Analysis of Quickselect

Helmut Prodinger
Technical University of Vienna

October 16, 1995

[summary by Bruno Salvy]

Abstract

Quickselect is an algorithm due to Hoare which uses the same partitioning process as Quicksort. As in Quicksort, there is a median-of-three version which reduces the number of comparisons and passes. This is analyzed as well as a variant called multiple Quickselect. All these analyses result in explicit expressions for the number of passes and comparisons.

Quicksort and Quickselect work as follows. The input is an array of n elements. First, one of these elements—the pivot—is selected at random. Then partitioning takes place: the array is rearranged so that its elements smaller than the pivot end up to the left of it, while the elements larger than the pivot end up to the right (see Fig. 1). It is an important hypothesis for the analysis that this partitioning should be *stable*, i.e. the order of the smaller elements and the order of the larger elements should not have been modified during the partitioning. In the next step, Quicksort and Quickselect differ. In Quicksort, whose aim is to sort the array, the same process is applied recursively to both sides of the array. In Quickselect, whose aim is to find the jth element of the array, the process is applied recursively to the side containing it.

In the case of Quicksort, the number of passes and the number of comparisons satisfy recurrences from which follow explicit formulæ in terms of the harmonic numbers $H_n = \sum_{k=1}^{n} 1/k$ [5].

A classical optimization of Quicksort is obtained by selecting the pivot by a median-of-three process: three elements of the array are selected at random, and the pivot is taken to be the median one. The analysis of this optimization is well-known [3, 2]. In [4], the analysis of Quickselect with this optimization is carried out. The same technique is applied to multiple Quickselect in [7]. We now summarize these works.

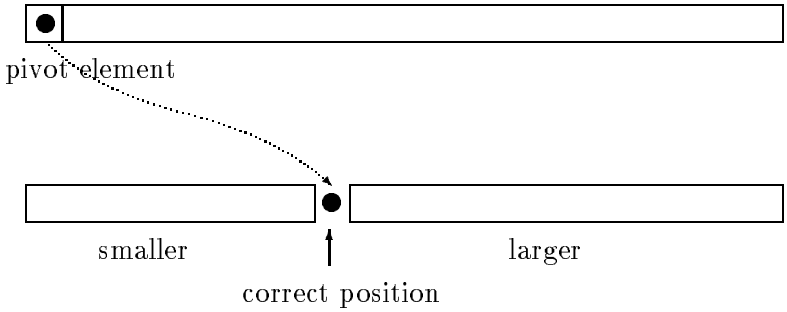

FIGURE 1. The partioning process

1. Number of passes and comparisons

After the pivot has been selected by the median of three process, the probability that the partitioning yields two sub-arrays of sizes $(k-1)$ and $(n-k)$ is

$$\pi_{n,k} = \frac{(k-1)(n-k)}{\binom{n}{3}}.$$

Let $F_{n,j}(z)$ denote the probability generating function of the number of passes necessary to select the jth element out of n under the assumption that all $n!$ permutations of the array are equally likely. Then by a simple generating function argument

$$F_{n,j}(z) = z\left[\sum_{k=1}^{j-1} \pi_{n,k} F_{n-k,j-k}(z) + \pi_{n,j} + \sum_{k=j+1}^{n} \pi_{n,k} F_{k-1,j}(z)\right] \tag{1}$$

for $n \geq 3$ while $F_{1,1}(z) = F_{2,1}(z) = F_{2,2}(z) = z$. The expected number of passes is obtained as $P_{n,j} = F'_{n,j}(1)$ and the generating function $P_j(z) = \sum_{n \geq j} P_{n,j} z^n$ satisfies the following mixed shift-differential equation derived from (1):

$$\frac{1}{6} P_j'''(z) = \frac{1}{(1-z)^4} - \sum_{k=3}^{j-1} \binom{k}{3} z^{k-3} + \sum_{k=2}^{j-1} (k-1) z^{k-2} P'_{j-k}(z) + \frac{P'_j(z)}{(1-z)^2}. \tag{2}$$

Since this is really an equation in P'_j, it is convenient to set $D_j = P'_j$. Then, with the help of Maple, it is possible to find closed-form formulæ for $D_1(z)$, $D_2(z)$, etc. All these functions are linear combinations of $(1-z)^{-2}\log(1-z)$, $\log(1-z)$, $(1-z)^{-2}$ and polynomials in z with simple rational coefficients. It is possible to spot patterns in these coefficients and this suggests studying the bivariate generating function $D(z,u)$ of the $D_j(z)$. From (2), it follows that $D(z,u)$ satisfies a linear differential equation:

$$\frac{1}{6}\frac{\partial^2 D}{\partial z^2} - \left(\frac{1}{(1-z)^2} + \frac{u^2}{(1-uz)^2}\right) D = \frac{u}{1-u}\left(\frac{1}{(1-z)^4} - \frac{u^3}{(1-uz)^4}\right),$$

with initial conditions $D(0,u) = u$, $D'_z(0,u) = 2u(1+u)$. This equation turns out to have a (several pages long) closed-form solution involving the logarithms of $(1-uz)$ and $(1-z)$ and rational functions in u and z. Extracting the coefficients then yields the following theorem.

THEOREM 1. *Given a random permutation of n elements and $5 \leq j \leq n-4$, the average* number of passes *needed to select the jth element using Quickselect with a median-of-three partition is*

$$\begin{aligned} P_{n,j} = {} & \frac{24}{35} H_n + \frac{18}{35} H_j + \frac{18}{35} H_{n+1-j} + \frac{12}{35j} + \frac{12}{35(n+1-j)} - \frac{304}{175} - \frac{6}{7n} + \frac{18j}{35n} - \frac{12(j-1)^{\underline{2}}}{35n^{\underline{2}}} \\ & - \frac{4(2j-3)(j-1)^{\underline{2}}}{35n^{\underline{3}}} - \frac{6(j-2)(j-1)^{\underline{3}}}{35n^{\underline{4}}} + \frac{6(2j-5)(j-1)^{\underline{4}}}{35n^{\underline{5}}} - \frac{4(j-3)(j-1)^{\underline{5}}}{35n^{\underline{6}}}, \end{aligned}$$

where $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$.

For instance, to compute the median of $2n+1$ elements requires a number of passes $P_{2n+1,n+1} = \frac{24}{35} H_{2n+1} + \frac{36}{35} H_{n+1} + O(1) = \frac{12}{7}\log n + O(1)$ instead of $2\log n$ in the classical case. The savings are thus about 14%.

The number of comparisons is obtained in a similar fashion. In (1), it is sufficient to replace the factor z by z^{n-1} to obtain the generating function of the number of comparisons (at each pass, there are $n-1$ comparisons during the partitioning). Then again, the bivariate generating function

of the number of comparisons to select the jth element out of a random permutation of n elements can be found explicitly, and extracting the coefficients yields the following theorem.

THEOREM 2. *Given a random permutation of n elements and $5 \le j \le n-4$, the average* number of comparisons *needed to select the jth element using Quickselect with a median-of-three partition is*

$$C_{n,j} = 2n + \frac{72}{35}H_n - \frac{156}{35}H_j - \frac{156}{35}H_{n+1-j} + \frac{36}{35j} + \frac{36}{35(n+1-j)} + \frac{88}{175} + \frac{24}{7n} + 3j - \frac{3(j-1)^{\underline{2}}}{n} - \frac{156j}{35n}$$
$$- \frac{36(j-1)^2}{35n^{\underline{2}}} - \frac{12(2j-3)(j-1)^{\underline{2}}}{35n^{\underline{3}}} - \frac{18(j-2)(j-1)^{\underline{3}}}{35n^{\underline{4}}} + \frac{18(2j-5)(j-1)^{\underline{4}}}{35n^{\underline{5}}} - \frac{12(j-3)(j-1)^{\underline{5}}}{35n^{\underline{6}}},$$

where $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$.

Computation of the median therefore requires $11n/2 + O(\log n)$ comparisons whereas the classical method requires $4(1+\log 2)n + O(\log n)$ comparisons. The savings are thus about 19%.

The same technique also applies to several variants, such as counting only $n-3$ comparisons per partition or selecting the smaller of two random elements as the pivot.

2. Multiple Quickselect

In *multiple Quickselect*, one searches simultaneously for the elements of indices $\{j_1, \ldots, j_p\}$ $(0 < j_1 < \cdots < j_p \le n)$. The analysis is very similar to the analyses above and results in *explicit* formulæ for the number of passes and the number of comparisons. With obvious notation, one has

$$P[n; j_1, \ldots, j_p] = H_{j_1} + H_{n+1-j_p} + 2\sum_{t=2}^{p} H_{j_t+1-j_{t-1}} - 2p + 1,$$
$$C[n; j_1, \ldots, j_p] = 2n + j_p - j_1 + 2(n+1)H_n - 2(j_1+2)H_{j_1} - 2(n+3-j_p)H_{n+1-j_p}$$
$$- 2\sum_{t=2}^{p} (j_t + 4 - j_{t-1})H_{j_t+1-j_{t-1}} + 8p - 2.$$

Of course, as a special case, we recover the analysis of Quicksort when $p = n$.

A recent work of Lent and Mahmoud [6] gives asymptotic estimates for so-called *grand averages*:

$$\mathcal{P}_{n,p} = \frac{1}{\binom{n}{p}} \sum_{1 \le j_1 < \cdots < j_p \le n} P[n; j_1, \ldots, j_p],$$
$$\mathcal{C}_{n,p} = \frac{1}{\binom{n}{p}} \sum_{1 \le j_1 < \cdots < j_p \le n} C[n; j_1, \ldots, j_p].$$

Using the formulæ above and summing the harmonic numbers by direct manipulations or standard generating function techniques [1], it is actually possible to derive closed-form formulæ for these averages in terms of harmonic numbers [7].

THEOREM 3.

$$\mathcal{P}_{n,p} = \frac{2p(n+1)^2}{(n+2-p)(n+1-p)}(H_{n+1} - H_p) + 1 - 2p - \frac{2(p-1)^2}{n+2-p},$$
$$\mathcal{C}_{n,p} = \frac{1}{(n+2-p)(n+1-p)}\left[(2H_p+1)n^3 - 8pH_n n^2 + 4((p+2)H_p + p)n^2\right.$$
$$\left. + 2p(p-9)H_n n + (2(4p+5)H_p - 5p^2 + p - 1)n + 2p(p-5)H_n + 4(p+1)H_p - p(p+7)\right].$$

Bibliography

[1] Graham (R. L.), Knuth (D. E.), and Patashnik (O.). – *Concrete Mathematics*. – Addison Wesley, 1989.

[2] Greene (D. H.) and Knuth (D. E.). – *Mathematics for the analysis of algorithms*. – Birkhauser, Boston, 1981.

[3] Hennequin (Pascal). – Combinatorial analysis of quicksort algorithm. *RAIRO Theoretical Informatics and Applications*, vol. 23, n° 3, 1989, pp. 317–333.

[4] Kirschenhofer (Peter), Martínez (Conrado), and Prodinger (Helmut). – Analysis of Hoare's find algorithm with median-of-three partition. – 1995. Preprint.

[5] Knuth (Donald E.). – *The Art of Computer Programming*. – Addison-Wesley, 1973, vol. 3: Sorting and Searching.

[6] Lent (J.) and Mahmoud (H. M.). – Average-case analysis of multiple quickselect: An algorithm for finding order statistics. *Statistics and Probability Letters*, vol. 28, n° 4, August 1996, pp. 299–310.

[7] Prodinger (Helmut). – Multiple Quickselect – Hoare's Find algorithm for several elements. *Information Processing Letters*, vol. 56, n° 3, November 1995, pp. 123–129.