第4卷第2期
1993年4月

软 件 学 报
JOURNAL OF SOFTWARE

Vol. 4, No. 2
Apr., 1993

分段快速排序法

唐 向 阳

(西南民族学院,成都 610041)

FAST SORTING METHOD OF SEPARATING SEGMENT

Tang Xiangyang

(*Southwest Nationalitis Institute, Chengdu* 610041)

Abstract In this paper, a fast sorting method of separating segment is given. For given N data recordings, the maximum mean of sorting time is O(N). The experiment results from mean distribution data recordings and normal distribution data recordings with three sorting methods on IBM—PC are given.

摘要 本文给出分段快速排序方法. 对于给定的N个数据记录,此方法的最大平均排序时间为O(N). 本文最后给出利用三种快速排序方法在IBM—PC机上分别关于均匀分布数据记录和正态分布数据记录进行排序的实验结果.

§1. 引 言

排序作为一项复杂而重要的技术在计算机科学中引起人们的广泛重视,不少专家、学者对排序问题进行了深入的探讨,给出了许多不同的排序方法[1]. Hoare 提出的快速排序法一直被认为具有最佳的平均性能,能达到的平均排序时间 $O(N \cdot \log_2 N)$[1],其中 N 为需排序的数据记录个数,并且在所有实施反复比较和交换这二种操作以及同数量级的排序方法中,Hoare 快速排序法在平均排序时间上是最佳的[1],文献[2]给出了一种采用增加 2N+1 个存储空间方式的快速分组排序方法,并从理论上证明了此方法平均排序时间为 O(N),优于 Hoare 快速排序法.

本文给出的分段快速排序法,在附加 2N+[(N+1)/2]+1 个存储空间的代价下,平均排序时间达到 O(N),并且优于快速分组排序方法,更优于 Hoare 快速排序方法,其中[X]表示将 X 的值取整;在附加 N+[(N+1)/2]+1 个存储空间的代价下,排序时间比快速分组排序方法所用的时间稍多一点,但平均排序时间仍为 O(N),仍然优于 Hoare 快速排序方法.

本文1990年9月8日收到,1991年3月5日定稿. 本课题是国家民委科研资助项目. 作者**唐向阳**,讲师,1989年硕士毕业于西安交通大学,主要研究领域为计算机辅助几何设计、计算机图形学、软件算法的研究.

§2. 排序算法描述

对给定的一组数据记录：$D_1, D_2, \cdots, D_N$ 排序，就是在计算机上经过一定的计算处理，把数据记录$\{D_i\}_{i=1}^N$排成递增或递减的数据记录序列. 分段快速排序法是将数据记录排成递增的数据记录序列：$D_{(1)} \leqslant D_{(2)} \leqslant \cdots \leqslant D_{(N)}$，其中 $D_{(j)} \in \{D_i\}_{i=1}^N$. 分段快速排序算法具体描述如下：

1. 给定 N 个数据记录$\{D_i\}_{i=1}^N$. 求出最大值 $D_{max} = \max\limits_{1 \leqslant i \leqslant N} \{D_i\}$和最小值 $D_{min} = \min\limits_{1 \leqslant i \leqslant N} \{D_i\}$. 若 $D_{max} - D_{min} = 0$. 则勿需进行排序；否则往下执行.

2. 取 $M = [(N+1)/2]$，将区间$[D_{min}, D_{max}]$等分成 M 个小区间，它们的长度均为：$H = (D_{max} - D_{min})/M$.

3. 开辟 N+M+1 个附加存储空间：N 个元素的数组 R 和 M+1 个元素的数组 P. 初始化数组 P，数组 R 和 P 的元素分别表示每个数据记录所在段的段号和某一段中数据记录的个数，它们分别由下列式子求出：

$$R(i) := [(D_i - D_{min})/H + 1] \qquad i = 1, 2, \cdots\cdots, N. \tag{1}$$

$$P(R(i)) := P(R(i)) + 1 \qquad i = 1, 2, \cdots\cdots, N. \tag{2}$$

这样便将数据记录$\{D_i\}_{i=1}^N$分成 M+1 段，其中第 M+1 段中的数据记录均等于 D_{max}.

4. 求出各段中第一个数据记录的位置，并将其结果仍放入数组 P：

$$P(M+1) := N - P(M+1) + 1, \qquad P(1) := 1, \tag{3}$$

$$P(j) := P(j+1) - P(j), \qquad j = M, M-1, \cdots\cdots, 2 \tag{4}$$

5. 开辟 N 个附加存储空间 Q，将数据记录$\{D_i\}_{i=1}^N$依照下列方式传入数组 Q：

$$Q(P(R(i))) := D_i, \qquad P(R(i)) := P(R(i)) + 1, i = 1, 2, \cdots\cdots, N \tag{5}$$

这样的数组 Q 中的数据记录以段为单位是排好序的，即第 j 段中的数据记录均小于第 j+1 段中的数据记录.

6. 如果第 j 段中数据记录个数≥2，则利用 Hoare 快速排序法对数组 Q 中该段内的数据记录进行排序，j=1,2,……,M.

完成以上各步骤，便完成对数据记录$\{D_i\}_{i=1}^N$的排序工作，并且结果依次存放在数组 Q 中.

§3. 算法分析

3.1 时间复杂性

引理 1：设 ξ 为[0,1]均匀分布的随机变量，数据$\{X_i\}_{i=1}^L$为 ξ 的 L 个相互独立的观测值. 将区间[0,1]等分成 K(≥2)个小区间，则在任一小区间内平均有不超过 L/(K+1)个观测值.

证明：数据$\{X_i\}_{i=1}^L$中落入任一个小区间的数据个数的数学期望为：$L \cdot (1-1/K)/K \leqslant L/(K+1)$. 证毕.

引理 2：[3]设连续分布的随机变量 η 有分布函数 $F(x)$，则 $\xi=F(\eta)$ 为[0,1]上均匀分布的随机变量.

定理 1：若数据记录 $\{D_i\}_{i=1}^N$ 为均匀分布随机变量 ξ 的 N 个独立观测值，则对 $\{D_i\}_{i=1}^N$ 进行分段快速排序所需的平均时间为 $O(N)$.

证明：显然完成算法中步骤 1——步骤 5 所需时间为 $O(N)$.

不失一般性，设 ξ 为[0,1]上均匀分布的随机变量，则 $D_{min}\geqslant 0, D_{max}\leqslant 1$. 区间 $[D_{min}, D_{max}]$ 已被等分成 M 个小区间. 记落入第 i 个小区间中观测值的个数为 N_i 和该小区间中数据记录用 Hoare 快速排序法排序所需平均时间为 T_i，由引理 1，有 $N_i\leqslant 2$，从而 $T_i=K_i\cdot N_i\cdot \log_2 N_i=K_iN_i, i=1,2,\cdots\cdots,M$，其中 K_i 均为常数因子. 于是算法中步骤 6 所需的平均时间不超过 $\sum_{j=1}^{m}T_j\leqslant\sum_{j=1}^{m}K_j\cdot N_j\leqslant K\cdot N$，其中 $K=\max_{1\leqslant j\leqslant M}\{K_j\}$.

因此，对于均匀分布的数据记录用分段快速排序法排序所需平均时间为 $O(N)$. 证毕.

定理 2：若数据记录 $\{D_i\}_{i=1}^N$ 为任意连续分布随机变量 η 的 N 个独立观测值，则对 $\{D_i\}_{i=1}^N$ 进行分段快速排序所需平均时间为 $O(N)$.

证明：设随变量 η 的分布函数为 $F(x)$. 由引理 2，$\xi=F(\eta)$ 为[0,1]上均匀分布的随机变量，$\{C_i: C_i=F(D_i)\}_{i=1}^N$ 为 ξ 的 N 个独立观测值. 显然，从 $\{D_i\}_{i=1}^N$ 变到 $\{C_i\}_{i=1}^N$ 所需时间为 $O(N)$. 由定理 1，对 $\{C_i\}_{i=1}^N$ 用分段快速排序法进行排序所需的平均时间为 $O(N)$. 设 $\{C_{(i)}\}_{i=1}^N$ 已排好序. 由于函数 $F^{-1}(y)$ 单调递增[3]，则从 $\{C_{(i)}\}_{i=1}^N$ 变回到 $\{D_{(i)}\}_{i=1}^N$ 后，$\{D_{(i)}\}_{i=1}^N$ 便是排好序的，所需时间为 $O(N)$.

故整个排序所需平均时间为 $O(N)$. 证毕.

在实际应用中，通常我们并不知道数据记录 $\{D_i\}_{i=1}^N$ 所服从概率分布的分布函数 $F(x)$，既使知道 $F(x)$，也很难计算 $C_i=F(D_i)$ 或 $D_{(i)}=F^{-1}(C_{(i)})$. 因此只能用分段快速排序法对数据记录 $\{D_i\}_{i=1}^N$ 直接进行排序. 对大多数服从连续概率分布的数据记录来说，直接用分段快速排序法进行排序的平均时间仍可达到 $O(N)$. 特别地有：

定理 3：若数据记录 $\{D_i\}_{i=1}^N$ 为正态分布随机变量 ξ 的 N 个独立观测值，则对 $\{D_i\}_{i=1}^N$ 直接进行分段快速排序所需的平均时间为 $O(N)$.

证明：不失一般性，设 $\{D_i\}_{i=1}^N$ 服从标准正态分布 $N(0,1)$，密度函数为

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

因为 $\int_{-5}^{5}\varphi(t)dt=0.9999994$，因此可近似地认为 $D_{max}-D_{min}\leqslant 10$. 设 $N\geqslant 10$.

首先利用算法的步骤 1——步骤 5，将记录 $\{D_i\}_{i=1}^N$ 分成 M+1 段，其中第 i 段中的数据记录均落入小区间 $[D_{min}+(i-1)\cdot H, D_{min}+i\cdot H]$ 内 $(i=1,2,\cdots,M)$，而落入每个小区间内的数据记录个数的最大期望值不大于：

$$\left.\begin{matrix}N\cdot\int_0^H\varphi(t)dt\\ N\cdot\int_{-H/2}^{H/2}\varphi(t)dt\end{matrix}\right\}\leqslant\frac{N\cdot H}{\sqrt{2\pi}}\leqslant\frac{20}{\sqrt{2\pi}}\cdot\frac{N}{N-1}$$

$$\leqslant 8.865384,\quad \begin{matrix} M\text{ 为偶数.} \\ M\text{ 为奇数.} \end{matrix}$$

此外，当小区间向期望值 $\mu=0$ 的两端移动时，落在此小区间内的数据记录个数将会迅速下降.

记落入第 j 个小区间中的数据记录个数为 N_j，$j=1,2\cdots\cdots,M$，则执行算法步骤 6 所需的平均时间不超过：$\sum_{j=1}^{m} K_j \cdot N_j \cdot \log_2 N_j < 3.15 \cdot K \cdot N$，$K=\max_{1\leqslant j\leqslant M}\{K_j\}$ 为常数.

故对正态分布的数据记录 $\{D_i\}_{i=1}^{N}$ 直接进行分段快速排序所需平均时间为 O(N). 证毕.

当然，对正态分布的数据记录直接用分段快速排序法排序所需的平均时间稍长于对均匀分布数据记录直接排序所需的平均时间. 下面的实验结果也说明了这一点.

如果需要排序的数据记录分布极不均匀，过分集中在某些段内，最坏的情况是 N−1 个数据记录(除 D_{max} 外)全部集中在第 j 段内($j\neq M+1$). 这时分段快速排序法就不如 Hoare 快速排序法. 关于这种情况，参考下面的实验结果，可作如下处理：如果某段内的数据记录个数＞100时，则对此段中的数据记录执行算法的步骤 6 之前，重复执行一遍算法步骤 1——步骤 5(需再增加 N+M+1 个附加存储空间)，然后对所得到的每个小段中的数据记录执行算法的步骤 6，这样能使排序的平均时间近似为 O(N).

3.2 空间复杂性

分段快速排序法需要 2N+M+1 个附加存储空间，以此为代价，与 Hoare 快速排序法相比，所需的平均排序时间从 $O(N\cdot\log_2 N)$ 降低到 O(N). 与快速分组排序法相比，在多增加 M 个存储空间的代价下，排序时间有所减少，可以从下面的实验结果看到这一点.

若用普通变量 R 代替数组 R，则所需的附加存储空间为 N+M+1 个，这时排序的平均时间仍为 O(N)，仍优于 Hoare 快速排序法. 因为在(5)式之前需要利用(1)式重新算出每个数据记录所在的段号，因此排序时间比快速分组排序法所需时间稍长一点，但与其相比附加存储空间减少了 M 个.

3.3 算法的稳定性

显然分段快速排序法是稳定的.

总之，分段快速排序法的平均排序时间为 O(N)，优于 Hoare 快速排序法. 就其阶数而言，已达到了排序算法运算量的下限. 但此算法在所需排序的数据记录分布极不均匀时，速度将受到较大影响，因此，此算法还有待改进.

§4. 实验结果

为了比较分段快速排序法、快速分组排序法和 Hoare 快速排序法的实际功能，笔者在 IBM-PC 微机上用机器本身的时钟测量时间，采用 PASCAL 语言编程，利用这三种排序方法分别对均匀分布的数据记录和正态分布的数据记录作了实验，下面表中的“分段快速排序法 I”采用附加 N+M+1 个存储空间的方式编程；“分段快速排序法 II”采用附加 2N+M+1 个存储空间的方式编程. 由表 1 和表 2 不难看出，实验结果与理论结论是一致的.

表1 对N个均匀分布随机数据记录三种排序方法时间对比

单位：秒

排序方法 \ 数据记录量(N)	100	500	1000	2000	3000	4000	5000
Hoare 快速排序法	0.38	2.80	6.59	14.22	22.50	32.30	40.54
快速分组排序法	0.60	2.86	5.71	11.43	17.80	23.13	29.11
分段快速排序法 I	0.62	2.98	5.88	11.70	18.13	23.45	29.60
分段快速排序法 II	0.39	1.98	3.90	7.69	11.97	15.54	19.66

表2 对N个正态分布随机数据记录三种排序方法时间对比

单位：秒

排序方法 \ 数据记录量(N)	100	500	1000	2000	3000	4000	5000
Hoare 快速排序法	0.40	2.87	6.20	14.67	23.07	34.17	43.89
快速分组排序法	0.60	2.91	5.77	11.53	18.03	23.96	30.07
分段快速排序法 I	0.66	3.25	6.68	13.12	20.09	25.81	31.27
分段快速排序法 II	0.44	2.25	4.58	9.29	14.12	19.66	24.38

参考文献

[1]严蔚敏，吴伟民，《数据结构》，清华大学出版社，292——318.

[2]张建中，快速分组排序，数值计算与计算机应用，Vol. 9, No. 2(1988)，139——143.

[3]复旦大学，《概率论》(第一册 概率论基础)，人民教育出版社，148——149.