Union-Find Data Structure

1 Introduction

A union-find data structure, also known as a disjoint-set data structure, is a data structure that can keep track of a collection of pairwise disjoint sets \( S = \{S_1, S_2, \ldots, S_r\} \) containing a total of \( n \) elements. Each set \( S_i \) has a single, arbitrarily chosen element that can serve as a representative for the entire set, denoted as \( \text{rep}[S_i] \).

Specifically, we wish to support the following operations:

- \( \text{MAKE-SET}(x) \), which adds a new set \( \{x\} \) to \( S \) with \( \text{rep}[\{x\}] = x \).
- \( \text{FIND-SET}(x) \), which determines which set \( S_x \in S \) contains \( x \) and returns \( \text{rep}[S_x] \).
- \( \text{UNION}(x, y) \), which replaces \( S_x \) and \( S_y \) with \( S_x \cup S_y \) in \( S \) for any \( x, y \) in distinct sets \( S_x, S_y \).

Furthermore, we want to make these as efficient as possible. We will go through the process of creating a data structure that, in an amortized sense, performs spectacularly.

2 Starting Out

2.1 Linked List Solution

A naïve way to solve this problem is to represent each set with a doubly linked list like so:

![Figure 1: A simple doubly linked list.](image)

We may also have a data structure, such as a hash table mapping elements to pointers, that allows us to access each linked list node in constant time.

We define \( \text{rep}[S_i] \) to be the element at the head of the list representing \( S_i \).

Here are the algorithms for each operation:

- \( \text{MAKE-SET}(x) \) initializes \( x \) as a lone node. This takes \( \Theta(1) \) time in the worst case.
Recitation 3: Union-Find Data Structure

- **FIND-SET(x)** walks left in the list containing \( x \) until it reaches the front of the list. This takes \( \Theta(n) \) time in the worst case.

- **UNION(x, y)** walks to the tail of \( S_x \) and to the head of \( S_y \) and concatenates the two lists together. This takes \( \Theta(n) \) time in the worst case.

### 2.2 Augmenting the Linked List

We can improve the behavior of our linked list solution by augmenting the linked lists such that each node has a pointer to its representative and that we also keep track of the tail of the list and the number of elements in the list at any time, which we will call its weight.

![Figure 2: An augmented linked list.](image)

Now, **FIND-SET(x)** can run in \( \Theta(1) \) time.

We also change the behavior of **UNION(x, y)** to concatenate the lists containing \( x \) and \( y \), updating the rep pointers for all the elements in \( y \).

However, this could still require \( \Theta(n) \) time in the worst case! Imagine if we called **MAKE-SET** for every integer between 1 and \( n \) and then called **UNION(n - 1, n)**, **UNION(n - 2, n - 1)**, \ldots, **UNION(1, 2)**, where at the \( i \)th union we modify a list of length \( i \). The total cost of all the calls to **UNION** is \( 1 + 2 + \ldots + (n - 1) = \Theta(n^2) \).

### 3 First Improvement: Smaller into Larger

However, we can solve this by forcing **UNION** to merge the smaller list into the larger list. If we do this, the above scenario will only modify a list of length 1 every time, resulting in \( \Theta(n) \) running time for these \( n - 1 \) **UNIONS**.

**Claim.** With the first improvement, the amortized running time of \( n \) calls to **MAKE-SET** followed by \( n \) calls to **UNION** is \( \widetilde{O}(n \lg n) \).

**Proof.** We use the accounting method. Suppose the cost of moving a rep pointer is 1. Monitor some element \( x \) and the set \( S_x \) that contains it. After we call **MAKE-SET(x)**, we have \( \text{weight}[S_x] = 1 \). When we call **UNION(x, y)**, one of two things will happen:
Recitation 3: Union-Find Data Structure

- If $weight[S_x] \leq weight[S_y]$, we pay 1 to update $rep[x]$, and $weight[S_x]$ at least doubles.
- If $weight[S_x] > weight[S_y]$, then $rep[x]$ stays unchanged, so we need not pay anything on $x$’s behalf, and $weight[S_x]$ will only increase.

Therefore we pay at most $\lg n$ to update $rep[x]$. Therefore across all $n$ elements, we get an amortized running time of $O(n \lg n)$. □

4 Forest of Trees Representation

One interesting observation we can make is that we don’t really care about the links between nodes in the linked list; we only care about the $rep$ pointers. Essentially, a list can be represented as a forest of trees, where $rep[x]$ will be the root of the tree that contains $x$:

![Figure 3: A forest of trees representation of a disjoint set data structure.](image)

The algorithms will now be as follows:
- **MAKE-SET**($x$) initializes $x$ as a lone node. This takes $\Theta(1)$ time in the worst case.
- **FIND-SET**($x$) climbs the tree containing $x$ to the root. This takes $\Theta(height)$ time.
- **UNION**($x, y$) climbs to the roots of the trees containing $x$ and $y$ and merges sets the parent of $rep[y]$ to $rep[x]$. This takes $\Theta(height)$ time.

4.1 Adapting the First Improvement

Our first trick can be modified to fit the forest of trees representation by merging the tree with the smaller height into the tree with the bigger height. One can then show that the height of a tree will still be $O(\lg n)$. 
Recitation 3: Union-Find Data Structure

5 Second Improvement: Path Compression

Let’s now improve FIND-SET. When we climb the tree, we learn the representatives of every intermediate node we pass. Therefore we should redirect the rep pointers so a future call to FIND-SET won’t have to do the same computations multiple times.

Claim. Let $n$ be the total number of elements we’re keeping track of. With the second improvement alone, the amortized running time of $m$ operations (including FIND-SET) is $O(m \lg n)$.

Proof. Amortization by potential function.
Let $\Phi(x_1, \ldots, x_n) = \sum_i \lg \text{weight}[x_i]$. UNION($x_i, x_j$) will increase the potential of root FIND-SET($x_i$) by at most $\lg n$. This is because since $x_j$’s subtree gets attached to $x_i$’s subtree’s root, we only change weight[root FIND-SET($x_i$)]. Let $W$ denote this quantity. Since $\lg W$ starts as nonnegative ($W \geq 1$ because the tree must contain $x_i$) and cannot exceed $\lg n$ due to the fact that there are only $n$ elements, the potential increases by no more than $\lg n$.

Each step down $p \rightarrow c$ made by FIND-SET($x_i$), except for the first, moves $c$’s subtree out of $p$’s subtree. If, at any particular step, $\text{weight}[c] \geq \frac{1}{2}\text{weight}[p]$, then the potential increases by at least 1, which pays for the step down. Furthermore, there can be at most $\lg n$ steps for which $\text{weight}[c] < \frac{1}{2}\text{weight}[p]$, since each one decreases the size of the tree we’re looking in by more than half. □

6 Why Don’t We Do Both?

If we do both improvements to the tree representation, we get spectacular behavior where the amortized cost of each operation is almost constant!

6.1 CLRS-Ackermann Function

We define a function, $A_k(j)$, with a similar structure to the well-known Ackermann function, as follows:

$$A_k(j) = \left\{ \begin{array}{ll} j + 1 & \text{if } k = 0 \\ A_k^{j+1}(j) & \text{if } k \geq 1 \end{array} \right.$$
Because you’re repeatedly iterating $A_{k-1}$ many times when $k \geq 1$, this function explodes very quickly:

\[
\begin{align*}
A_0(1) &= 2 \\
A_1(1) &= 3 \\
A_2(1) &= 7 \\
A_3(1) &= 2047 \\
A_4(1) &> 2^{2^{2047}}
\end{align*}
\]

Now consider its “inverse” as follows:

\[
\alpha(n) = \min\{k : A_k(1) \geq n\}
\]

While not technically constant, this value is pretty darn close, even for very large values of $n$. For practical purposes, you can easily assume that $\alpha(n) \leq 4$.

### 6.2 How Spectacular?

**Theorem.** With both improvements, the amortized running time of $m$ operations is $O(m\alpha(n))$.

The proof is very long and very tricky. If you’re interested, check out Section 21.4 in CLRS.