快速排序算法的平均计算量分析

排序问题

给定 n 个数字，$(x_1, x_2, \cdots, x_n)$，按照从小到大的顺序将其排列。

冒泡法：

冒泡法的基本想法是从数列的开始，对元素进行两两比较，如果前面的数字大于后面的数字，就交换它们的顺序，否则的话，保持原来的顺序不变。这样将全体数字扫描一遍，最右端就得到了整个数列的最大值元素，再对前 n-1 个元素做同样的操作，得到第二大的元素。依次进行 n-1 次扫描，就得到了原数列按升序排列的结果。

****

冒泡法实现过程示例

第 1 轮比较：

(5 1 4 2 8) → (1 5 4 2 8)， 5 > 1，交换位置；

(1 5 4 2 8) → (1 4 5 2 8)， 5 > 4，交换位置

(1 4 5 2 8) →(1 4 2 5 8)， 5 > 2，交换位置

(1 4 2 5 8) → (1 4 2 5 8)， 5 < 8，不作交换

(1 4 2 5 8)

****

第 2 轮比较

(1 4 2 5 8) → (1 4 2 5 8)， 1 < 4，不作交换

(1 4 2 5 8) → (1 2 4 5 8)， 4 > 2，交换位置

(1 2 4 5 8) → (1 2 4 5 8)， 4 < 5，不作交换

（ 1 2 4 5 8 ）

****

第 3 轮比较

（ 1 2 4 5 8 ）→（ 1 2 4 5 8 ）

（ 1 2 4 5 8 ）→（ 1 2 4 5 8 ）

（ 1 2 4 5 8 ）

第 4 轮比较

（ 1 2 4 5 8 ）→（ 1 2 4 5 8 ）

（ 1 2 4 5 8 ）

****

冒泡法需要的比较运算的次数

$$(n-1)+(n-2)+\cdots+2+1=\frac{n(n-1)}{2}$$

计算机速度：10^{12} 次基本运算/秒

Google 搜索引擎的运行需要对至少 10^{10} 个数字进行排序：$\left(10^{10}\right)^2/2=\frac{10^{20}}{2}$

$\frac{10^{20}/2}{10^{12}}=\frac{10^8}{2}$ 秒，　1 年 $=365\times24\times3600=31536000\approx3.15\times10^7$ 秒

用冒泡法对 100 亿个网页的得分进行排名，每秒万亿次功能的计算机至少需要运行 1 年多的时间！

****

快速排序（quicksort）算法，英国学者霍尔 (C.A.R.Hoare)1962 年提出

问题：将 n 个互不相同的数按升序排列。

算法思想描述：

首先在n个数中，任选一个数，作为基准元素。将剩下的$n-1$个数与基准元素进行比较，大于基准元素的数放在右边，小于基准元素的数放在左边；

对左右两个子集再各选一个基准元素，比基准元素小的放左边，比其大的放右边，重复同样的操作；

继续在基准元素划分出的各部分选定新的基准元素，重复同样的操作；

直至所有元素排定位置

****

快速排序算法实现过程示例

3 7 8 4 5 2 1 9 6

3 7 8 4 5 2 1 9 6 → 3 2 1 4 7 8 5 9 6

3 2 1 4 7 8 5 9 6 → 1 3 2 4 7 5 6 8 9

1 3 2 4 7 5 6 8 9 → 1 2 3 4 5 6 7 8 9

****

第 1 步 *

第 2 步 *

第 3 步 *

....

第 m 步 *

$$1+2+2^2+\cdots+2^{m-1}=2^m-1, \qquad 2^m-1=n \;\Rightarrow\; m\approx\log_2 n$$

总比较次数：最理想的情况，不超过$n\log_2 n$；而最坏情况，$\dfrac{n(n-1)}{2}$

****

平均计算量分析结果

假设随机化快速排序算法对n个不同的数，$x_1,x_2,\cdots,x_n$，进行排序，每一次都是从所有可能的元素中独立且均匀地选取基准元素，那么所做比较次数的期望为$2n\ln n+O(n)$。

证明：设$y_1,y_2,\cdots,y_n$是$x_1,x_2,\cdots,x_n$ 按照升序排列后的结果，$y_1<y_2<\cdots<y_n$

对$1\le i<j\le n$，定义随机变量X_{ij}。

如果在算法过程中y_i与y_j进行了比较$X_{ij}=1$，否则$X_{ij}=0$。

那么总的比较次数$X=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}X_{ij}$，

****

$$E(X)=E\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}X_{ij}\right)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}E(X_{ij})。$$

考虑$\{y_i,y_{i+1},\cdots,y_j\}$，当且仅当$y_i$或$y_j$是集合$\{y_i,y_{i+1},\cdots,y_j\}$中第一个被选作基准的元素时，$y_i$与$y_j$进行比较。

所以$P(X_{ij}=1)=\dfrac{2}{j-i+1}$， $E(X_{ij})=\dfrac{2}{j-i+1}$。

$$E(X)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}E(X_{ij})=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{2}{j-i+1}\overset{\sum_{j=i+1}^{n}\frac{2}{j-i+1}=\sum_{k=2}^{n-i+1}\frac{2}{k}}{=}\sum_{i=1}^{n-1}\sum_{k=2}^{n-i+1}\frac{2}{k}$$

交换 i 和 k 求和顺序

$$=\sum_{k=2}^{n}\sum_{i=1}^{n-k+1}\frac{2}{k}=\sum_{k=2}^{n}\left(n-k+1\right)\frac{2}{k}=\left(n+1\right)\sum_{k=2}^{n}\frac{2}{k}-2\left(n-1\right)$$

$$=2n\sum_{k=1}^{n}\frac{1}{k}-2n-2\left(n-1\right)+2\sum_{k=2}^{n}\frac{1}{k}\overset{\sum_{k=1}^{n}\frac{1}{k}=\ln n+O\left(1\right)}{\sim}2n\ln n+O\left(n\right)。$$