Comparison between Radix and Merge Sorting Methods on Their Processing Efficiencies

YUKIO MIZUNO AND RYUJI KATO*

Abstract

The purpose of this paper is to describe and discuss the processing efficiencies in sorting operation. The typical methods adopted to large-volume sorting are radix sorting and merge sorting. The following note is to compare the processing efficiencies of these methods and to decide the most efficient sorting condition.

1. *The Fundamental Characteristics for Each Sorting Method*

To compare a certain number of methods for sorting, one must determine how long each method takes to solve a problem. In this paper, we discuss the total processing time including input and output.

The fundamental characteristics calculated for radix sorting are as follows:

$$\text{number of comparison} = N \log_m k_1 N,$$
$$\text{number of passes (input and output)} = N \log_m k_1 N,$$

where

m = number of tapes (pockets),
N = number of items,
r = number of digits,
m^n = number of keys,
$k_1 = m^r/N$.

For n-way merging, there are n input magnetic tape files and n output magnetic tape files. In general the average number of strings on the random sequence is $N/2$. Therefore, the number of passes by one item is described as

$$S = \log_n \frac{N}{2}.$$

(i) 2 way merge sorting: —

$$\text{number of comparison} = \frac{8}{3} N \log_2 \frac{N}{2},$$

This paper first appeared in Japanese in the Journal of Information Processing Society of Japan, Vol. 2, No. 6 (1961), pp. 326-332.

* Electronics Joint Division, Nippon Electric Co., Ltd., Kawasaki.

number passes $=N\log_2\frac{N}{2}$.

(Number of comparison for one item is 8/3.)

(ii) 3 way merge sorting: —

number of comparison $=3N\log_3\frac{N}{2}$,

number of passes $=N\log_3\frac{N}{2}$,

2. *Distribution Function of String*

We denote the length of string as follows.

Let $X_1, X_2\cdots X_l, Y_1, Y_2$ be the numbers of keys. If

$$Y_1>X_1<X_2<X_3<\cdots<X_l>Y_2$$

or

$$Y_1<X_1>X_2>X_3>\cdots>X_l<Y_2,$$

then the length of string is l.

If the data $X_1, X_2\cdots X_l, Y_1, Y_2$ are sampled from a population which has a probability density function $f(X)$, the distribution function of length of string $P(l)$ is described as follows.

$$P(l)=\frac{1}{A}\int_{-\infty}^{\infty}\int_{Y_1}^{\infty}\int_{-\infty}^{X_1}\int_{-\infty}^{X_2}\cdots\int_{-\infty}^{X_{l-1}}\int_{X_l}^{\infty}f(Y_1)f(X_1)f(X_2)\cdots f(X_l)f(Y_2)dY_2dX_l\cdots dX_3dX_2dX_1dY_1 \qquad (1)$$

where A is

$$A=\int_{-\infty}^{\infty}\int_{X_1}^{\infty}f(Y_1)f(X_1)dX_1dY_1=\frac{1}{2}.$$

Then

$$P(l)=\frac{2(l^2+l-1)}{(l+2)!}. \qquad (2)$$

As merging operations are repeated, this probability function of length of string will be altered.

After the n-th process of merging pass has finished, the distribution function of length of string $P_n(l)$ will be formed as follows.

$$P_n(l)=\sum_{i=1}^{l-1}P_{n-1}(i)P_{n-1}(l-i). \qquad (3)$$

The expectation of l and l^2 in regard to $P_n(l)$ are

$$\begin{aligned}E_n(l)&=\sum_{l=1}^{\infty}l\sum_{i=1}^{l-1}P_{n-1}(i)P_{n-1}(l-i)\\&=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i+j)P_{n-1}(i)P_{n-1}(j)\\&=2E_{n-1}(l),\end{aligned} \qquad (4a)$$

$$E_n(l^2)=\sum_{l=1}^{\infty} l^2 \sum_{i=1}^{l-1} P_{n-1}(i)P_{n-1}(l-i)$$
$$=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i+j)^2 P_{n-1}(i)P_{n-1}(j)$$
$$=2E_{n-1}(l^2)-2E_{n-1}^2(l)\,. \tag{4b}$$

The initial conditions of equations (4a) and (4b) are derived from equation (2) as follows.

$$E(l)=\sum_{l=1}^{\infty}\frac{2l(l^2+l^{-1})}{(l+2)!}=2\,,$$
$$E(l^2)=\sum_{l=1}^{\infty}\frac{2l^2(l^2+l-1)}{(l+2)!}=4e-6\,.$$

Thus we have

$$E_n(l)=2^{n+1}\,,$$
$$E_n(l^2)=2^n(4e-6)+2^{2(n+1)}\,.$$

The distribution of the sums of independent variables sampled from the same population is known to be approximated by a normal distribution owing to the central limit theorem. Therefore, for n large, the distribution of length of string $P_n(l)$ will be close to the normal distribution:

$$P_n(l)\fallingdotseq\frac{1}{\sqrt{2^{a+1}\pi(4e-10)}}e^{-\frac{(l-2^{n+1})^2}{2^n(4e-10)}}\,. \tag{5}$$

3. *Comparison of Sorting Methods*

To compare these methods we must define a measure of efficiency. One of the measures concerning with the efficiency of sorting will be the product of speed and cost.

Therefore, we define the product as W

$$W=(N_t t_t+N_c t_c)(am+b)\,, \tag{6}$$

N_t: number of passes, N_c: number of comparisons,
t_t: time to pass, t_c: time to compare,
m: essential number of magnetic tapes,
a: constant to indicate the cost coefficient of the computer,
b: constant to indicate the cost coefficient of the magnetic tape,

and use it as a measure of efficiency of sorting. These values are given for each method. For example

(i) $W_{m_2}=8/3t_c+t_t N\log_2\frac{N}{2}(4a+b)$ (2 way merge),

(ii) $W_{m_3}=(3t_c+t_t)N\log_3\frac{N}{2}(6a+b)$ (3 way merge),

(iii) $W_{d_m}=(t_c+t_m)N\log_n k_1N\{(m+1)a+b\}$ (radix sort)

Regarding m (in w-way merge and m radix sort) as a continuous variable x, we have

$$W_d(x)=t_cN(a+b)(1+x-xk_2)\log_x k_1N\,, \tag{7a}$$

$$W_m(x)=t_cN(a+b)(2x-(2x-1)k_2)\log_x\frac{N}{2}\,, \tag{7b}$$

$$k_2=b/(a+b)\,, \qquad k_3=t_c/t_t=0\,.$$

From the equations (7a) and (7b) we can decide the most efficient value of x by differentiating with x.

$$k_2=1-\frac{1}{x(\log x-1)} \tag{7a'}$$

$$k_2=1+\frac{1}{2x(1-\log x)-1} \tag{7b'}$$

Equations (7a') and (7b') show the optimal value of radix $x=m$ corresponding to k_2.

From the above results, we know that the optimal value of m monotonically increases according as k_2 increases; furthermore when $k_2=0$, the optimal value of m is 4 in radix sort and 3 in merge sort. And if the comparison time cannot be neglected, the condition for the optimal 2-way merge sort is $k_3>0.31$.

4. *Conclusion*

Thus, we can decide the optimal method of sorting from the given N, k_2 and k_3. However, some of more ideal sorting methods will be found. We have shown the results of comparison between one sorting method and others and did not take up the difference of a method from the ideal method. Therefore, in order to discuss the comparison with the ideal sorting method, it is necessary to analyze the information value on the way of sorting process, and to define the sorting method which has the least loss of information.