Лекция 5 от 26.01.2016

Быстрая сортировка. Продолжение

Говоря об алгоритме быстрой сортировки (QSort), мы рассматривали только случаи, когда все элементы различны. Однако это далеко не всегда так. Если в входном массиве есть равные элементы, то алгоритм может застопориться. Для того, чтобы избежать этого, изменим алгоритм Partition. Попытаемся преобразовывать массив таким образом, чтобы в левой части стояли элементы строго меньшие опорного, в правой — строго большие, а в середине — равные ему:

	$\dots$	x	$\dots$		$\longrightarrow$	$< x$	$= x$	$> x$

Обозначим за опорный элемент последний. Будем проходиться по массиву от начала до конца, выставляя элементы в нужном порядке (? — ещё не просмотренные элементы):

$< x$	$> x$	?	$= x$	x
$[1, i)$	$[i, j)$	$[j, k)$	$[k, n)$	n

Algorithm 1 Модифицированный алгоритм Partition

```
function Partition(a)
    i := 1
    j := 1
    k := n − 1
    while j < k do
        if a[j] = a[n] then
            k := k − 1
            a[j], a[k] := a[k], a[j]
        else
            if a[j] < a[n] then
                a[i], a[j] := a[j], a[i]
                j := j + 1
                i := i + 1
            else
                j := j + 1
```

Заметим, что $j = k$ (так как алгоритм не закончит работу до тех пор, пока это не станет верно). Тогда на выходе получится массив вида:

$< x$	$> x$	$= x$
$[1, i)$	$[i, j)$	$[k, n]$

Остаётся только переставить части массива:

```
j := n
while i < k and j ⩾ n do
    a[i], a[j] := a[j], a[i]
    i := i + 1
    j := j − 1
```

В: Самая быстрая из наших сортировок — $O(n \log n)$. А можно ли быстрее?
О: На основе только сравнений — нет.

Использовать разобранные нами сортировки можно на любых сущностях, для которых определена операция сравнения.

Предположим теперь, что мы сортируем натуральные числа, не превосходящие некоторого числа C.

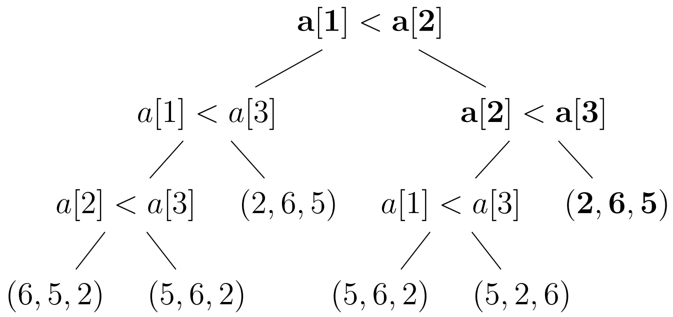
Создадим массив b размера C, заполненный нулями. Будем проходить по исходному массиву a и на каждом шаге будем добавлять 1 к соответствующему элементу массива b:

$$b[a[i]] := b[a[i]] + 1$$

Потом, проходя по получившемуся массиву b, будем восстанавливать исходный массив уже в отсортированном виде.

Такая сортировка будет работать за $O(n)$, однако, она не универсальна.

Вернёмся к универсальным сортировкам. Рассмотрим дерево для массива $a = [6, 5, 2]$:



Подобное дерево можно составить для любого детерминированного[1] алгоритма сортировки, зафиксировав n. Сложность алгоритма будет являться высота h дерева. Посчитаем это h:

- Так как алгоритм должен работать на любой перестановке из n элементов, то у дерева не может быть меньше, чем $n!$ листов.
- Так как сравнение — бинарная операция, то у каждой вершины не более двух потомков. Тогда в дереве не может быть больше, чем 2^h листьев.
- Тогда $2^h \geqslant n! \iff h \geqslant \log_2 n!$. Заметим, что:

$$n! = 1 \cdot 2 \cdot \ldots \cdot \left\lfloor \frac{n}{2} \right\rfloor \cdot \underbrace{\left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right) \cdot \ldots (n-1) \cdot n}_{\text{каждый из } \frac{n}{2} \text{ элементов не меньше } \frac{n}{2}} \geqslant \left(\frac{n}{2}\right)^{\frac{n}{2}}$$

  Тогда $h \geqslant \log_2 \left(\frac{n}{2}\right)^{\frac{n}{2}} = \frac{n}{2} \log_2 \frac{n}{2} = \Omega(n \log n)$. Из этого следует, что отсортировать произвольный массив с помощью только сравнений меньше, чем за $\Omega(n \log n)$ операций, невозможно.

Поиск медианы

Медиана — такой элемент массива, что не меньше половины элементов меньше неё, и не меньше половины — больше.

Для отсортированного массива размера n медиана будет находиться под номером $\dfrac{n+1}{2}$ для нечётных n и $\dfrac{n}{2}$ для чётных n. Пример: для массива (8, 1, 3, 5, 6, 9) медианой будет являться 5.

Как же найти медиану? Очевидно, что можно отсортировать и взять средний — $\Theta(n)$.

[1] Детерминированный алгоритм — алгоритмический процесс, который выдаёт предопределённый результат для заданных входных данных. Например, QSort, выбирающий опорный элемент случайным образом, не является детерминированным.

А можно ли найти медиану ли за линейное время? Можно. Напишем алгоритм, находящий элемент, стоящий на k-ом месте в массиве, получающемся из входного после сортировки. Это называется поиском k-ой порядковой статистики. Составим этот алгоритм, немного модифицировав QSort:

Algorithm 2 Поиск k-ой порядковой статистики

function $\textsc{Select}(a, k)$
    choose pivot $a[p]$
    $i := \textsc{Partition}(a, p)$
    if $i := k$ **then**
        return $a[i]$
    if $i > k$ **then**
        return $\textsc{Select}(a[1 \dots i-1], k)$
    else
        return $\textsc{Select}(a[i+1 \dots n], k-i)$

Как и в быстрой сортировке, неправильно выбранный опорный элемент портит скорость до n^2. Будем выбирать опорный элемент случайным образом. Попробуем посчитать время работы в среднем случае.

j-подзадача размера n'. $\left(\frac{3}{4}\right)^{j+1} n < n' \leqslant \left(\frac{3}{4}\right)^{j} n$

Как и в QSort, в среднем мы потратим две попытки на переход к следующему j.

Максимальное j — $O(\log_{\frac{4}{3}} n)$

$T(n) \leqslant \sum\limits_{j=0}^{\log_{\frac{4}{3}} n} 2 \cdot c \cdot \left(\frac{3}{4}\right)^{j} n = 2cn \sum\limits_{j=0}^{\log_{\frac{4}{3}} n} \left(\frac{3}{4}\right)^{j} \leqslant 2cn$

Время работы алгоритма в худшем случае всё ещё $O(n^2)$. Худший случай — когда на каждом шаге мы отщеплем всего один элемент. Для достижения лучшего случая, на каждом шаге нужно выбирать в качестве опорного элемента медиану.

Медиана медиан

Попробуем несколько модифицировать наш алгоритм. Разобьём входной массив на группы по 5 элементов. Отсортируем каждую такую группу. Так как размер каждой группы зафиксирован, время сортировки не зависит от n. Зависит только количество сортировок. Возьмём медиану в каждой группе и применим алгоритм нахождения медианы к получившемуся массиву медиан. Выберем её в качестве опорного элемента.

Algorithm 3 Поиск k-ой порядковой статистики 2

function $\textsc{Select}(a, k)$
    Divide a into groups of 5
    Choose medians $m_1, \dots m_{\frac{n}{5}}$
    $x = \textsc{Select}([m_1, \dots, m_{\frac{n}{5}}], \frac{n}{10})$
    choose x as pivot $a[p]$
    $i := \textsc{Partition}(a, p)$
    if $i := k$ **then**
        return $a[i]$
    if $i > k$ **then**
        return $\textsc{Select}(a[1 \dots i-1], k)$
    else
        return $\textsc{Select}(a[i+1 \dots n], k-i)$

$T(n) \leqslant cn + T\left(\frac{n}{5}\right) + T\left(\frac{7}{10}n\right)$
$T(n) \leqslant ln$ для некоторого l
$T(n) \leqslant cn + T(\frac{n}{5}) + T(\frac{7}{10}) \leqslant cn + \frac{ln}{5} + \frac{7}{10}ln$