Vol. 34 No. 1 Transactions of Information Processing Society of Japan Jan. 1993

Regular Paper

An Alternative Analysis of the Algorithm for Separate Chaining Technique of the Hashing Method

RYOZO NAKAMURA†

A mathematical analysis is proposed to evaluate exactly the number of probes for the separate chaining technique of the hashing method in consideration of the frequency of access on each key. The proposed analysis is compared to the traditional one, so the reasons why the evaluation formulae are different even in the uniform probing case are discussed. It is then shown that the proposed analysis makes it possible to evaluate accurately the number of probes of the separate chaining technique.

1. Introduction

The hashing method is a frequently used technique for storing and retrieving records maintained in the form of a table. The key of each record is mapped by a hashing function to a sequence of table locations, and records are inserted and retrieved by following this sequence. In this case, there will probably be two or more keys which hash to one identical location. Such an occurrence is called a collision; there are several techniques for collision handling. These techniques are chiefly classified into two categories: chaining technique and open addressing technique.[2)]

The time required to solve a problem is one of the most important measures in evaluating an algorithm. This time is called the search cost in this case. The search cost is defined as the product of the number of probes and the frequency of access on a key. When frequency of access is uniform, the search cost of the separate chaining technique is independent of the order of inserting a key. However, when frequency of access on a key is not uniform, the inserting order plays an essential role.

Search cost considering frequency of access on keys has never been analyzed, so it is impossible to evaluate the search cost in accordance with the reality of searching. It is necessary to derive the evaluation formulae of search cost considering the frequency of access on an individual key.

In this paper, the algorithm of search cost for the separate chaining technique is analyzed mathematically according to a model that considers the frequency of access on keys. The evaluation formulae of the search cost are then derived with the concrete probability distribution of the number of probes. The proposed analysis is compared to the traditional one, and it is then shown that the proposed analysis is appropriate for evaluating correctly the search cost in conformity with the real behavior, and for offering the accurate and systematic analysis in the evaluation of search cost.

2. Basic Concepts of the Separate Chaining Technique

The hashing method is a technique for storing and retrieving records maintained in the form of a table. There is a special field in each record, called the key, that uniquely identifies it. The key is mapped by a hashing function to a sequence of table locations (indices), and the records are inserted and searched for by following these sequences.[2)-4)]

In a basic data structure for the separate chaining technique, the hash table called the bucket table is indexed by the bucket numbers 0, 1, 2, $\cdots$, $M-1$ according to their positions on the table of size M, and contains the headers for M linked lists. The elements of the i-th list are the class of key x in the set such that the value

† Department of Electrical Engineering and Computer Science, Faculty of Engineering, Kumamoto University

of its hashing function $h(x)$ is equal to i, that is, the hash address of x is i, $(i=0, 1, \cdots, M-1)$.

We assume that the N keys are uniformly mapped into the hash table of size M by a hashing function, and each of the M^N possible hash sequences

$$a_1, a_2, \cdots, a_N, \quad 0 \leq a_j < M$$

is equally likely, where a_j denotes the initial hash address of the j-th key to be inserted into the table.

Let p_{Nk} be the probability that the number of keys on any list is equal to k, $(k=0, 1, \cdots, N)$. There are $\binom{N}{k}$ ways to choose the set of j such that k keys among a_j, $(j=1, \cdots, N)$ have an identical value and $(M-1)^{N-k}$ ways to assign value to the other a's. Therefore, p_{Nk} is the binominal probability as follows,[1]

$$p_{Nk} = \binom{N}{k}(M-1)^{N-k}/M^N = \binom{N}{k}\left(\frac{1}{M}\right)^k\left(1-\frac{1}{M}\right)^{N-k} \quad (1)$$

and its generating function $P(z)$ is given by

$$P(z) = \sum_{k=0}^{N}\binom{N}{k}\left(\frac{z}{M}\right)^k\left(\frac{M-1}{M}\right)^{N-k} = \left(1+\frac{z-1}{M}\right)^N \quad (2)$$

The above generating function is used to calculate the average and the variance of the search cost.

We shall attempt to derive the evaluation formulae of the search cost. Let the average and the variance of the search cost be denoted by S_N and V_N in the successful search, and by $\bar{S}_N$ and $\bar{V}_N$ in the unsuccessful search, respectively. In the remainder of this paper, α denotes the loading factor N/M.

3. Traditional Analysis

For the separate chaining technique, the traditional analysis has been derived by Knuth[2] in conformity to the assumption that all keys are uniformly accessed, namely, a uniform probing is presumed. The analysis results have been referred to by many articles, even in the analysis of dynamic hashing.[5]

In the above mentioned analysis, N keys are equally scattered over M lists, and k_i denotes the number of keys on i-th list, that is, k_1 is the length of the first list, k_2 is the length of the second list and so on. Then, the number of ways to distribute N keys to M lists is given by $\binom{N}{k_1, \cdots, k_M}$.

Searching for a desired key requires the scanning of the list from the beginning. The total number of probes to find all keys of a list of length k_i is $\binom{k_i+1}{2}$.

Therefore, the average and the variance of the search cost have been derived for the uniform probing case as follows.

1) A successful searching

$$S_N = \sum_{k_1+\cdots+k_M=N}\left\{\frac{\binom{k_1+1}{2}+\cdots+\binom{k_M+1}{2}}{N}\right\}\frac{\binom{N}{k_1, \cdots, k_M}}{M^N}$$

$$= \sum_{k=0}^{N} M\frac{\binom{k+1}{2}}{N}\frac{\binom{N}{k}(M-1)^{N-k}}{M^N} = \frac{M}{N}\sum_{k=1}^{N}\binom{k+1}{2}\binom{N}{k}\left(\frac{1}{M}\right)^k\left(1-\frac{1}{M}\right)^{N-k}$$

$$= \frac{M}{N}\sum_{k=1}^{N}\binom{k+1}{2}p_{Nk} = \frac{M}{N}\left\{p_{N1}+3p_{N2}+\cdots+\frac{N(N+1)}{2}p_{NN}\right\}$$

$$= \frac{M}{N}\left\{\frac{1}{2}P''(1)+P'(1)\right\} = \frac{1}{2}\frac{N-1}{M}+1 \doteqdot 1+\frac{\alpha}{2} \quad (3)$$

$$V_N = \frac{\sum_{k_1+\cdots+k_M=N}\binom{N}{k_1, \cdots, k_M}}{M^N}\left\{\frac{\binom{k_1}{2}+\binom{k_2}{2}+\cdots+\binom{k_M}{2}}{N}\right\}^2 - \left\{\frac{N-1}{2M}\right\}^2$$

$$= \frac{1}{M^N N^2}\left\{M(M-1)\Sigma\binom{N}{k_1, k_2, \cdots, k_M}\times\binom{k_1}{2}\binom{k_2}{2}+M\Sigma\binom{N}{k_1, \cdots, k_M}\binom{k_1}{2}^2\right\} - \left\{\frac{N-1}{2M}\right\}^2$$

$$= \frac{1}{M^N N^2}\left\{M(M-1)\times\left(\frac{1}{4}M^{N-4}N^{\underline{4}}\right)+M(M^{N-4})\times\left(\frac{1}{4}N^{\underline{4}}+N^{\underline{3}}M\right)\right.$$

$$+\frac{1}{2}N^2M^2\Big)\Big\}-\Big\{\frac{N-1}{2M}\Big\}^2$$
$$=\frac{(M-1)(N-1)}{2NM^2} \qquad (4)$$

where $N^{\underline{k}}$ is $\prod_{0\le j<k}(N-j)$.

2) An unsuccessful searching

In this case, either a list having no key is searched or all of the keys linked on a list are searched. The former searching is counted as one probe. The average and the variance of the number of probes are given as follows;

$$\bar{S}_N=\sum_{k=0}^{N}(k+\delta_k)p_{Nk}$$
$$=P'(1)+P(0)$$
$$=\frac{N}{M}+\Big(1-\frac{1}{M}\Big)^N$$
$$\fallingdotseq\alpha+\exp(-\alpha) \qquad (5)$$
$$\bar{V}_N=\sum_{k=0}^{N}(k+\delta_k)^2p_{Nk}-S_N^2$$
$$=p_{N0}+\sum_{k=1}^{N}k^2p_{Nk}-S_N^2$$
$$=P(0)+P'(1)+P''(1)-\{P'(1)+P(0)\}^2$$
$$=\frac{N(M-1)}{M^2}+\Big(1-\frac{1}{M}\Big)^N\Big\{1-\frac{2N}{M}$$
$$-\Big(1-\frac{1}{M}\Big)^N\Big\}$$
$$\fallingdotseq\alpha+\exp(-\alpha)(1-2\alpha-\exp(-\alpha)) \qquad (6)$$

where $\delta_k=\begin{cases}1, & (k=0)\\ 0, & (k>0)\end{cases}$

4. Proposed Analysis

The traditional analysis is unable to evaluate the search cost even if the probability of frequency of access on a key is given. An analysis is therefore proposed in a more general situation considering the frequency of access on an individual key. In this case, the inserting order of a key plays an essential role.

For the analysis, we must clarify the relation between the inserting order of a key and its locating position in a list for constructing a more exact model. First let S_{ijk} be the probability that the i-th key inserted will be located in the j-th position from the head of a list having k keys. Assuming that keys are inserted successively at the head of a list, there are $\binom{N-i}{j-1}$ possible combinations to distribute the $N-i$ keys into the front $j-1$ positions of a list with k keys, and similarly $\binom{i-1}{k-j}$ ways to distribute the $i-1$ keys into the rear $k-j$ positions of the list. Therefore the probability S_{ijk} can be expressed as follows.

$$S_{ijk}=\binom{N-i}{j-1}\binom{i-1}{k-j}\Big/\sum_{j=1}^{k}\binom{N-i}{j-1}\binom{i-1}{k-j}$$
$$=\binom{N-i}{j-1}\binom{i-1}{k-j}\Big/\binom{N-1}{k-1} \qquad (7)$$

On the other hand, assuming that keys are inserted at the tail of a list, S_{ijk} is given by,

$$S_{ijk}=\binom{i-1}{j-1}\binom{N-i}{k-j}\Big/\binom{N-1}{k-1}. \qquad (8)$$

Here we have

$$\sum_{j=1}^{k}S_{ijk}=1.$$

N keys will be inserted in random order by the insertion algorithm, so the probability that any key will be located in the j-th position from the head of a list with k keys is represented as follows. In case that keys are inserted at the head of a list, the probability is given by

$$\frac{1}{N}\sum_{i=1}^{N}S_{ijk}=\frac{1}{N}\sum_{i=1}^{N}\binom{N-i}{j-1}\binom{i-1}{k-j}\Big/\binom{N-1}{k-1}$$
$$=\frac{1}{N}\binom{N}{k}\Big/\binom{N-1}{k-1}=\frac{1}{k},$$

and in the case of inserting at the tail of a list, that probability also becomes $1/k$. These results suggest that the probability of each of the keys being located in any position of a list is equal whether keys are inserted at the head or at the tail of a list.

In the analysis of the search algorithm, let ρ_i be the probability that the i-th key inserted will be retrieved, and let γ_{kj} be the probability that the j-th key from the head of a list having k keys is probed. The probability γ_{kj} can be expressed using the probability S_{ijk},

$$\gamma_{kj}=\sum_{i=1}^{N}S_{ijk}\rho_i \qquad (9)$$

and $\sum_{j=0}^{k}\gamma_{kj}=1$.

Here $\sum_{j=0}^{k}\gamma_{kj}=1$ is proved as follows. First $\gamma_{00}=1$ and $\gamma_{k0}=0$ for $k>0$.

$$\sum_{j=1}^{k}\gamma_{kj}=\sum_{j=1}^{k}\sum_{i=1}^{N}\binom{N-i}{j-1}\binom{i-1}{k-j}\rho_i\Big/\binom{N-1}{k-1}$$
$$=\Big\{\rho_1\sum_{j=1}^{k}\binom{N-1}{j-1}\binom{0}{k-j}$$
$$+\rho_2\sum_{j=1}^{k}\binom{N-2}{j-1}\binom{1}{k-j}+\cdots$$

$$+\rho_N\sum_{j=1}^{k}\binom{0}{j-1}\binom{N-1}{k-j}\Big\}\Big/\binom{N-1}{k-1}$$
$$=\Big\{\rho_1\sum_{j=0}^{k-1}\binom{N-1}{j}\binom{0}{k-1-j}+\rho_2\sum_{j=0}^{k-1}\binom{N-2}{j}\binom{1}{k-1-j}+\cdots+\rho_N\sum_{j=0}^{k-1}\binom{0}{j}\binom{N-1}{k-1-j}\Big\}\Big/\binom{N-1}{k-1}$$
$$=\Big\{\rho_1\binom{N-1}{k-1}+\rho_2\binom{N-1}{k-1}+\cdots+\rho_N\binom{N-1}{k-1}\Big\}\Big/\binom{N-1}{k-1}$$
$$=\rho_1+\rho_2+\cdots+\rho_N$$
$$=1$$

Let q_{Nk} be the probability that the k-th key from the head of a list of any length is probed. q_{Nk} is equal to the probability that the k-th key from the head of a list of length k or longer is probed. Hence, using the probability previously introduced, the probability q_{Nk} is given by

$$q_{Nk}=\sum_{j=k}^{N}\gamma_{jk}p_{Nj}. \tag{10}$$

Here

$$\sum_{k=0}^{N}q_{Nk}=\sum_{k=0}^{N}\sum_{j=k}^{N}\gamma_{jk}p_{Nj}=\sum_{k=0}^{N}\sum_{j=0}^{k}\gamma_{kj}p_{Nk}=\sum_{k=0}^{N}p_{Nk}=1.$$

Finally, we conceive the number of probes as a random variable and its probability distribution in order to construct the evaluation formulae for the search cost; we also attempt to derive the average and the variance of the search cost for both successful and unsuccessful search.

1) A successful search

In this case, there must exist at least one key in a list for successful searching. The evaluation formulae are thus represented as follows.

$$S_N=\sum_{k=1}^{N}kq_{Nk}\Big/\sum_{k=1}^{N}p_{Nk}=\sum_{k=1}^{N}k\sum_{j=k}^{N}\gamma_{jk}p_{Nj}\Big/\sum_{k=1}^{N}p_{Nk} \tag{11}$$

$$V_N=\sum_{k=1}^{N}k^2\sum_{j=k}^{N}\gamma_{jk}p_{Nj}\Big/\sum_{k=1}^{N}p_{Nk}-S_N^2 \tag{12}$$

2) An unsuccessful search

In this case, either a list having no key is searched or all of the keys linked in the list are searched. The following evaluation formulae are the same as the traditional ones.

$$\bar{S}_N=\sum_{k=0}^{N}(k+\delta_k)p_{Nk} \tag{13}$$

$$\bar{V}_N=\sum_{k=0}^{N}(k+\delta_k)^2p_{Nk}-S_N^2 \tag{14}$$

5. Comparison of Both Analyses

In this section, the proposed analysis is compared to the traditional one under the assumption that the probability of frequency of access on each key is equally likely. We only discuss the successful searching case, since both of the evaluation formulae of an unsuccessful searching are identical.

From the above assumption, ρ_i becomes $1/N$, $(i=1, 2, \cdots, N)$ independently, and then the probability γ_{jk} of equations (11), (12) becomes as follows,

$$\gamma_{jk}=\sum_{i=1}^{N}S_{ikj}\rho_j=\frac{1}{N}\sum_{i=1}^{N}\binom{N-i}{k-1}\binom{i-1}{j-k}\Big/\binom{N-1}{j-1}=\frac{1}{N}\binom{N}{j}\Big/\binom{N-1}{j-1}=\frac{1}{j}$$

Thus, we can derive the following evaluation formulae of S_N and V_N.

$$S_N=\sum_{k=1}^{N}k\sum_{j=k}^{N}\frac{1}{j}p_{Nj}\Big/\sum_{k=1}^{N}p_{Nk}=\frac{1}{2}\{P'(1)+P(1)-P(0)\}/\{1-P(0)\}=\frac{1}{2}\Big[\frac{N}{M}\Big/\Big\{1-\Big(1-\frac{1}{M}\Big)^N\Big\}+1\Big]\doteqdot\frac{1}{2}\Big\{\frac{\alpha}{1-\exp(-\alpha)}+1\Big\} \tag{15}$$

$$V_N=\sum_{k=1}^{N}k^2\sum_{j=k}^{N}\frac{1}{j}p_{Nj}\Big/\sum_{k=1}^{N}p_{Nk}-S_N^2=\frac{1}{\sum_{k=1}^{N}p_{Nk}}\Big\{p_{N1}+\frac{15}{6}p_{N2}+\cdots+\frac{(N+1)(2N+1)}{6}p_{NN}\Big\}-S_N^2$$
$$=\frac{1}{6\{1-P(0)\}}\{1-P(0)+5P'(1)+2P''(1)\}-\Big\{\frac{1}{2\{1-P(0)\}}\{P'(1)+1-P(0)\}\Big\}^2$$
$$=\Big\{\frac{N}{M}\frac{N-1}{M}+\frac{N}{M}\Big\}\Big/3\Big\{1-\Big(1-\frac{1}{M}\Big)^N\Big\}-\Big(\frac{N}{M}\Big)^2\Big/4\Big\{1-\Big(1-\frac{1}{M}\Big)^N\Big\}^2-\frac{1}{12}$$

$$\fallingdotseq \frac{\alpha(\alpha+1)}{3(1-\exp(-\alpha))} - \frac{\alpha^2}{4(1-\exp(-\alpha))^2} - \frac{1}{12} \qquad (16)$$

The above evaluation formulae (15), (16) of the average and the variance of the search cost are represented concisely by the function of the load factor α alone, but differ from the traditional equations (3) and (4). We shall discuss the reasons why these evaluation formulae are different.

First, in the proposed analysis the number of probes itself is regarded as a random variable and its probability distribution is derived concretely. On the other hand, in the traditional analysis described in section 3, N keys are equally scattered over M lists, and k_i denotes the number of keys on i-th list. The random variable expressed by $\left\{\binom{k_1+1}{2} + \cdots + \binom{k_M+1}{2}\right\}/N$ denotes the number of probes per key, that is, counting the total number of probes to find all keys on M lists and then dividing that by N keys. Here its probability distribution $\binom{N}{k_1, \cdots, k_M}/M^N$ is the same as the probability of Maxwell-Boltzmann statistics.[6]

From the above arguments, the difference is caused by considering what the random variable and its probability distribution are. Besides this, it is important to emphasize that only lists having more than one key in a list are probed for successful searching in the proposed analysis but this matter is not regarded in the traditional analysis.

As a result, in particular the variance (4) derived by the traditional analysis can not be expressed in terms of the load factor α alone. As the table size $M \gg 1$ and the number of keys $N \gg 1$, the variance becomes approximately $1/2M$, that is, it is a function of the only M size table and does not depend on the number of keys N, so that the formula (4) can not evaluate correctly the actual behavior.

6. Numerical Tests

The proposed evaluation formulae (11), (12) can evaluate the search cost in accordance with any probability distribution of the frequency of access on a key. In the empirical tests, let us assume the following three probability distributions of the frequency of access on a key.

i) The probability of the frequency of access on each key is equally likely, called "Uniform," the probability ρ_i holds the relation $\rho_i = 1/N$, $(i=1, 2, \cdots, N)$.
The proposed evaluation formulae (15), (16) are compared to the traditional ones (3), (4).

ii) The probability of the frequency of access on a key is reduced in half according to the order of inserting a key, called "Binary," the probability ρ_i holds the relation $\rho_i = c/2^{i-1}$, $(i=1, 2, \cdots, N)$, where $c = 1/(2-2^{1-N})$.

iii) The probability of the frequency of access on a key is reduced harmonically according to the order of inserting a key, typically called "Zipf's law,"[2], the probability ρ_i holds the relation $\rho_i = c/i$, $(i=1, 2, \cdots, N)$ where $c = 1/H_N$ and H_N is the harmonic number, $H_N = \sum_{k=1}^{N} 1/k$.

The numerical results with all three probability distributions are shown in **Table 1**, provided that keys are inserted successively at the head of a list.

The numerical behavior mentioned above shows that it is possible to evaluate the search cost appropriately and consistently in accordance with the frequency of access on a key by the proposed analysis.

7. Conclusions

In this paper, we have analyzed mathematically the evaluation formulae of the average number of probes and the variance of search cost in consideration of the frequency of access on an individual key for the separate chaining method. The proposed evaluation formulae have been compared to the traditional ones assuming uniform probing. As a result, the difference between them has been indicated, and it has been shown that this difference is caused by the way in which a random variable and its probability distribution are defined.

Finally, it has been shown that the proposed analysis makes it possible to evaluate exactly the performance of the separate chaining technique, and it also suggests how to analyze the algo-

Table1 Comparisons of the search cost in a successful search under the consideration of the probability distributions. (table size $M=50$)

Number of key (N)	Load factor (α)	Traditional analysis		Proposed analysis					
		i) Uniform		i) Uniform		ii) Binary		iii) Zipf's law	
		S_N	V_N	S_N	V_N	S_N	V_N	S_N	V_N
		(3)*	(4)	(15)	(16)	(11)	(12)	(11)	(12)
25	0.5	1.25	0.0094	1.14	0.15	1.26	0.26	1.20	0.21
50	1.0	1.50	0.0096	1.29	0.35	1.59	0.63	1.45	0.52
100	2.0	2.00	0.0097	1.66	0.89	2.29	1.54	2.06	1.35
150	3.0	2.50	0.0097	2.08	1.63	3.13	2.59	2.78	2.39
200	4.0	3.00	0.0097	2.54	2.56	4.06	3.68	3.56	3.57
250	5.0	3.50	0.0097	3.02	3.65	5.02	4.75	4.38	4.85

*(k) denotes the evalution formula (k).

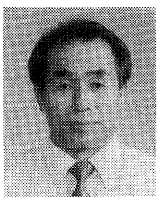
rithm taking account of the intrinsic nature of the problem.

Acknowledgements The author would like to thank the anonymous referees for their helpful comments in improving the clarity of this paper.

References

1) Knuth, D. E.: *The Art of Computer Programming,* Vol. 1, *Fundamental Algorithms,* pp. 51-73, Addison-Wesley, Reading, Mass. (1973).
2) Knuth, D. E.: *The Art of Computer Programming,* Vol. 3, *Sorting and Searching,* pp. 506-549, Addison-Wesley, Reading, Mass. (1973).
3) Wirth, N.: *Algorithms+Data Structures=Programs,* pp. 264-279, Prentice-Hall, Inc, Englewood Cliffs, New Jersey (1976).
4) Aho, A. V., Hopcroft, J. E. and Ullman, J. D.: *Data Structures and Algorithms,* pp. 122-134, Addison-Wesley, Reading, Mass. (1987).
5) Larson, P.-A.: Dynamic Hash Table, *Comm. ACM,* Vol. 31, No. 4, pp. 446-457 (1988).
6) Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. 1, pp. 39-40, John Wiley & Sons, New York (1957).

(Received October 17, 1991)
(Accepted October 21, 1992)



Ryozo Nakamura (Member)

Ryozo Nakamura received the M. E. degree from Kumamoto University in 1968 and the D. E. degree in computer science from Kyushu University in 1985. From 1968 to 1974, he joined Chubu Electric Power Company. Since 1975 he has joined in Faculty of Engineering of Kumamoto University, and is presently a professor in Department of Electrical Engineering and Computer Science. His current research interests include the design and analysis of algorithms and data structures.