CMSC 420: Disjoint-Set Data Structures

Justin Wyss-Gallifent

December 6, 2022

1 Introduction ............................................ 2
2 Implementation ........................................ 2
3 Three Basic Operations Version 1 ..................... 3
   3.1 Introduction ..................................... 3
   3.2 Creating a New Subset with a New Element ....... 3
   3.3 Finding Subset Representatives .................. 3
   3.4 Union of Subsets ................................ 4
   3.5 Time Complexity ................................ 5
3 Three Basic Operations Version 2 .................... 5
   4.1 Introduction ..................................... 5
   4.2 Finding Subset Representatives Revisited ...... 5
   4.3 Union of Subsets Revisited ....................... 6
   4.4 Time Complexity Revisited ...................... 7
5 Simple Graph Management ............................... 7
   5.1 Introduction ..................................... 7
   5.2 Representing Connected Components .............. 7
   5.3 Edge Addition ................................... 8
   5.4 Does Adding an Edge Form a Cycle? .............. 8
   5.5 Is There a Cycle in the Graph? ................... 8
6 Kruskal’s Algorithm .................................. 9
1 Introduction

Disjoint-set data structures, also called union-find data structures, are a class of data structure which stores elements according to disjoint sets they are in in a way which allows us to do things such as:

- Discover if two elements are in the same disjoint set.
- Merge two disjoint sets to form a new one.

Disjoint-set data structures are heavily used in Kruskal’s Algorithm for finding a minimal spanning tree, specifically for checking if there is a cycle in the graph.

2 Implementation

The most common way to implement a disjoint-set data structure is a forest, or a collection of trees. We’ll call this a disjoint set forest or just a forest for short. Each tree will correspond to a subset and each node within that tree to an element. The trees will be a bit non-traditional in that each node will only contain a pointer to its parent. For each subset, one node will be chosen as the root. This is a representative of the subset. The root nodes have no parent of course and it’s traditional to have their parent pointers point to themselves.

There is no real limit to the number of children a node could have since we aren’t storing child data.

Example 2.1. Consider the set of elements \{0, 1, 2, 3, 4, 5, 6, 7\} divided into three disjoint subsets \{0, 1, 6\}, \{2, 4, 7\}, and \{3, 5\}.

We’ll store this by creating three trees, one for each subset. Here they are. In these pictures I’ve not drawn the parent pointers (pointing to themselves) out of habit:

```
1
0 6 4

2
4 7 5

3

```

In such case we chose the representatives 1, 2, and 3 but this was arbitrary. In fact this forest would work, too:

```
0
6
2

4 5
7 3

1
```
In addition the structure of these trees can be stored in a simple list $A$ indexed with the number of elements whereby $A[i]$ is equal to the parent index. In such a case for a root $r$ we assign $A[r] = r$.

**Example 2.2.** The second picture in the above example would be stored easily as:

$$A = [0, 6, 4, 5, 4, 5, 0, 7]$$

This is because:

- $A[0] = 0$ since 0 is a root.
- And so on...


Before proceeding, we have:

**Definition 2.0.1.** For an element $x$ in the set, define the root of $x$, denoted $\text{root}(x)$, to be the root of the tree containing $x$.

In this sense the representative of a subset can be found by taking $\text{root}(x)$ for any $x$ in the subset.

3 Three Basic Operations Version 1

3.1 Introduction

There are three basic operations we can easily perform with our disjoint subset data structure.

3.2 Creating a New Subset with a New Element

If a new element is introduced we simply create a node (or list entry) which points to itself.

3.3 Finding Subset Representatives

To find the representative for a subset means to find the root. The following pseudocode will do this. This pseudocode is premised on the fact that the root’s parent is itself.

The pseudocode would look like:
Observe that finding subset representatives allows us to easily see if two elements are in the same subset. Given two elements \( x \) and \( y \) we can check if they’re in the same set by checking if \( \text{findrep}(x) == \text{findrep}(y) \).

**Note 3.3.1.** You might wonder why we didn’t just call this function \( \text{root} \), since we’re finding the root. The answer is that in a bit we’ll tweak it so that it adjusts the tree and we’d like to keep the \( \text{root}(x) \) so that it simply returns the root.

### 3.4 Union of Subsets

Given two subsets it’s easy to merge them. Suppose \( x \) and \( y \) are elements and we wish to merge the subsets which contain them. We check \( \text{root}(x) \) and \( \text{root}(y) \). If they’re the same then there’s nothing to do. If they’re not the same then we simply set the parent of \( \text{root}(x) \) (which was originially \( \text{root}(x) \)) to be \( \text{root}(y) \).

**Example 3.1.** For example:

```
function findrep(x)
    if x.parent == x
        return(x)
    else
        return(findrep(x.parent))
    end if
end function
```

The pseudocode would look like:

```
function union(x,y)
    if findrep(x) != findrep(y)
        findrep(x).parent = findrep(y)
    end if
end function
```
3.5 Time Complexity

The major issue with the above operations is that it’s possible for the trees representing the subsets to get very unbalanced. For example if we have \( n \) elements total we might end up with one subset with \( n \) elements for which the tree is a list of length \( n \), or we might end up with two subsets with \( n/2 \) elements for which the trees are lists of length \( n/2 \).

In such cases the second two operations run with time \( O(n) \) which is less than ideal.

4 Three Basic Operations Version 2

4.1 Introduction

We’d like to speed up our operations!

We might suggest some ideas like - keeping the trees balanced, but that takes time itself. Can we do something else?

The answer is yes and it’s anchored in the fact that our trees have two interesting properties. First, there is no child limit, and second, we have parent pointers. This allows us to modify our second two operations so that they keep the trees “short”.

4.2 Finding Subset Representatives Revisited

When we are finding a subset representative of the subset containing the element \( x \) we follow the graph from \( x \) to \( root(x) \). While we’re doing this we can actually easily modify all the nodes along the root so that their parents are \( root(x) \).

Here is the modified pseudocode. Note that it runs just as fast as it did before.

```pseudocode
function findrep(x)
    if x.parent == x
        return x
    else
        return (x.parent = findrep(x))
    end if
end function
```

Here’s an example to see what it does:

Example 4.1. This example demonstrates what our updated `findrep(5)` will do:
It turns out that if we do this every time we look for the representative we keep the tree short enough to reduce the time complexity significantly. We call this path compression.

4.3 Union of Subsets Revisited

Taking the union of subsets is what can lead to trees getting rather tall so perhaps there’s a way to carefully join the trees so that this doesn’t happen. There are, and we’ll look at one of them.

Given two elements $x$ and $y$ When we took the union earlier we simply set $\text{root}(x) = y$. However suppose the tree containing $x$ has height $h_x$ and the tree containing $y$ has height $h_y$. If we set $\text{root}(x) = y$ then the new tree has height $h_x + 1$ whereas if we set $\text{root}(y) = x$ then the new tree has height $h_y + 1$. We could of course record the heights of the trees and choose the shorter option but keeping track of tree heights takes time and care, especially given the fact that we’re repeatedly messing with them in this case.

Instead a reasonable proxy for tree height is the number of elements. A tree with more elements tends to be higher. Moreover keeping track of element counts is easy. When we create a tree with one element we store its size as 1 and when we merge two trees we add their sizes.

So what we will do is pick the tree with fewer elements and attach the root of that tree to the root of the tree with more elements. We call this the weighted union. It turns out that this small change has a massive impact.
Here is the modified pseudocode. Note that it runs just as fast as it did before.

```plaintext
function union(x, y)
    if findrep(x) != findrep(y)
        if x.size < size.y
            findrep(x).parent = findrep(y)
        else
            findrep(y).parent = findrep(x)
        end if
    end if
end function
```

Example 4.2. For example:

![Diagram of sets](attachment:image.png)

4.4 Time Complexity Revisited

These small changes punch above their weight. If we implement them then amortized analysis shows that any series of our three operations runs in $O(\alpha(n))$ amortized time, where $\alpha$ is the inverse of the Ackermann function which is "essentially constant". The Ackermann function is an increasing function which is less than 4 for all $n$ less than approximately $10^{665}$.

This means that our set operations essentially run in constant amortized time!

The proof of this amortized time complexity is not trivial. If you are interested you can find it in the classic CLRS (Cormen, Leiserson, Rivest, Stein) algorithms textbook.

5 Simple Graph Management

5.1 Introduction

Here are some tools for managing simple graphs. In what follows we’ll assume that we have some fixed number $n$ of vertices and the only thing that changes is the edges.

5.2 Representing Connected Components

Given a (not necessarily connected) graph with $n$ vertices suppose we partition the set of vertices into subsets according to which connected component. These subsets will be disjoint and their union will be the set of all vertices.
Example 5.1. Consider the following graph:

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 \\
\downarrow & \downarrow & \downarrow & \downarrow \\
\downarrow & \downarrow & \downarrow & \downarrow \\
4 & 5 & 6 & 7 \\
\end{array}
\]

The disjoint sets of vertices are:

\[
\{0, 1, 2\}, \{3, 4, 5\}, \{6, 7\}
\]

These can be represented by the forest:

\[
\begin{array}{cccc}
1 & 5 & 6 \\
\downarrow & \downarrow & \downarrow \\
2 & 0 & 4 & 3 \\
\end{array}
\]

5.3 Edge Addition

Suppose now we add an edge \((u, v)\) to the graph which wasn’t there before.
We check if \(\text{root}(u) = \text{root}(v)\). If so then the addition of the edge \((u, v)\) would form a cycle and we disallow this for a simple graph, so respond accordingly. If \(\text{root}(u) \neq \text{root}(v)\) then we take the union of the tree containing \(u\) with the tree containing \(v\).

5.4 Does Adding an Edge Form a Cycle?

There are two distinct questions here, whether adding an edge forms a cycle and whether a graph already has a cycle.

The first of these we addressed in the previous subsection. Namely when adding the edge \((u, v)\) we check if \(\text{root}(u) = \text{root}(v)\) and if so, a cycle is formed. We have seen that we can check this in \(O(\alpha(n))\) amortized time, which is almost constant.

Example 5.2. Returning to our example:

Does addition of the edge 3 – 6 form a cycle? Well the root of 3 is 5 and the root of 6 is 6 so no.

Does addition of the edge 3 – 4 form a cycle? Well the root of 3 and the root of 4 are both 5 so yes.

5.5 Is There a Cycle in the Graph?

Given the forest for a graph detecting whether a cycle exists is not possible. This is because the forest does not contain this information, it only contains information about whether the vertices are connected by a path or not.
6 Kruskal’s Algorithm

We now discuss how Kruskal’s algorithm can work in $O(E\alpha(V))$ amortized time. In the following pseudocode assume the graph has $V$ nodes and $E$ edges and that $EL$ is a list of the edges in increasing order by weight. The forest $F$ is our disjoint set forest. When the code ends $K$ will contain all the edges in the minimal spanning tree.

We assume that $F$ has been initialized as a forest with $V$ isolated nodes and uses path compression and weighted union. We assume $K$ is an empty list.

\begin{verbatim}
for each edge(u,w) in EL:
    if findrep(u) != findrep(w)
        K.append(edge(u,w))
        F.union(edge(u,w))
    end if
end for
\end{verbatim}

The for loop iterates $E$ times.

In the body of the for loop we need to calculate $\text{root}(u)$ and $\text{root}(v)$ and with $V$ nodes it takes time $\alpha(V)$ for each. If the if is satisfied then we also need to do $F$.union($u,w$) which takes time $\alpha(V)$.

Consequently the body of the for loop takes $O(\alpha(V))$ time so the entire algorithm takes $O(E\alpha(V))$ time.