EcoLab1

Создание компонента, реализующего алгоритм сортировки подсчетом

Выполнила
Слепнева Мария
21ПИЗ

1. Алгоритм

Сортировка подсчетом - это алгоритм сортировки, который может использоваться для сортировки элементов в пределах определенного диапазона, основанный на количестве каждого сортируемого элемента в исходном массиве. При сортировке подсчетом поддерживается вспомогательный массив, который значительно увеличивает требования к памяти для реализации алгоритма. Эта сортировка базируется на целочисленных значениях исходного массива, также она не является in-place сортировкой и считается стабильной (сохраняет относительный порядок равных элементов).

Это работает следующим образом. Сначала мы вычисляем максимальное значение во входном массиве. Затем мы подсчитываем количество вхождений каждого элемента массива от 0 до длины-1 и присваиваем его вспомогательному массиву. Затем этот массив снова используется для извлечения отсортированной версии входного массива (проходимся по вспомогательному массиву и вставляем значения в исходный на основании частоты

```
countingSort(arr, n)
  maximum <- find the largest element in arr
  create a count array of size maximum+1
  initialise count array with all 0's
  for i <- 0 to n
    find the total frequency/ occurrences of each element and
    store the count at ith index in count arr

  for i <- 1 to maximum
    find the cumulative sum by adding current(i) and prev(i-1) count and store
        it in count arr itself

  for j <- n down to 1
    copy the element back into the input array
    decrement count of each element copied by 1
```

встречания элементов).
Псевдокод

2. Сложность

Поскольку подсчет вхождения каждого элемента во входном диапазоне занимает k времени, а затем поиск правильного значения индекса каждого элемента в отсортированном выходном массиве занимает n времени, таким образом, общая временная сложность становится равной O(N + K).

Временная сложность в наихудшем случае

Временная сложность в наихудшем случае возникает тогда, когда данные искажены, то есть самый большой элемент значительно больше, чем другие элементы. Это увеличивает диапазон, который придется проходить, чтобы в конечном итоге обработать все числа.

Временная сложность в лучшем случае

Временная сложность в лучшем случае возникает, когда все элементы находятся в небольшом диапазоне, то есть когда массив состоит из какого-то одного повторяющегося элемента. В этом случае подсчет вхождения каждого элемента во входной диапазон занимает постоянное время, а затем поиск правильного значения индекса каждого элемента в отсортированном выходном массиве занимает n раз, таким образом, общая временная сложность уменьшается до O(1 + n), т.е. O(n), то есть сложность становится линейной.

Средняя временная сложность

Чтобы вычислить среднюю временную сложность, сначала фиксируем N и принимаем разные значения k от 1 до бесконечности, в этом случае k вычисляется как (k + 1) / 2, а средний случай будет равен N + (K + 1) / 2. Но поскольку K стремится к бесконечности, K является доминирующим фактором. Аналогично, теперь, если изменим N, то увидим, что и N, и K одинаково доминируют, и, следовательно, мы имеем O (N + K) в качестве среднего случая.

Пространственная сложность сортировки подсчетом

При реализации подсчета сортировки требуется дополнительное пространство, поскольку необходимо создать вспомогательный массив размером K + 1. Следовательно, сложность пространства равна: O(K).

Выводы

Алгоритм сортировки подсчетом - это алгоритм сортировки, не основанный на сравнении, т.е. расположение элементов в массиве не влияет на ход выполнения алгоритма. Независимо от того, отсортированы ли элементы в массиве уже, отсортированы в обратном порядке или случайным образом, алгоритм работает одинаково для всех этих случаев, и, следовательно, временная сложность для всех таких случаев одинакова, т.е. O(N + K).

Сортировка с подсчетом наиболее эффективна, если диапазон входных значений не превышает количество значений, подлежащих сортировке. В этом сценарии сложность сортировки с подсчетом намного ближе к O(N), что делает ее алгоритмом линейной сортировки.

Хоть данный алгоритм имеет линейную временную сложность, я не могу сказать, что это лучший подход, потому что пространственная сложность довольно высока, и он подходит для использования только в сценарии, где диапазон элементов входного массива близок к размеру массива. Также минусом является тот факт, что данная сортировка работает только на массивах с целочисленными типами, что делает ее не универсальной.

3. Улучшения

- Случай, который я описала выше, подходит только для сортировки неотрицательных чисел. Но что делать, если в исходном массиве присутствуют отрицательные числа? Для работы сортировки в этом случае необходимо сделать следующие шаги:

1. Определить минимальное и максимальное значение в исходном массиве.
2. Создать дополнительный массив для хранения количества каждого элемента от минимального до максимального значения + 1.
3. Преобразовать числа, добавив к каждому элементу значение минимального элемента (для того, чтобы индексы не были отрицательными).
4. Посчитать количество каждого элемента в исходном массиве и заполнить соответствующие ячейки дополнительного массива.
5. Восстановить отсортированный массив из дополнительного массива, помня, что необходимо снова прибавить минимальное значение к каждому элементу, чтобы вернуться к исходным числам.

В этом случае сложность остается та же, но K здесь - это диапазон значений (наибольшее минус наименьшее число в исходном массиве).

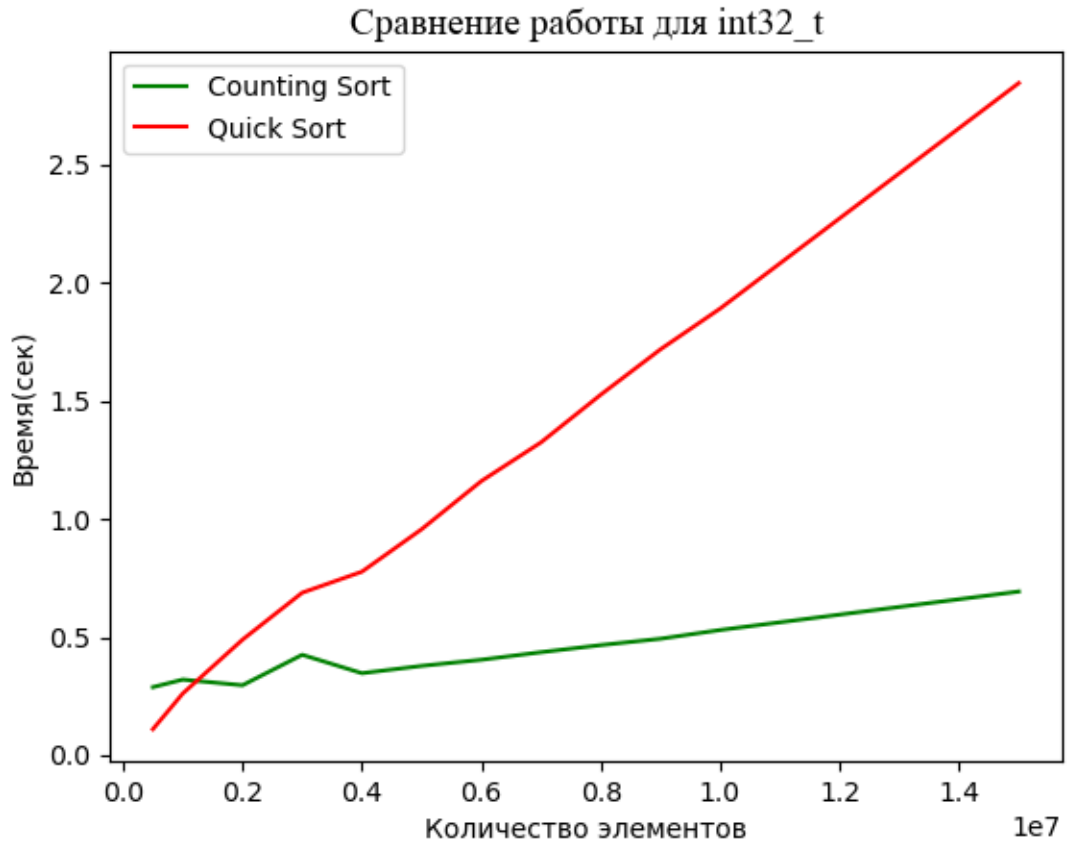
- Также был разработан отдельный компонент для сортировки строки.

Компонент предоставляет пользователю сортировать строку за линейное время.

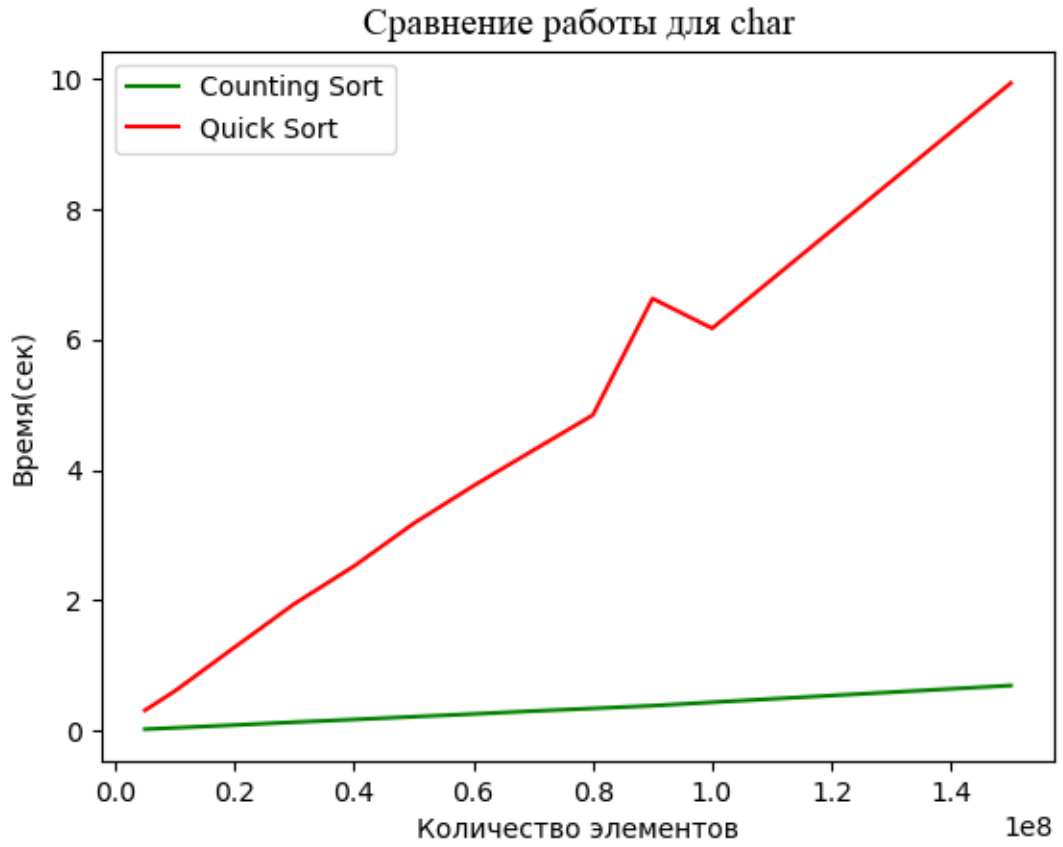
- Таким образом, в ходя работы было разработано два компонента
  - **1F5DF16EE1BF43B999A434ED38FE8F3A** - Компонент с версией сортировки подсчетом для массива целых чисел(int32_t*)
  - **1F5DF16EE1BF43B999A434ED38FE8F3B** - Компонент с версией сортировки подсчетом для строки(char*)

4. Результаты тестирования

Так как данный алгоритм подразумевает работу исключительно на целочисленных значениях, то тестирование программы было проведено именно на этом типе данных. Был рассмотрен только int32, поскольку при попытке сгенерировать доп массив для чисел типа int64 моя программа падала. Думаю, что это зависит от ЭВМ моего компьютера. Очень возможно, что на другом компьютере памяти бы хватило, но все-таки данный эксперимент доказал, что сортировка подсчетом требует больших пространственных затрат, а значит это не самый универсальный способ упорядочивания значений в массиве. Для чистоты эксперимента использовался рандомный int32. Результаты тестирования можно увидеть на графике ниже (время указано в секундах):



Также было проведено тестирование второго компонента. В данном случае сортировка подсчетом показывает лучшие результаты, так как диапазон значений символов составляет 256 элементов.



На основании этого я могу сделать следующий вывод - как видно, сортировка подсчетом действительно имеет линейную сложность. Особенно хорошо это видно в сравнении с qsort. Причем на больших данных сортировка подсчетом показала себя лучше! Но, конечно же стоит учитывать, что за быстроту работы алгоритма пришлось платить очень большими затратами памяти.