TR-85-003

Some Combinatorial Identities in Computer Sorting

by

Jun S. Huang

Institute of Information Science
Academia Sinica
Taipei, Taiwan, Republic of China
March 1985

中研院資訊所圖書室
3 0330 03 000049 6

0049

Abstract

Some combinatorial identities are presented here, where these identities are derived from the complex analysis of parallel sorting a large data set by computer. The parallel sorting requires statistical data partitioning and method of quick sort.

1. Introduction

Combinatorial identities are widely tabulated in mathematical, statistical and computational textbooks [1,2,3,4,] . These identities are important in many fields of research and application, especially those require exact forms of solutions or calculations. Here some combinatorial identities are derived through the analysis of parallel sorting a large data set by computer. The parallel sorting requires statistical data partitioning and method of quick sort [5] .

Frazer and Mckellar [6] have earlier derived a combinatorial equation in their sample sort analysis, where Knuth simplifies it and obtains an identity:

$$\sum_{k=m}^{n} \binom{k}{m} \frac{1}{n+1-k} = \binom{n+1}{m} \sum_{k=m+1}^{n+1} \frac{1}{k}$$

The proof is easy by induction but the identity is hard to find, due to the complexity of harmonic series. The proof is based on a popular identity:

$$\sum_{i=m}^{n} \binom{i}{m} = \binom{n+1}{m+1}$$

which will be used in the subsequent identities.

2. Some Identities from Parallel Sorting

Let the data set to be sorted parallely on n processors be denoted by X, and the size of X by N, where $N > n$. To partition X we first take a random sample of size $n\ell - 1$ (the choice of ℓ will be discussed later), and order this sample in ascending order to get order statistics :

$$Y_1 < Y_2 < \ldots < Y_\ell < \ldots < Y_{2\ell} < \ldots < Y_{(n-1)\ell} < \ldots < Y_{n\ell-1}$$

Secondly, we use $n - 1$ points $Y_\ell, Y_{2\ell}, \ldots, Y_{(n-1)\ell}$ as pivot nodes and form a balanced binary tree having these $n - 1$ nodes. At the bottom of this tree are n buckets. Each data is steered to its correct bucket as it descends the tree (see Figure 1). Thus from

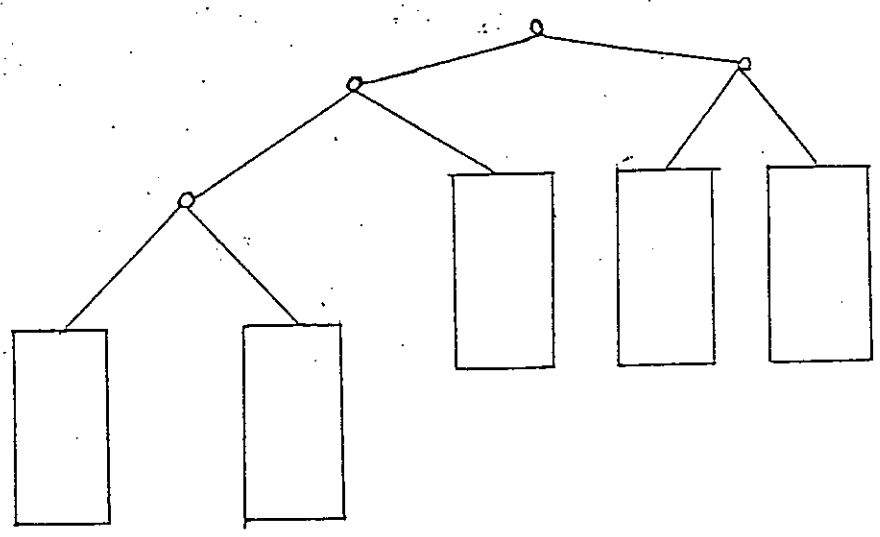

Figure 1 : Binary Tree with n = 5 Buckets.

this binary tree we are able to partition X into n components such that all data in the i^{th} component are less than each data in the $i + 1^{st}$ component, $i = 1, 2, \ldots, n - 1$. Let the i^{th} component be denoted by Q_i, $i = 1, 2, \ldots, n$. Then

$$Q_1 = \{x : x < Y_\ell\} ,$$

$$Q_i = \{x : Y_{(i-1)\ell} < x < Y_{i\ell}\} , \text{ for } 2 \le i \le n-1,$$

$$Q_n = \{x : Y_{(n-1)\ell} < x\} .$$

Now we can use the sample sort method proposed by Frazer and McKellar [6] to sort these n Q_i's on n processors simultaneously. To explain this more clearly, we note that there are $\ell - 1$ sample points between $Y_{(i-1)\ell}$ and $Y_{i\ell}$. These $\ell - 1$ sample points are again used as random sample taken from Q_i. Thus we can apply Frazer and McKellar's procedure to sort each Q_i on the i^{th} processor. Their procedure is a variation of Quick Sort. After parallel sorting, we can easily insert these n pivot nodes into Q_i and then concatenate all together with very little effort to obtain the full sorted data set X. The entire sorting consists of sampling and insertion of pivot points, parallel sorting on each processor, and the final concatenation of the sorted components.

Let $q_i(j)$ be the probability that $y_i = x_j$ where y_i is the i^{th} order statistic of the sample and x_j is the j^{th} elements of the sorted set of X. It is easy to see

$$q_i(j) = \binom{j-1}{i-1}\binom{N-j}{n\ell-i-1}\Big/\binom{N}{n\ell-1} .$$

Let $P_i(j)$ be the probability that the number of elements in Q_i is j. Then we have

LEMMA 1.

$$P_i(j) = \binom{N-j-1}{(n-1)\ell-1}\binom{j}{\ell-1}\Big/\binom{N}{n\ell-1} , \text{ for } j \geq \ell-1 .$$

This probability is independent of i = 1, 2, ..., n.

Proof. For i = 1, $P_1(j) = q_\ell(j+1) = \binom{j}{\ell-1}\binom{n-j-1}{(n-1)\ell-1}\Big/\binom{N}{n\ell-1}$.

For i=n, $P_n(j) = q_{(n-1)\ell}(N-j) = \binom{N-j-1}{(n-1)\ell-1}\binom{j}{\ell-1}\Big/\binom{N}{n\ell-1}$.

For $2 \leq i \leq n-1$,

$$P_i(j) = \sum_{t=(i-1)\ell}^{N-(n-i)\ell-j} q_{(i-1)\ell}(t)\, q_{i\ell}\,(t+j+1 \mid Y_{(i-1)\ell} = x_t)\,,$$

$$= \sum_{t=(i-1)\ell}^{N-(n-i)\ell-j} \frac{\binom{t-1}{(i-1)\ell-1}\binom{N-t}{(n-i+1)\ell-1}}{\binom{N}{n\ell-1}} \cdot \frac{\binom{j}{\ell-1}\binom{N-t-j-1}{(n-i+1)\ell-\ell-1}}{\binom{N-t}{(n-i+1)\ell-1}}$$

$$= \frac{\binom{N-j-1}{(n-1)\ell-1}\binom{j}{\ell-1}}{\binom{N}{n\ell-1}}\,,$$

where $q_{i\ell}(t+j+1 \mid y_{(i-1)\ell} = X_t)$ equals to the probability $q_\ell(j+1)$ for a sample of size $(n-i+1)\ell - 1$ from a set of size $N - t$. □.

From this lemma, we get the distribution function $P_i(j)$, $j = \ell - 1, \ell, \ldots, N - (n-1)\ell$. In fact this distribution is called the negative hypergeometric distribution. The mean of this distribution, or the mean size of Q_i is

$$E(j) = \frac{N - n + 1}{n}$$

and the variance of this distribution is

$$\mathrm{Var}(j) = \frac{(N - n\ell + 1)(n-1)}{(n\ell + 1)\, n}$$

Thus an approximate 95% confidence interval for the size of Q is

$$\frac{N+1}{n} - 1 \pm 3\sqrt{\frac{(N-n\ell+1)(n-1)}{(n\ell+1)\cdot n}} .$$

This holds for all i and also this formula sets an approximate lower limit of the size of core storage of each processor for fast processing without disk I/O delay.

Let $E(C_1)$ be the expected number of comparisons required to sort the sample of size $n\ell - 1$ by using the minimum storage Quicksort, then

$$E(C_1) = 2n\ell \sum_{1}^{n\ell-1} \frac{1}{i+1} - 2(n\ell - 1) \qquad (1)$$

Now we can treat $Y_{(i-1)\ell+1} < Y_{(i-1)\ell+2} < \cdots < Y_{i\ell-1}$ as $\ell - 1$ order statistics from a population of size j given that Q_i has size j. We can extend the sample sort proposed by Frazer and McKellar to sort Q_i. The expected number of comparisons required to sort Q_i given that Q_i has isze j, $j \geq \ell - 1$ is

$$E[C(Q_i \mid j)] = E(C_2) + E(C_3)$$

whree C_2 is the number of comparisons required to insert the sample, and C_3 is the number of comparisons to sort the segments of Q_i.

Similarly with Frazer and McKellar's analysis, it can be shown that

$$(j-\ell+1)\log_2 \ell \leq E(C_2) \leq (j-\ell+1)[0.0861 + \log_2 \ell] ,$$

$$\text{and } E(C_3) = 2(j+1) \sum_{i=1}^{j} \frac{1}{i+1} - 2(j-\ell+1).$$

Thus the expected number of comparisons required to sort Q_i is

$$E\,[\,C(Q_i)] = E\,E\,[\,C(Q_i \mid j)]$$

and therefore

$$E\,[\,C(Q_i)] = E\,E(C_2) + E\,E(C_3).$$

After further derivation we obtain

$$\frac{N-n\ell+1}{n} \log_2 \ell \le E\,E(C_2) < \frac{N-n\ell+1}{n}\,[\,0.0861 + \log_2 \ell] \qquad (2)$$

and

$$E\,E(C_3) = \sum_{j=\ell-1}^{N-(n-1)\ell} \frac{\binom{N-j-1}{(n-1)\ell-1}\binom{j}{\ell-1}\left[2(j+1)\sum_{i=\ell}^{j}\frac{1}{i+1}\right]}{\binom{N}{n\ell-1}}$$

$$- \sum_{j=\ell-1}^{N-(n-1)\ell} \frac{\binom{N-j-1}{(n-1)\ell-1}\binom{j}{\ell-1}}{\binom{N}{n\ell-1}} \cdot 2(j-\ell+1).$$

To simplify the calculation we need the following genius identity due to Knuth.

LEMMA 2 (Knuth).

$$\sum_{j=\ell}^{N-a} \binom{N-j-1}{a-1}\binom{j}{\ell-1}\Big[\,(j+1)\sum_{i=\ell}^{j}\frac{1}{i+1}\,\Big] = \ell\binom{N+1}{a+\ell}\sum_{a+\ell}^{N}\frac{1}{i+1}\,.$$

PROOF. $\sum_{j=\ell}^{N-a} \binom{N-j-1}{a-1}\binom{j}{\ell-1}\Big[\,(j+1)\sum_{\ell}^{j}\frac{1}{i+1}\,\Big]$

$$= \ell\sum_{\ell}^{N-a}\binom{N-j-1}{a-1}\binom{j+1}{\ell}(H_{j+1} - H_\ell) \text{ where } H_j = \sum_1^j \frac{1}{i},$$

$$= \ell\sum_{0}^{N}\binom{N-j}{a-1}\binom{j}{\ell}(H_j - H_\ell).$$

Now $\sum_0^\infty \binom{k}{a-1} z^k = \frac{z^{a-1}}{(1-z)^a}$ and

$$\sum_0^\infty \binom{j}{\ell}(H_j - H_\ell)z^j = \frac{z^\ell}{(1-z)^{\ell+1}}\log\left(\frac{1}{1-z}\right).$$

Multiply these two power series together and look at the coefficient of Z^N;

$$\frac{z^{a+\ell-1}}{(1-z)^{a+\ell+1}}\log\left(\frac{1}{1-z}\right) = \sum_0^\infty \binom{N+1}{a+\ell}(H_{N+1} - H_{a+\ell})Z^N$$

and hence the given sum is $\ell\binom{N+1}{a+\ell}\sum_{a+\ell}^{N}\frac{1}{i+1}$. □.

By putting $a = (n-1)\ell$ in Lemma 2, we have

$$EE(C_3) = 2\left[\frac{N\ell-N-1}{n} + \ell\frac{\binom{N+1}{n\ell}}{\binom{N}{N\ell-1}}\sum_{n\ell}^{N}\frac{1}{i+1}\right],$$

$$= 2\left[\ell + \frac{N+1}{n}\left(-1 + \sum_{n\ell}^{N}\frac{1}{i+1}\right)\right]. \qquad (3)$$

Since $\sum_{n\ell}^{N}\frac{1}{i+1} \le \log(N/(n\ell-1)) - 1/n\ell + 2/(N+1)$

$$EE(C_3) \le 2\left[\ell + \frac{2}{n} + \frac{N+1}{n}\left(-1 - \frac{1}{n\ell} + \log\left(\frac{N}{n\ell-1}\right)\right)\right].$$

From the above results we have

<u>THEOREM 1.</u> The expected number of comparisons (or computing time), E(C), on processing Q_i is given by the sum of Eqs.(1),(2),(3), which is

$$2n\ell\sum_{1}^{n\ell-1}\frac{1}{i+1} + \frac{N+1}{n}\left(\log_2\ell - 2 + 2\sum_{n\ell}^{N}\frac{1}{i+1}\right) - \ell\log_2\ell + 2(\ell-n\ell+1)$$

$$\le E(C)$$

$$\le 2n\ell\sum_{1}^{n\ell-1}\frac{1}{i+1} + \frac{N+1}{n}\left(\log_2\ell - 1.9139 + 2\sum_{n\ell}^{N}\frac{1}{i+1}\right) - \ell\log_2\ell + 1.9139\ell$$

$$+ 2(1-n\ell),$$

$$\le \frac{N+1}{n}\left(2\log\frac{N}{n\ell-1} + \log_2\ell - 1.9139 - \frac{2}{n\ell}\right) + 2n\ell\log(n\ell-1) - \ell\log_2\ell$$

$$+ 1.9139\ell + \frac{4}{n} + 6.$$

From the discussion of negative hypergeometric distribution [2] it is easy to see that

$$\sum_{n=k}^{N-M+K} \binom{n-1}{k-1} \binom{N-n}{M-k} = \binom{N}{M} \quad \text{for } N \geqslant M \geqslant k.$$

The left hand side of this identity is a complete sum. But to consider the partial sum properties, the general approach is by normal approximation and in most delicated cases no exact formulas can be found. However the following identity can be useful in many fields:

Lemma 3 (Huang). For $i \geqslant b$

$$\sum_{t=i+1}^{M-a+1} \binom{M-t}{a-1} \binom{t}{b} = \sum_{t=0}^{b} \binom{i+1}{b-t} \binom{M-i}{a+t} .$$

Proof. By induction on b. When b=0, $\sum_{t=i+1}^{M-a+1} \binom{M-t}{a-1} = \binom{a-1}{a-1}$

$+ \binom{a}{a-1} + \ldots\ldots + \binom{M-i-1}{a-1} = \binom{M-i}{a}$.

Assume the identity is true for b; we try to prove it true for b+1, and in this case the left hand side of the identity is equal to $\sum_{j=i}^{M-a} \binom{M-j-1}{a-1} \binom{j+1}{b+1} = \sum_{j=i}^{M-a} \binom{M-j-1}{a-1} \left[\sum_{k=b}^{j} \binom{k}{b} \right]$,

$$= \sum_{k=b}^{i} \sum_{j=i}^{M-a} \binom{M-j-1}{a-1} \binom{k}{b} + \sum_{k=i+1}^{M-a} \sum_{j=k}^{M-a} \binom{M-j-1}{a-1} \binom{k}{b},$$

$$= \sum_{k=b}^{i} \binom{M-i}{a} \binom{k}{b} + \sum_{k=i+1}^{M-a} \binom{M-k}{a} \binom{k}{b},$$

$$= \binom{M-i}{a} \binom{i+1}{b+1} + \sum_{k=0}^{b} \binom{i+1}{b-k} \binom{M-i}{a+1+k},$$

$$= \sum_{k=0}^{b+1} \binom{i+1}{b+1-k} \binom{M-i}{a+k} . \square .$$

From this Lemma and Lemma 2 we have the following identity:

$$\sum_{i=b}^{M-a} \frac{1}{i+1} \sum_{t=0}^{b} \binom{i+1}{b-t} \binom{M-i}{a+t} = \binom{M+1}{a+b} \sum_{a+b}^{M} \frac{1}{i+1}.$$

REFERENCES

1. Y. V. Geronimus and Y. Tseytlin (1965), Table of Integrals, Series and Products, Academic Press.
2. W. Feller (1970), An Introduction to Probabilities and Its Applications, Vol. 1 and 2, Wiley and Sons Inc.
3. D. Knuth (1980), The Art of Computer Programming, Vol. 1,2,3, Addison-Wesley Co., Reading, MA.
4. C. L. Liu (1968), Introduction to Combinatorial Mathematics, McGraw-Hill Inc., New York.
5. J. S. Huang and Y. C. Chow (1983), Parallel sorting and data partitioning by sampling, Proceeding of Comp SAC, Chicago, U. S. A..
6. W. Frazer and A. Mckellar (1970), Sample sort: A sampling approach to minimal tree sorting, J. ACM, 17, 496-507.