

Union Find - Recitation 11

January 11, 2023

1 Union Find.

We have mentioned that for finding efficiently the minimal spanning tree using kruskal, one has to answer quickly about whether a pair of vertices v, u share the same connectivity component. In this recitation we will present a data structure that will allow us both querying the belonging of given item and merging groups at efficient time cost.

The problem defined as follows. Given n items $x_1 \dots x_n$ we would like to maintain the partition of them into disjoint sets by supporting the following operations:

1. $\text{Make-Set}(x)$ create an empty set whose only member is x . We could assume that this operation can be called over x only once.
2. $\text{Union}(x, y)$ merge the set which contains x with the one which contains y .
3. $\text{Find-Set}(x)$ returns a pointer to the set holding x .

Notice that the naive implementation using pointers array, A , defined to store at place i a pointer to the set containing x can perform the Find-Set operation at $O(1)$. The bottleneck of that implementation is that the merging will require from us to run over the whole items and change their corresponding pointer at A one by one. Namely, a running time cost of $\Theta(n)$ time. Let's review a different approach:

Linked Lists Implementation. One way to have a non-trivial improvement is to associate for each set a linked list storing all the elements belonging to the set. Each node of those linked lists contains, additionally to its value and its sibling pointer, also a pointer for the list itself (the set). Consider again the merging operation. It's clear that having those lists allow us to union sets by iterating and updating only the elements belong to them. Still one more trick is needed for achieving a good running cost.

Union(x, y)

```
1 if size A[x] ≥ size A[y] then
2   size A[x] ← size A[x] + size A[y]
3   for z ∈ A[y] do
4     A[z] ← A[x]
5   A[x] ← A[x] ∪ A[y] // O(1) concatenation of linked lists.
6 else
7   Union(y, x)
```

Clearly, executing the above over sets at linear size require at least linear time. Let's analyze what happens when merging n times. As we have already seen at graphs, runtime can be measured by counting the total number of operations that each item/vertex do along the whole running. So we can ask ourselves how many times does an item change his location and his set pointer. Assume that at the time when x were

changed $A[x]$ contained (before the merging) t elements then immediately after that $A[x]$ will store at least $2t$ elements. In other words

$$\text{size}A^{(t+1)}[x] \leftarrow \text{size}A^{(t)}[x] + \text{size}A^{(t)}[y] \geq 2A^{(t)}[x]$$

Union By Rank.

Path Compression. Let's analyse the cost of queries m times by counting the edges on which the algorithm went over. Let's denote by $\text{find}(v^{(t)})$ the query which requested at time t and let $P^{(t)} = v, v_2, \dots, v_k$ be the vertices path on which the algorithm climbs from v up to its root. Now, observe that by compressing the path the ranks of the vertices in P must be distinct. Now consider any partition of the line into buckets $B_i = [b_i, b_{i+1}]$. $\sum_v \sum_{B_i} \sum_{v \rightarrow u} \mathbf{1}_{r[v]=r[u]} \leq \sum_v \sum_B |B|$

$$\begin{aligned} T(n, m) &= \text{direct parent move} + \text{climbing moves} = \\ &= \text{direct parent move} + \text{stage exchange} + \text{inner stage} = \\ &\leq m + m \cdot |\mathcal{B}| + \sum_{B \in \mathcal{B}} \text{steps inside } B \\ &\leq m + m \cdot |\mathcal{B}| + \sum_{B \in \mathcal{B}} \sum_{u \in B} \text{steps inside } B \text{ started at } u \\ &\leq m + m \cdot |\mathcal{B}| + \sum_{B \in \mathcal{B}} \sum_{u \in B} |B| \\ &\leq m + m \cdot |\mathcal{B}| + \sum_{B \in \mathcal{B}} \frac{n}{\min B} |B| \end{aligned}$$