ECE430.217 Data Structures

Bucket Sort

Textbook: Weiss Chapter 7.11

Byoungyoung Lee

https://compsec.snu.ac.kr

byoungyoung@snu.ac.kr

Lecture slides were prepared based on Douglas W. Harder's notes (**dwharder@alumni.uwaterloo.ca).**

Outline

Recall:

- All exchange (swapping) based sorting takes $\Theta(n^2)$
- All comparison based sorting takes $\Theta(n \ln(n))$

Bucket sort makes **assumptions about the data** being sorted

- Consequently, we can achieve **better than $\Theta(n \ln(n))$** run times

We will look at:

- a supporting example
- the algorithm
- run times (no best-, worst-, or average-cases)
- summary and discussion

Supporting example

Suppose we are sorting a large number of local phone numbers, for example, all residential phone numbers in the 02 area code region

Goal: We want something faster than heap/merge/quicksort!

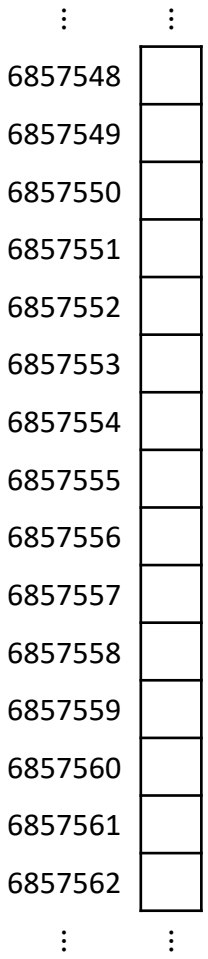
Supporting example

Consider the following scheme:

- Create a bit vector with $10\,000\,000$ bits
  - This requires $10^7/1024/1024/8 \approx 1.2$ MB
- Initially set each bit to 0 (indicating false)
- For each phone number, set the bit indexed by the phone number to 1 (true)
- Once each phone number has been checked, walk through the array and for each bit which is 1, record that number

Supporting example

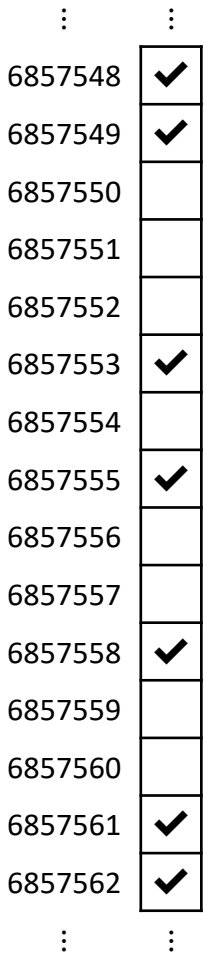
For example, consider this section within the bit array



Supporting example

For each phone number, set the corresponding bit

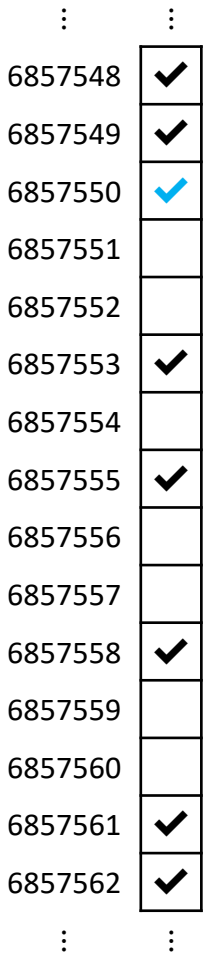
- For example, 685-7550 is a residential phone number



Supporting example

For each phone number, set the corresponding bit

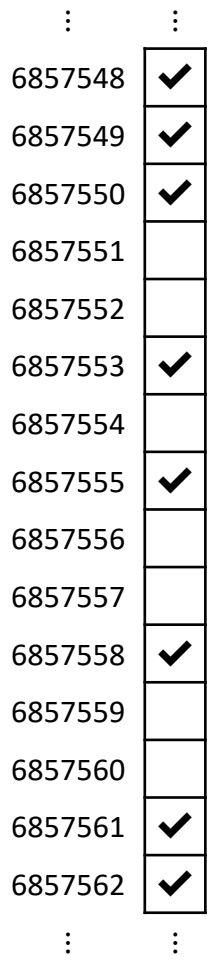
- For example, 685-7550 is a residential phone number



Supporting example

At the end, we just take all the numbers out that were checked:

…,685-7548, 685-7549, 685-7550,
685-7553, 685-7555, 685-7558,
685-7561, 685-5762, …



Supporting example

In this example, the number of phone numbers ($4\,000\,000$ or n) is similar to the size of the array ($10\,000\,000$ or m)

The runtime of such an algorithm is $\Theta(n+m)$:

- we make one pass through the data (n),
- we make another pass through the array and extract the phone numbers which are true (m)

Summary

By assuming that the data falls into a given range, we can achieve $\Theta(n)$ sorting run times

As any sorting algorithm must access any object at least once, the run time must be $\Theta(n)$

References

Wikipedia, http://en.wikipedia.org/wiki/Bucket_sort

[1] Donald E. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching*, 2nd Ed., Addison Wesley, 1998, §5.1, 2, 3.

[2] Cormen, Leiserson, and Rivest, *Introduction to Algorithms*, McGraw Hill, 1990, p.137-9 and §9.1.

[3] Weiss, *Data Structures and Algorithm Analysis in C++, 3rd Ed.*, Addison Wesley, §7.1, p.261-2.

[4] Gruber, Holzer, and Ruepp, *Sorting the Slow Way: An Analysis of Perversely Awful Randomized Sorting Algorithms*, 4th International Conference on Fun with Algorithms, Castiglioncello, Italy, 2007.