$ aunt/uncle of $x$}

if $color[y] = RED$

then \langle Case 1 \rangle

else if $x = right[p[x]]$

then \langle Case 2 \rangle \Comment{Case 2 falls into Case 3}

\langle Case 3 \rangle

else \langle“then” clause with “left” and “right” swapped\rangle

$color[root[T]] \leftarrow$ BLACK
Graphical notation

Let \( \triangle \) denote a subtree with a black root.

All \( \triangle \)'s have the same black-height.
Case 1

(Or, children of $A$ are swapped.)

Push $C$’s black onto $A$ and $D$, and recurse, since $C$’s parent may be red.
Case 2

\textbf{LEFT-ROTATE}(A)

Transform to Case 3.
Case 3

**Right-Rotate(C)**

Done! No more violations of RB property 3 are possible.
Analysis

- Go up the tree performing Case 1, which only recolors nodes.
- If Case 2 or Case 3 occurs, perform 1 or 2 rotations, and terminate.

**Running time:** $O(lg \ n)$ with $O(1)$ rotations.

**RB-DELETE** — same asymptotic running time and number of rotations as **RB-INSERT** (see textbook).
Other ideas

• Balancing can be an independent process
  – at night?

• Many search&insert&delete processes, and few rebalancing processes

• Local locking. Must ensure no deadlocks occur!
Properties of Red-Black trees

• No overhead for searching – efficient

• 100-200 lines of code, many symmetric cases

• Left-Leaning Red-Black trees (LLRB)
  – Robert Sedgewick:
Insert red node:

Passing a red link up in a LLRB tree
Splay trees (Sleator, Tarjan 1985)

- Self-adjusting BST
- Recently accessed elements are brought to the top of the tree
- **Repeated accesses will be executed faster!**
- No extra bookkeeping

- Average $\log(n)$ but worst case $O(n)$
2-3, 2-3-4, **B-trees**

- Binary trees are useful for memory-based data structures

- Large databases and disk based systems would benefit of fewer reads of larger block sizes

- Organise data in a search tree that minimizes disk accesses
Red-black representation of a 2-3-4 tree
B-tree (m-way)

\[ h = O(\log_m n) \]

In practice: 3-5 accesses to disk ..
B-tree properties

A B-tree of order $m$ (the maximum number of children for each node) is a tree which satisfies the following properties:

- Every node has at most $m$ children.
- Every node (except root and leaves) has at least $m/2$ children.
- The root has at least two children if it is not a leaf node.
- All leaves appear in the same level, and carry information.
- A non-leaf node with $k$ children contains $k-1$ keys.
Half-full property ensures that ...

- two half-full nodes can be joined to make a legal node, and one full node can be split into two legal nodes (if there is room to push one element up into the parent).
Example: Two-level Insertion

- Inserting 29
- Leaf node is full, so we split it into two
Example: Two-level Insertion

- Parent node is full, so we must split it
Example: Two-level Insertion

• The root node must be updated
Example: Root Insertion

• Insert 67
• Leaf is full, so split it into two
Example: Root Insertion

- Parent is full, so split it into two
Example: Root Insertion

- Root is full, so split it into two
Example: Root Insertion

- Create a new root node
• The creators of the B-tree structure, [Rudolf Bayer](https://en.wikipedia.org/wiki/Rudolf_Bayer) and Ed McCreight, have not explained what, if anything, the $B$ stands for. [Douglas Comer](https://en.wikipedia.org/wiki/Douglas_Comer) suggests a number of possibilities:

• "Balanced," "Broad," or "Bushy" might apply [since all leaves are at the same level]. Others suggest that the "B" stands for Boeing [since the authors worked at [Boeing](https://www.boeing.com) Scientific Research Labs in 1972]. Because of his contributions, however, it seems appropriate to think of B-trees as "Bayer"-trees.[1]
Variants of B-trees

- Keys and data in leaves or internal nodes
- Order statistics
- ...

...
Analogy between R-B and B-trees
**k-d tree**

- Multi-dimensional data
  - 2-dim \((x,y)\)
  - 3D \((x,y,z)\)
  - d-dim \((x_1,...,x_d)\)
- Does a point belong to a set?
- What is the closest point? (other data structures)
- ...
2-D tree (x,y coordinates)
kd tree
\( kd \)-Trees

- Suppose we wish to partition the following points in a 2-dimensional \( kd \)-tree:

\[(0.03, 0.90), (0.37, 0.04), (0.56, 0.78), (0.01, 0.48), (0.41, 0.89), (0.95, 0.07), (0.97, 0.09), (0.54, 0.65), (0.04, 0.61), (0.73, 0.69), (0.46, 0.58), (0.08, 0.89), (0.04, 0.41), (0.94, 0.02), (0.33, 0.07), (0.55, 0.54), (0.06, 0.05), (0.04, 0.06), (0.74, 0.97), (0.29, 0.15), (0.05, 0.88), (0.23, 0.23), (0.55, 0.02), (0.02, 0.97), (0.05, 0.07), (0.06, 0.28), (0.09, 0.55), (0.02, 0.91), (0.05, 0.97), (0.68, 0.42), (0.97, 0.18)\]
kd-Trees

- The first step is to order the points based on the 1st coordinate and find the median:

(0.01, 0.48), (0.02, 0.91), (0.02, 0.97), (0.03, 0.90),
(0.04, 0.06), (0.04, 0.41), (0.04, 0.61), (0.05, 0.07),
(0.05, 0.88), (0.05, 0.97), (0.06, 0.05), (0.06, 0.28),
(0.08, 0.89), (0.09, 0.55), (0.23, 0.23), (0.29, 0.15),
(0.33, 0.07), (0.37, 0.04), (0.41, 0.89), (0.46, 0.58),
(0.54, 0.65), (0.55, 0.02), (0.55, 0.54), (0.56, 0.78),
(0.68, 0.42), (0.73, 0.69), (0.74, 0.97), (0.94, 0.02),
(0.95, 0.07), (0.97, 0.09), (0.97, 0.18)
$kd$-Trees

- The median point, $(0.29, 0.15)$, forms the root of our $kd$-tree
kd-Trees

• This partitions the remaining points into two sets:

\{(0.01, 0.48), (0.02, 0.91), (0.02, 0.97), (0.03, 0.90),
(0.04, 0.06), (0.04, 0.41), (0.04, 0.61), (0.05, 0.07),
(0.05, 0.88), (0.05, 0.97), (0.06, 0.05), (0.06, 0.28),
(0.08, 0.89), (0.09, 0.55), (0.23, 0.23)\}

\{(0.33, 0.07), (0.37, 0.04), (0.41, 0.89), (0.46, 0.58),
(0.54, 0.65), (0.55, 0.02), (0.55, 0.54), (0.56, 0.78),
(0.68, 0.42), (0.73, 0.69), (0.74, 0.97), (0.94, 0.02),
(0.95, 0.07), (0.97, 0.09), (0.97, 0.18)\}
• Starting with the first partition, we order these according to the 2nd coordinate:

(0.06, 0.05), (0.04, 0.06), (0.05, 0.07), (0.23, 0.23),
(0.06, 0.28), (0.04, 0.41), (0.01, 0.48), (0.09, 0.55),
(0.04, 0.61), (0.03, 0.90), (0.02, 0.91), (0.02, 0.97),
(0.05, 0.88), (0.08, 0.89), (0.05, 0.97)
kd-Trees

• This point creates the left child of the root
kd-Trees

- Starting with the second partition, we also order these according to the $2^{nd}$ coordinate:

  $(0.55, 0.02), (0.94, 0.02), (0.37, 0.04), (0.33, 0.07),
  (0.95, 0.07), (0.97, 0.09), (0.97, 0.18), (0.68, 0.42),
  (0.55, 0.54), (0.46, 0.58), (0.54, 0.65), (0.73, 0.69),
  (0.56, 0.78), (0.41, 0.89), (0.74, 0.97)$
$kd$-Trees

- This point creates the right child of the root
kd-Trees

- Next, ordering the partitioned elements by the 1\textsuperscript{st} coordinate, we choose the medians to find the children of the left child (0.09, 0.55):

  (0.01, 0.48), (0.04, 0.06), (0.04, 0.41), \textbf{(0.05, 0.07)},
  (0.06, 0.05), (0.06, 0.28), (0.23, 0.23),
  (0.02, 0.91), (0.02, 0.97), (0.03, 0.90), \textbf{(0.04, 0.61)},
  (0.05, 0.88), (0.05, 0.97), (0.08, 0.89)
kd-Trees

- Doing the same with the two right partitions, we get the children of the right child of the root:

  (0.33, 0.07), (0.37, 0.04), (0.55, 0.02), (0.94, 0.02),
  (0.95, 0.07), (0.97, 0.09), (0.97, 0.18)

  (0.41, 0.89), (0.46, 0.58), (0.54, 0.65), (0.55, 0.54),
  (0.56, 0.78), (0.73, 0.69), (0.74, 0.97)
kd-Trees

• At the next level, we order the points again based on the 2\textsuperscript{nd} coordinate and choose the medians:

  (0.04, 0.06), (0.04, 0.41), (0.01, 0.48)
  (0.06, 0.05), (0.23, 0.23), (0.06, 0.28)
  (0.03, 0.90), (0.02, 0.91), (0.02, 0.97)
  (0.05, 0.88), (0.08, 0.89), (0.05, 0.97)
  (0.55, 0.02), (0.37, 0.04), (0.33, 0.07)
  (0.95, 0.07), (0.97, 0.09), (0.97, 0.18)
  (0.46, 0.58), (0.54, 0.65), (0.41, 0.89)
  (0.73, 0.69), (0.56, 0.78), (0.74, 0.97)
$kd$-Trees

- The result is a 2-dimensional $kd$-tree of the given 31 points
kd-Trees

- Finally, the last point, a leaf node, falls within the given box
$kd$-Trees

- A useful application of a $kd$-tree provides an efficient data structure for counting the number of points which fall within a given $k$-dimensional rectangle
kd-Trees

• This is used in image processing: locating objects within a scene, ray tracing, etc.
• Find the points which lie in the quadrant 
  \([0.5, 1] \times [0, 0.5]\)
kd-Trees

• The traversal rules we will follow are:
  – we always match the coordinate corresponding to the level we are current at
  – if that coordinate is less than the corresponding interval of the box, we only need to visit the right sub-tree
  – if that coordinate is greater than the corresponding interval, we need only visit the left sub-tree
  – otherwise, we check if the root is in the box and we visit both sub-trees
**kd-Trees**

- Starting with the left sub-tree:
  \[0.94 \in [0.5, 1]\]

- We note that
  \[(0.94, 0.02) \in [0.5, 1] \times [0, 0.5]\]

  and we visit both sub-trees.
Nearest neighbour search
• **kd-trees** are not suitable for efficiently finding the nearest neighbour in high dimensional spaces.

• As a general rule, if the dimensionality is \( D \), then number of points in the data, \( N \), should be \( N >> 2^D \).

• Otherwise, when **kd-trees** are used with high-dimensional data, most of the points in the tree will be evaluated and the efficiency is no better than exhaustive search.

• The problem of finding NN in high-dimensional data is thought to be **NP-hard**,[2] and approximate nearest-neighbour methods are used instead.
High dimensionality

• Data often comes in high-dimensional form

• Curse of dimensionality
  – K-d tree nodes become empty after a few levels already...

• Everything is “far” from everything else
  – Difference along even one dimension makes them far from each other
K-dimensional tree

- Partitions space into multidimensional “buckets”
- If done in classical way ends up with a lot of empty buckets to search through (in fact exponentially many)
- Needs to keep track of number of points per bucket
Fast approximate max-correlation queries

Dmytro Fishman
K-dimensional tree
K-dimensional tree
K-dimensional tree
K-dimensional tree
K-dimensional tree
Random projection tree
Random projection tree

A(2, 4)

B(6, 2)
Random projection tree

A(2, 4)

?x + ? = y

B(6, 2)
Random projection tree

\[ m = \frac{\text{change in } y}{\text{change in } x} \]

A(2, 4)
B(6, 2)
Random projection tree

\[ m = -\frac{1}{2} \]
Random projection tree

\[ y - y_0 = m(x - x_0) \]
Random projection tree

A(2, 4)

C(?, ?)

B(6, 2)
Random projection tree

A(2, 4)  C(4, 3)  B(6, 2)
Random projection tree

\[ (-1/2)x + 5 = y \]

A(2, 4)
C(4, 3)
B(6, 2)
Random projection tree

\[ m = -\frac{1}{m_{\text{perpendicular}}} \]

- A(2, 4)
- B(6, 2)
Random projection tree
Random projection tree

2x - 5 = y
Random projection tree

\[2x - 5 = y\]

\[2 \times 5 - 5 - 3 = ?\]
Random projection tree

\[ \frac{1}{2}x + 1 = y \]

\[ 2 \times 5 - 5 - 3 = 2 > 0 \]
Random projection tree

\[1/2 \times x + 1 = y\]

\[Y (4,4)\]

\[2 \times 4 - 5 - 4 = ?\]
Random projection tree

\[
\frac{1}{2} x + 1 = y
\]

\[
Y (4,4)
\]

\[
2 \times 4 - 5 - 4 = -1 < 0
\]
Random projection tree

\[
\frac{1}{2}x + 1 = y
\]
Random projection tree
Random projection tree
Random projection tree
Random projection tree
Random projection tree
## Methods average complexity

<table>
<thead>
<tr>
<th></th>
<th>Building Time</th>
<th>Search Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>$O(n^2)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>K-d Tree</td>
<td>$O(n \log(n^2))$</td>
<td>$O(n \log(n) + nd)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>RP Tree</td>
<td>$O(nd \log(n))$</td>
<td>$O(nd \log(n))$</td>
<td>$O(n)$</td>
</tr>
</tbody>
</table>
Random projection tree

- Places **hyper planes** between two arbitrary points each time dividing hyperspace **into two equal parts**.
- Also keeps track of number of points in the node.
See also (Wikipedia)

- implicit *kd-tree*
- min/max *kd-tree*
- Quadtree
- Octree
- Bounding Interval Hierarchy
- Nearest neighbor search
- Klee's measure problem
- *kd-trie*
Quadtree – divide area recursively to quadrants
Quadtree

- 2-dimensional
- 4 quadrants
- Either point or area based
http://nns.tume-maailm.pri.ee/
Octree – 3D, 8 children
R-Tree

Overlapping

Minimal bounding boxes

B-Tree analog on k-D
R-Tree
Variants

- R, R+, R*

**Difference between R+ trees and R trees**

R+ trees are a compromise between R-trees; and kd-trees; they avoid overlapping of internal nodes by inserting an object into multiple leaves if necessary.

**R+ trees differ from R trees in that:**
- Nodes are not guaranteed to be at least half filled
- The entries of any internal node do not overlap
- An object ID may be stored in more than one leaf node

**Advantages**

Because nodes are not overlapped with each other, point query performance benefits since all spatial regions are covered by at most one node. A single path is followed and fewer nodes are visited than with the R-tree
**R* tree** topological split.[1]

The pages overlap very little since the R*-tree tries to minimize page overlap, and the reinsertions further optimized the tree. The split strategy also does not prefer slices, the resulting pages are much more useful for common map applications.
25.10

- Priority queue (Binary Heap), Heapsort
- Augmenting data structures
- Union-Find...

- Heaps (mergeable...)

Priority queue

• Insert Q, x

• Retrieve next x from Q s.t. x.value is largest

• Sorted list implementation:
  – O(n) to insert x into right place
  – O(1) access, O(1) delete
Binary heap

Complete – missing nodes only at the lowest level

Heap property – on any path parent has higher priority

Typically: min-heaps

Priority queue
insert (Q, x)
pop Q
Binary heap - Insert

Insert into a next allowed place

Make sure heap property is restored
Binary heap – Insert – “Bubble up”
Use Array based implementation

left = i*2;
right = i*2 + 1;
parent = i/2 ;
Insert

insert( int A[], int x, int *last) {
    (*last)++;
    A[*last] = x;
    bubbleUp( A, *last );
}

Bubble up

BubbleUp( int A[], int i ) {
    while ( ( i>1 ) && A[i] > A[i/2] ){
        swap( A, i, i/2 ) ;
        i=i/2;
    }
}
Delete (max)
• Remove top value (make free space)

• Remove last element

• Insert to top value location, then bubble down to the correct place
Binary heap – Delete – “Bubble down”
Cost

• Insert – \( O(\log n) \)
• Delete – \( O(\log n) \)
Heap-sort

- Heapify the array
- while not empty
  - pop_largest
  - copy to next free place
• Build heap
  – n times insert to heap = $O(n \log n)$

• “Sort”
  – n times repeat remove largest = $O(n \log n)$

• Total: $O(n \log n)$ method
Heapify... in linear time

- last \( \frac{n}{2} \) – ignore
- \( \frac{n}{4} \) - bubble down (at most by 1 level)
- \( \frac{n}{8} \) – bubble down (at most by 2 levels)
- ...

\[
\sum_{i=1}^{\log n} \frac{in}{2^{i+1}} \leq \frac{n}{2} \sum_{i=1}^{\infty} \frac{in}{2^i}
\]
\[
\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots = 1
\]
\[
\quad + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \ldots = \frac{1}{2}
\]
\[
\qquad + \frac{1}{8} + \frac{1}{16} + \ldots = \frac{1}{4}
\]
\[
\ldots
\]
\[
\text{Sum} = 2
\]
Lecture 11
Augmenting Data Structures
- Dynamic order statistics
- Methodology
- Interval trees

Prof. Charles E. Leiserson
Dynamic order statistics

\textbf{OS-Select}(i, S): returns the \textit{i}th smallest element in the dynamic set \textit{S}.

\textbf{OS-Rank}(x, S): returns the rank of \textit{x} \in \textit{S} in the sorted order of \textit{S}'s elements.

\textbf{Idea}: Use a red-black tree for the set \textit{S}, but keep subtree sizes in the nodes.

Notation for nodes: \begin{tabular}{c|c}
\textit{key} & \textit{size} \\
\hline
\end{tabular}
Example of an OS-tree

\[ size[x] = size[left[x]] + size[right[x]] + 1 \]
Selection

Implementation trick: Use a sentinel (dummy record) for NIL such that $\text{size}[\text{NIL}] = 0$.

OS-SELECT($x$, $i$) $\triangleright$ $i$th smallest element in the subtree rooted at $x$

\[ k \leftarrow \text{size}[\text{left}[x]] + 1 \quad \triangleright k = \text{rank}(x) \]

if $i = k$ then return $x$

if $i < k$

then return OS-SELECT($\text{left}[x]$, $i$)

else return OS-SELECT($\text{right}[x]$, $i - k$)

(OS-RANK is in the textbook.)
Example

OS-SELECT(root, 5)

Running time = $O(h) = O(\lg n)$ for red-black trees.
Data structure maintenance

Q. Why not keep the ranks themselves in the nodes instead of subtree sizes?

A. They are hard to maintain when the red-black tree is modified.

Modifying operations: Insert and Delete.
Strategy: Update subtree sizes when inserting or deleting.
Example of insertion

\textbf{INSERT(”K”)}
Handling rebalancing

Don’t forget that RB-INSERT and RB-DELETE may also need to modify the red-black tree in order to maintain balance.

- **Recolorings**: no effect on subtree sizes.
- **Rotations**: fix up subtree sizes in $O(1)$ time.

**Example:**

$\therefore$ RB-INSERT and RB-DELETE still run in $O(\lg n)$ time.
Data-structure augmentation

**Methodology:** *(e.g., order-statistics trees)*

1. Choose an underlying data structure *(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 *(RB-INSERT, RB-DELETE — don’t forget rotations)*.
4. Develop new dynamic-set operations that use the information *(OS-SELECT and OS-RANK)*.

These steps are guidelines, not rigid rules.
Interval trees

**Goal:** To maintain a dynamic set of intervals, such as time intervals.

\[ i = [7, 10] \]

\[ low[i] = 7 \quad 10 = high[i] \]

**Query:** For a given query interval \( i \), find an interval in the set that overlaps \( i \).
• In computer science, an interval tree, also called a segment tree or segtree, is an ordered tree data structure to hold intervals. Specifically, it allows one to efficiently find all intervals that overlap with any given interval or point. It is often used for windowing queries, for example, to find all roads on a computerized map inside a rectangular viewport, or to find all visible elements inside a three-dimensional scene.

• The trivial solution is to visit each interval and test whether it intersects the given point or interval, which requires $\Theta(n)$ time, where $n$ is the number of intervals in the collection. Since a query may return all intervals, for example if the query is a large interval intersecting all intervals in the collection, this is asymptotically optimal; however, we can do better by considering output-sensitive algorithms, where the runtime is expressed in terms of $m$, the number of intervals produced by the query.
Following the methodology

1. Choose an underlying data structure.
   - Red-black tree keyed on low (left) endpoint.

2. Determine additional information to be stored in the data structure.
   - Store in each node $x$ the largest value $m[x]$ in the subtree rooted at $x$, as well as the interval $int[x]$ corresponding to the key.
Example interval tree

\[ m[x] = \max \left\{ \text{high}[\text{int}[x]], \text{m[left}[x]], \text{m[right}[x]] \right\} \]
Modifying operations

3. Verify that this information can be maintained for modifying operations.
   • **INSERT**: Fix $m$’s on the way down.
   • **Rotations** — Fixup = $O(1)$ time per rotation:

```
   11,15   6,20
   ----    ----
     30    30
   6,20   11,15
   ----    ----
     30    30
   30     19
   14     19
```

**Total INSERT time** = $O(\log n)$; **DELETE** similar.
New operations

4. Develop new dynamic-set operations that use the information.

`INTERVAL-SEARCH(i)`

\[ x \leftarrow \text{root} \]

\[ \text{while } x \neq \text{NIL} \text{ and } (\text{low}[i] \succ \text{high}[\text{int}[x]]) \]
\[ \text{or } \text{low}[\text{int}[x]] \succ \text{high}[i]) \]

\[ \text{do } i \text{ and } \text{int}[x] \text{ don’t overlap} \]

\[ \text{if } \text{left}[x] \neq \text{NIL} \text{ and } \text{low}[i] \leq \text{m}[\text{left}[x]] \]
\[ \text{then } x \leftarrow \text{left}[x] \]
\[ \text{else } x \leftarrow \text{right}[x] \]

\[ \text{return } x \]
Cases for overlap:

Queries $i$

$\text{Int}(X)$

$max(\text{tree})$
Example 1: \textsc{Interval-Search}([14,16])

\[
x \leftarrow \text{root}
\]

\[ [14,16] \text{ and } [17,19] \text{ don't overlap} \]

\[ 14 \leq 18 \Rightarrow x \leftarrow \text{left}[x] \]
Example 1: \textsc{Interval-Search}([14,16])

\[ [14,16] \text{ and } [5,11] \text{ don’t overlap} \]

\[ 14 > 8 \Rightarrow x \leftarrow \text{right}[x] \]
Example 1: `INTERVAL-SEARCH([14,16])`

[14,16] and [15,18] overlap

return [15,18]
Example 2: \texttt{INTERVAL-SEARCH}([12,14])

\begin{itemize}
\item \texttt{x} \leftarrow \texttt{root} \\
\item \texttt{[12,14]} and \texttt{[17,19]} don't overlap \\
\item 12 \leq 18 \Rightarrow x \leftarrow \texttt{left}[x]
\end{itemize}
Example 2: \textsc{Interval-Search}([12,14])

$[12,14]$ and $[5,11]$ don't overlap

$12 > 8 \Rightarrow x \leftarrow \text{right}[x]$
Example 2: \textsc{Interval-Search}([12,14])

\begin{itemize}
  \item [12,14] and [15,18] don’t overlap
  \item 12 > 10 \Rightarrow x \leftarrow \text{right}[x]
\end{itemize}
Example 2: \textsc{Interval-Search}([12, 14])

\[
\begin{array}{c}
5, 11 \\
\underline{18}
\end{array}
\quad
\begin{array}{c}
17, 19 \\
\underline{23}
\end{array}
\quad
\begin{array}{c}
22, 23 \\
\underline{23}
\end{array}
\quad
\begin{array}{c}
4, 8 \\
\underline{8}
\end{array}
\quad
\begin{array}{c}
15, 18 \\
\underline{18}
\end{array}
\quad
\begin{array}{c}
7, 10 \\
\underline{10}
\end{array}
\quad
x
\]

\[x = \text{NIL} \Rightarrow \text{no interval that overlaps } [12, 14] \text{ exists}\]
Analysis

Time = $O(h) = O(\lg n)$, since INTERVAL-SEARCH does constant work at each level as it follows a simple path down the tree.

List all overlapping intervals:
- Search, list, delete, repeat.
- Insert them all again at the end.

Time = $O(k \lg n)$, where $k$ is the total number of overlapping intervals.

This is an output-sensitive bound.

Best algorithm to date: $O(k + \lg n)$. 

October 24, 2005
Correctness

**Theorem.** Let \( L \) be the set of intervals in the left subtree of node \( x \), and let \( R \) be the set of intervals in \( x \)’s right subtree.

- If the search goes right, then
  \[
  \{ i' \in L : i' \text{ overlaps } i \} = \emptyset.
  \]

- If the search goes left, then
  \[
  \{ i' \in L : i' \text{ overlaps } i \} = \emptyset \Rightarrow \{ i' \in R : i' \text{ overlaps } i \} = \emptyset.
  \]

*In other words, it’s always safe to take only 1 of the 2 children: we’ll either find something, or nothing was to be found.*
cases

Queries i

Int(X)

max( tree )
Go right:
Correctness proof

Proof. Suppose first that the search goes right.

- If $\text{left}[x] = \text{NIL}$, then we’re done, since $L = \emptyset$.
- Otherwise, the code dictates that we must have $\text{low}[i] > m[\text{left}[x]]$. The value $m[\text{left}[x]]$ corresponds to the high endpoint of some interval $j \in L$, and no other interval in $L$ can have a larger high endpoint than $\text{high}[j]$.

\[ \text{high}[j] = m[\text{left}[x]] \]

- Therefore, $\{i' \in L : i' \text{ overlaps } i\} = \emptyset$. 
Left:

\[ m[\text{left}[x]] \]
Proof (continued)

Suppose that the search goes left, and assume that 
{\(i' \in L : i' \) overlaps \(i\) } = \(\emptyset\).

• Then, the code dictates that \(\text{low}[i] \leq m[\text{left}[x]] = \text{high}[j]\) for some \(j \in L\).

• Since \(j \in L\), it does not overlap \(i\), and hence \(\text{high}[i] < \text{low}[j]\).

• But, the binary-search-tree property implies that for all \(i' \in R\), we have \(\text{low}[j] \leq \text{low}[i']\).

• But then \{\(i' \in R : i' \) overlaps \(i\) \} = \(\emptyset\).
Treap

- Tree + Heap = Treap

- BST + Heap

  of (key,priority) pair at the same time.
Combining info: Treap

• Heap and binary search tree properties together

• Treap
Combining info: insert (18,20)

- insert 18, priority 20
Delete max

- delete max priority...
Combining info

• delete max priority...
Combining info

• delete max priority...
Combining info

• delete max priority...
Treap

- Can be used to make a random-like tree: priorities can be assigned by random, unique values...
- In [computer science](http://en.wikipedia.org/wiki/Computer_science), a **treap** is a [binary search tree](http://en.wikipedia.org/wiki/Binary_search_tree) that orders the nodes by adding a random **priority** attribute to a node, as well as a key. The nodes are ordered so that the keys form a binary search tree and the priorities obey the max heap order property. The name treap is a **portmanteau** of **tree** and **heap**.
- A **portmanteau word** (pronounced /pɔːtˈmæn.təʊ/ (help·info)) is used broadly to mean a **blend** of two (or more) words,[1][2][3] and narrowly in **linguistics** fields to mean only a blend of two or more **function words**
Union-find

• Domain $X = \{ x_1, \ldots, x_n \}$
• $x_i$ belongs to a set $S_j$
• Non-intersecting sets.

• **Union** of sets: $S_i' = S_i \cup S_j$

• **Find:** Which set $S_i$ does an element $x_j$ belong to?
• Sets = \{ \{1\}, \{2\}, \ldots \{n\} \}
• Non-overlapping, each value belongs to a set

• Merge sets i, j (give new set id i, remove j)
  – Union

• Which set does x belong to?
  – Find
Set
Value

\[
\begin{array}{cccccccc}
1 & 2 & 4 & 4 & 5 & 6 & 7 & 6 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{array}
\]

\[4=\{3, 4\}\]
\[6=\{6, 8\}\]

**Union 4, 6**

Find = \(O(1)\)
Union = \(O(n)\)
Link elements until “0”

Set
Value

Find  =  O(n)
Union  =  O( 1 )

s1 = { 3, 4 }
s2 = { 6, 8 }

Union 4, 6
Set Value

0 0 0 3 0 0 0 6
1 2 3 4 5 6 7 8

Union 4, 6

Set Value

0 0 0 3 0 4 0 6
1 2 3 4 5 6 7 8

Find 6

Set Value

0 0 0 3 0 3 0 6
1 2 3 4 5 6 7 8
• At every find – “flatten the tree”
Union-Find
A data structure for maintaining a collection of disjoint sets

Course: Data Structures
Lecturer: Uri Zwick
March 2008
Union-Find

- **Make(x):** Create a set containing $x$
- **Union(x,y):** Unite the sets containing $x$ and $y$
- **Find(x):** Return a representative of the set containing $x$
Union Find

make \( O(1) \)
union \( O(\alpha(n)) \)
find \( O(\alpha(n)) \)

Amortized
Fun applications: Generating mazes

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>make(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>make(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>make(16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>find(6)=find(7) ?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>union(6,7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>find(7)=find(11) ?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>union(7,11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Choose edges in random order and remove them if they connect two different regions
Fun applications: Generating mazes

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

[Grid image of a 4x4 grid with numbers from 1 to 16]
Generating mazes – a larger example

Construction time -- $O(n^2 \alpha(n^2))$
Graphics) example

A module XML::Writer

viewer, for example Adobe SVG Viewer
More serious applications:

- Maintaining an equivalence relation
- Incremental connectivity in graphs
- Computing minimum spanning trees
- ...


Union Find

Represent each set as a rooted tree

Union by rank   Path compression

The parent of a vertex $x$ is denoted by $p[x]$

$\text{Find}(x)$ traces the path from $x$ to the root
Union by rank on its own gives $O(\log n)$ find time.

A tree of rank $r$ contains at least $2^r$ elements.

If $x$ is not a root, then $\text{rank}(x) < \text{rank}(p[x])$
Path Compression
Union Find - pseudocode

**Function** make-set($x$)

$p[x] \leftarrow x$
$rank[x] \leftarrow 0$

**Function** union($x, y$)

$link(find(x), find(y))$

**Function** find($x$)

\[
\text{if } p[x] \neq x \text{ then} \\
\quad p[x] \leftarrow \text{find}(p[x]) \\
\text{return } p[x]
\]

**Function** link($x, y$)

\[
\text{if } rank[x] > rank[y] \text{ then} \\
\quad p[y] \leftarrow x \\
\text{else} \\
\quad p[x] \leftarrow y \\
\quad \text{if } rank[x] = rank[y] \text{ then} \\
\quad \quad rank[y] \leftarrow rank[y] + 1
\]
# Union-Find

## Worst case

<table>
<thead>
<tr>
<th>make</th>
<th>link</th>
<th>find</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(1)$</td>
<td>$O(1)$</td>
<td>$O(\log n)$</td>
</tr>
</tbody>
</table>

## Amortized

<table>
<thead>
<tr>
<th>make</th>
<th>link</th>
<th>find</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O(1)$</td>
<td>$O(\alpha(n))$</td>
<td>$O(\alpha(n))$</td>
</tr>
</tbody>
</table>
Nesting / Repeated application

\[ f^{(i)}(n) = f(f(\ldots(f(n))\ldots)) \]
\[ \text{\textit{i times}} \]

\[ f^{(0)}(n) = n \]

\[ f^{(i)}(n) = f(f^{(i-1)}(n)) \text{, for } i > 0 \]

\[ f(n) = n + 1 \quad f^{(5)}(n) = n + 5 \]

\[ f(n) = 2n \quad f^{(7)}(n) = 2^7 n \]

\[ f(n) = 2^n \quad f^{(3)}(n) = 2^{2^n} \]

\[ f(n) = \log n \quad f^{(2)}(n) = \log \log n \]
Ackermann’s function

\[ A_k(n) = \begin{cases} 
  n + 1 & \text{if } k = 1, \\
  A_{k-1}^{(n+1)}(n) & \text{if } k > 1.
\end{cases} \]

\[ A_1(n) = n + 1 \]
\[ A_2(n) = 2n + 1 \]
\[ A_3(n) = 2^{n+1}(n + 1) - 1 \]
\[ A_4(n) = ? \]
Ackermann’s function (modified)

\[ A_k(n) = \begin{cases} 
  n + 1 & \text{if } k = 1, \\
  A_{k-1}^{(n+1)}(n) & \text{if } k > 1.
\end{cases} \]

\[ \bar{A}_k(n) = \begin{cases} 
  2n & \text{if } k = 2, \\
  \bar{A}_{k-1}^{(n)}(1) & \text{if } k > 2.
\end{cases} \]

\[
\bar{A}_2(n) = 2n \\
\bar{A}_3(n) = 2^n \\
\bar{A}_4(n) = tower(n) = 2^{2^{\cdots^{2}}} 
\]
Inverse functions

\[ F(n) \implies f(n) = \min\{k \geq 1 \mid F(k) \geq n\} \]

\begin{align*}
F(n) &= n + 1 & f(n) &= n - 1 \\
F(n) &= 2n & f(n) &= \left\lfloor \frac{n}{2} \right\rfloor \\
F(n) &= 2^n & f(n) &= \left\lfloor \log_2 n \right\rfloor \\
F(n) &= \text{tower}(n) & f(n) &= \log^* n
\end{align*}
Inverse Ackermann function

$$\alpha_r(n) = \min\{ k \geq 1 \mid A_k(r) \geq n \}$$

$$\alpha(n) = \alpha_1(n) = \min\{ k \geq 1 \mid A_k(1) \geq n \}$$

$$\alpha(n)$$ is the inverse of the function $$A_n(1)$$

$$A_n(1) = A_{n-1}(1) = A_{n-1}(A_{n-1}(1)) > A_{n-1}(n)$$

The first “column”

A “diagonal”
Amortized cost of **make**

Actual cost: $O(1)$

$\Delta \Phi$: 0

Amortized cost: $O(1)$
Amortized cost of **link**

Actual cost: $O(1)$

The potentials of $y$ and $z_1$, $\ldots, z_k$ can only decrease

The potentials of $x$ is increased by at most $\alpha(n)$

$$\Delta \Phi \leq \alpha(n)$$

Actual cost: $O(\alpha(n))$
Amortized cost of $\text{find}$

$y = p'[x]$

- $\text{rank}[x]$ is unchanged
- $\text{rank}[p[x]]$ is increased

- $\text{level}(x)$ is either unchanged or is increased

If $\text{level}(x)$ is unchanged, then $\text{index}(x)$ is either unchanged or is increased

If $\text{level}(x)$ is increased, then $\text{index}(x)$ is decreased by at most $\text{rank}[x] - 1$

$$\phi(x) = (\alpha(n) - \text{level}(x)) \cdot \text{rank}[x] - \text{index}(x)$$

is either unchanged or is decreased
Amortized cost of \textit{find}

Suppose that:

\[ 0 < i < j < \ell \]
\[ \text{level}(x_i) = \text{level}(x_j) \]

\[ A_{\text{level}(x_i)}^{(\text{index}(x_i)+1)}(\text{rank}[x_i]) \]
\[ = A_{\text{level}(x_i)}(A_{\text{level}(x_i)}^{(\text{index}(x_i))}(\text{rank}[x_i])) \]
\[ \leq A_{\text{level}(x_i)}(\text{rank}[p[x_i]]) \]
\[ \leq A_{\text{level}(x_j)}(\text{rank}[x_j]) \]
\[ \leq \text{rank}[p[x_j]] \]
\[ \leq \text{rank}[x_\ell] \]

\( \phi(t) \) is decreased!
Amortized cost of find

The only nodes that can retain their potential are: the first, the last and the last node of each level.

Actual cost:

\[ l + 1 \]

\[ \Delta \Phi \leq (\alpha(n) + 1) - (l + 1) \]

Amortized cost:

\[ \alpha(n) + 1 \]