クイックソートにおける複数ピボットの選択法

2005MM079 高井正人
指導教員：尾崎俊治

1　はじめに

クイックソートは，1960 年にアントニー・ホーアが開発したソーティングアルゴリズムで，一般に最も高速であるとされている．

それは，分割統治法に基づいており，再帰的に整列を繰り返すことにより，整列される．また，クイックソートは，これまでに十分な研究がなされており，平均時間計算量が $O(n\log n)$ であるが，最悪の場合には $O(n^2)$ になることが知られている．

また，クイックソートの性能を図る上で，ピボットの選び方というのはとても重要である．

ピボットは通常，データ列からランダムに選んだり，データ列中の適当な 3 つの数の中央値を選んだりする．こうすることによって，最悪計算時間がかかってしまうような可能性を抑える工夫をするのである．

そこで，今回はピボットの選び方を 1 つの値に絞らず 2 個と 3 個選び，その選び方を工夫することによって，クイックソートの計算時間の短縮を図る．

2　クイックソートの計算量解析

まず，比較できるようピボットが 1 個の場合について解析する．これは単純なクイックソートである．

2.1　平均時間計算量

プログラム実行中に最も多く行われる処理は比較である．よって，ソート 1 回あたりにかかる計算コストは，比較が何回行われているのかを求めることによって抑えられる．そこで，比較回数を計算するためにアルゴリズムに沿って解析してみる．

n 個のデータ列を並べ替えるのに必要な比較回数を X_n とする．ピボット p を $1,2,...,n$ の中からランダムに選び，その値を a_p とすると，X_n はその他の要素との比較回数 $n-1$ 回と，a_p より「小さい」グループと「大きい」グループの比較回数の合計で求められる．p は $a[0],a[1],...,a[n-1]$ から一様に $\frac{1}{n}$ の確率で選ばれるとし，a_p が i 番目に小さい値であったとすると，「小さい」グループの比較回数は X_{i-1}，「大きい」グループの比較回数は X_{n-i} と表すことができる．従って，X_n は以下のように定義できる．

$$\begin{aligned}X_n = n-1+\frac{1}{n}\{&(X_0+X_{n-1})\\&+(X_1+X_{n-1})+\cdots+(X_{n-1}+X_0)\}\end{aligned}\quad(1)$$

(1) の平均を計算すると，$Q_n = E[X_n]$ を用いて，

$$Q_n = E[X_n] = 2\,(n+1)\,H_n - 4n \qquad (2)$$

となる [1]．ただし，H_n は調和数を表す．すなわち

$$H_n = \sum_{k=1}^{n}\frac{1}{k} = 1+\frac{1}{2}+\cdots+\frac{1}{n}$$

である．式 (2) より，クイックソートの平均計算量が $O(n\log n)$ となることが確認できた．

2.2　分散

期待値と考え方を同様に，それぞれ $1,2,...,n$ までピボットを選んだときの確率を期待値から引いて 2 乗し，その平均をとると，

$$\begin{aligned}V[X_n] = \frac{1}{n}\Big[&\{E[X_n]-(n-1+Q_0+Q_{n-1})\}^2\\&+\{E[X_n]-(n-1+Q_1+Q_{n-1})\}^2\\&+\cdots\\&+\{E[X_n]-(n-1+Q_{n-1}+Q_0)\}^2\Big]\end{aligned}$$

となり，これを計算すると，

$$V[X_n] = -4\,(n+1)^2\,H_n^{(2)} - 2\,(n+1) + H_n\,(7n+13)\,n \qquad (3)$$

$$\approx \left(7-\left(2\pi^2/3\right)\right)n^2 \qquad (4)$$

となり，以上のように近似できる [2][3].

2.3　最悪計算量

最悪計算量とは，クイックソートの時間計算量の取り得る値の下限である．

最悪計算量を B_n とすると，B_n はピボットがすべて最小か最大の時であり，

$$\begin{aligned}B_n &= n-1+2B_{n-1}\\&= \frac{1}{2}n^2-\frac{1}{2}n\end{aligned}$$

となる．これで最悪時の $O(n^2)$ となることが確認できた．

3　複数ピボット

図 1 は，データの個数 $n=1000$ を $\frac{1}{2},\frac{1}{3}$ にしたときに，それらを並び替えるのに必要な比較回数を示したグラフである．左から順に $n=333,500,1000$ のときのグラフである．

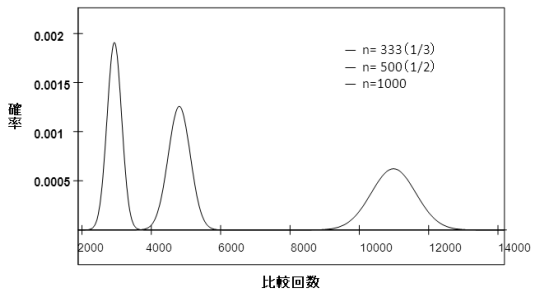

図 1　$n=333,500,1000$ のときの比較回数の分布

図 1 から分かるように，グラフは正規分布となり，n の値が小さくなるほど比較回数が少なくて済む．それに加え，比較回数のばらつきが少なくなるので，最悪の場合の比較回数 ($O(n^2)$) になる確率が低くなる．よって，大きな数のソートを行う場合は，早い段階でより少ない数の問題へと分割することが重要といえる. そこで，ピボットを増やすことにより，早い段階で複数の小さな問題へと分割することのできる複数ピボットの選択は有効だと考えることができる．

ここで，ピボットを複数に増やした場合の計算量を求める．

3.1 2 個の場合

ピボットを 2 個に増やしたときは，グループの分割が 3 等分になることが望ましい．そこで，ピボットを 2 個選んだときの計算量を $X_n^{(2)}$ とし，ピボットを $p_1 = a_1, p_2 = a_2$（ただし $a_1 < a_2$）とすると，

$$\begin{aligned} X_n^{(2)} &= (n-1) + (n-i-1) + X_{i-1}^{(2)} + X_{j-i-1}^{(2)} + X_{n-j}^{(2)} \\ &= 2n - i - 2 + X_{i-1}^{(2)} + X_{j-i-1}^{(2)} + X_{n-j}^{(2)} \end{aligned}$$

となる．この平均計算量を求めると，$R_n = E[X_n^{(2)}]$ を用いて，

$$R_n = \frac{2(5n-8)}{n(n-1)} + \frac{2(n-2)}{n} R_{n-1} - \frac{(n-5)}{(n-1)} R_{n-2}$$

という漸化式が求まる．これより一般項

$$R_n = 2\,(n+1)\,H_n - (\frac{19}{5}n + \frac{3}{10}) \qquad (n \geq 4) \tag{5}$$

が求まる．

3.2 3 個の場合

さらに，ピボットを 3 個に増やしたときには，グループの分割が 4 等分になることが望ましい．これは，単純なクイックソートを 2 段階まで進めた時と同じことである．そこで，ピボットを 3 個選んだときの計算量を $X_n^{(3)}$ としピボットを $p_1 = a_1, p_2 = a_2, p_3 = a_3$（ただし $a_1 < a_2 < a_3$）とし計算すると，

$$\begin{aligned} X_n^{(3)} &= \alpha_3 + (n-1) + (j-2) + (n-j-1) \\ &\quad + X_{i-1}^{(3)} + X_{j-i-1}^{(3)} + X_{k-j-1}^{(3)} + X_{n-k}^{(3)} \\ &= \alpha_3 + 2(n-2) \\ &\quad + X_{i-1}^{(3)} + X_{j-i-1}^{(3)} + X_{k-j-1}^{(3)} + X_{n-k}^{(3)} \end{aligned}$$

となる．この平均計算量を求めると，$S_n = E[X_n^{(3)}]$ を用いて，

$$\begin{aligned} S_n = {} & \frac{12(4n-11)+\alpha_3}{n(n-1)(n-2)} + \frac{3(n-3)}{n} S_{n-1} \\ & + \frac{3(n-3)(n-4)}{n(n-1)} S_{n-2} + \frac{(n^2-11n+36)}{n(n-2)} S_{n-3} \end{aligned}$$

という漸化式が求まる．

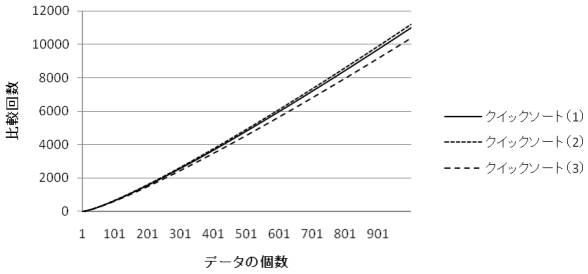

図 2 計算量の比較

4 複数ピボットとの比較

図 2 は，単純なクイックソートと，ピボットを 2 個，3 個に増やしたときの計算量を比較したグラフである．上から順に，ピボットが 2 個，単純なクイックソート，ピボットが 3 個のときの計算量である．ピボットが 2 個の場合，式 (2) と式 (5) から分かるように，$\frac{1}{5}n - \frac{3}{10}$ だけ計算量が増えた．これは，$n = 1000$ のときで約 1.8 %の増加である．

ピボットが 3 個の場合は，n について閉じた式を求めることはできなかったが，Excel で値を求めたところ $\alpha_3 = 2.5$ とすると，$n = 1000$ のときで約 5.5 %減少していることが分かった．

5 おわりに

複数ピボットとして，ピボットをを 2 個と 3 個選んだときの平均計算量を求めることができた．ピボットが 2 個の場合は計算量が増えてしまい改善されたとは言えないが，ピボットが 3 個の場合は計算量が n が十分に大きいとき 5 %以上減少することが分かった．これは 3 要素の中央値で行うクイックソート [4] と構造が非常によく似ているためであると考えられる．また，n が小さいときに他の効率のよいアルゴリズムに換えることで，もっと実行時間を縮められると思われる．

参考文献

[1] GRAHAM/KNUTH/PATASHNIK/著，有澤 誠/安村通晃/萩野達也/石畑 清/訳:『コンピュータの数学』. 共立出版株式会社，東京，1993.

[2] 溝井直史, 尾崎俊治:『クイックソートの時間計算量の確率的解析』. 電子情報通信学会論文誌 '95/9，Vol.J78-A，No.9，pp.1142-1148

[3] Luc Devroye,James Allen Fill,Ralph Neininger：*PerfectSimulationfromtheQuicksortLimit Distribution*. Elect. Comm. in Probab. 5 (2000) http://www.emis.ams.org/journals/EJP-ECP/_ejpecp/ECP/include/getdoc807c.pdf?id=3455&article=1607&mode=pdf

[4] R. セジウィック（野下浩平，星守，佐藤創，田口東訳）:『アルゴリズム　第 1 回=基礎・整列』，第 2 版. 近代科学社，東京，1990.