PUC

Série: Monografias em Ciência da Computação

Nº 1/77

(antiga/formerly: Monographs in Computer Science and Computer Applications)

AVERAGE LOWER BOUNDS FOR OPEN-ADRESSING HASH CODING

by

Gaston Gonnet

Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro

Rua Marquês de São Vicente 225 — ZC 19

Rio de Janeiro — Brasil

Série: Monografias em Ciência da Computação

Nº 1/77

(antiga/formerly: Monographs in Computer Science and Computer Applications)

AVERAGE LOWER BOUNDS FOR OPEN-ADDRESSING HASH - CODING*

by

Gaston H. Gonnet

Editor: Michael F. Challis February, 1977

* Work partially sponsored by FINEP.

For copies contact:

Rosane T. L. Castilho
Head, Setor de Docuementação e Informação
Depto. de Informática - PUC/RJ
Rua Maqrues de São Vicente, 209 - Gávea
20.000 - Rio de Janeiro - RJ - Brasil

INFORMATION TO OUR READERS

We have decided to change the title of our series "Monographs in Computer Science and Computer Applications" to "Monografias em Ciência da Computação", to take effect from the first issue of 1977.

The aim and scope have been maintained, but besides including works in English we will also include works in Portuguese. As we are aware that problems concerning language barrier can affect our purpose of communicating our findings to our foreign audience, whenever we publish a work in Portuguese an abstract in English will be included.

Thank you for your interest concerning our publications during the last nine years.

DEPARTAMENTO DE INFORMÁTICA
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

INFORMAÇÕES AOS NOSSOS LEITORES

A nossa série "Monographs in Computer Science and Computer Applications" teve o seu título mudado para " Monografias em Ciência da Computação ", a partir do primeiro fascículo de 1977.

A série de " Relatórios Técnicos " foi absorvida pela série de "Monografias...", tendo sido o seu último fascículo, RT-01-77, publicado em janeiro deste ano. Será, portanto, através da série "Monografias em Ciência da Computação" que o Departamento publicará, em inglês ou português , todos os seus trabalhos de pesquisa de interesse à comunidade técnico-científica.

Agradecemos a todos pelo interesse demonstrado por nossas publicações durante os últimos nove anos.

DEPARTAMENTO DE INFORMÁTICA
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RESUMO:

Neste trabalho são derivados limites inferiores do número médio de acessos e do número máximo médio de acessos para a pesquisa em tabelas organizadas com "open-addressing". Acha-se que a média do máximo número de acessos para tabelas completas é $\ell n(n) + O(1)$, e quando $n \to \infty$ é $\lceil - \alpha^{-1} \ell n(1-\alpha) \rceil$ onde α é o fator de carga. Acha-se que o limite inferior da média de acessos é 1.668... Resultados da simulação indicaram que os limites do máximo são muito precisos.

PALAVRAS CHAVES:

Limites inferiores, pesquisa, número médio de acessos, hashing, open-addressing, número de Stirling de 2ª. ordem, minimax.

ABSTRACT:

In this paper we derive average lower bounds on the, worst case and average, number of accesses for open-addressing hashing tables. The average worst case for full tables is found to be $\ell n(n) + O(\ell)$ while for tables with an α occupation factor the lower bound when $n \to \infty$ is $\lceil -\alpha^{-1} \ell n(1-\alpha) \rceil$. We find that an average lower bound on the average number of accesses is 1.668... Simulation results indicate that the worst case bounds are very tight.

KEY WORDS:

Hashing, lower bounds, open-addressing, searching, Stirling numbers 2nd. class, minimax, average number of accesses.

CONTENTS

Average Lower Bounds for Open-Addressing Hash Coding

by

Gaston H. Gonnet
P.U.C. Rio de Janeiro

Abstract: In this paper we derive average lower bounds on the, worst case and average, number of accesses for open-addressing hashing tables. The average worst case for full tables is found to be $ln(n)+O(1)$, while for tables with an α occupation factor the lower bound when $n\to\infty$ is $\lceil -\alpha^{-1}ln(1-\alpha)\rceil$. We find that an average lower bound on the average number of accesses is 1.668... Simulation results indicate that the worst case bounds are very tight.

Introduction.

We consider the problem of finding lower bounds on the average behaviour of open-addressing hash coding techniques. We derive lower bounds for the average number of accesses and for the maximum number of accesses needed to find any element in the table. We will denote by m the size of the table where we insert n keys; $\alpha=n/m \leqslant 1$ is the load factor. Furthermore we assume that our hashing functions do not show clustering of any order.

Average lower bound for the maximum number of accesses.

The *minimax* hash coding problem is defined as follows: given a file and a hashing function, find the order in which elements have to be inserted so that the maximum number of accesses to locate any single element is minimized. This minimum maximum number of accesses will be called the *minimax* value. A lower bound on the minimax is also a lower bound on the worst case of any other table construction scheme.

We will first consider the case of a full table, i.e. $\alpha=1$. Our model for this analysis will require that the hashing function produce for any key, an independent random sequence of probes distributed discrete rectangular in (1,m). This is slightly different from the usual hashing functions since we allow probe positions to be repeated.

A necessary, but not sufficient, condition to solve the minimax problem for n=m with k accesses is that the first k probe positions of each key i.e. the k×n probe positions "occupy" all possible values from 1 to m. Since this is a necessary condition, minimax≥k.

Given k, the probability of all table positions (1,m) appearing in between k×n probe positions, is an occupancy distribution, also known as Arfwedson's distribution [Arfwedson 51, Stevens 37, Johnson 69]

$$\Pr\{\text{low-bound} \leqslant k\} = \sum_{i=0}^{n} (-1)^i \binom{n}{i} (1-i/n)^{kn}$$

$$= n! n^{-kn} \left\{ {kn \atop n} \right\},$$

where the braces denote a Stirling number of the second kind.

The expected value of low-bound for the above distribution is

$$E[\text{low-bound}] = \sum_{k=1}^{\infty} k[\Pr\{\text{low-bound} \leqslant k\} - \Pr\{\text{low-bound} \leqslant k-1\}]$$

$$= \sum_{k=1}^{\infty} (1 - n! n^{-kn} \left\{ {kn \atop n} \right\}).$$

Using asymptotic expansions of the Stirling numbers of the second kind [Moser & Wyman 58, David & Barton 62], we find

Theorem:

$$E[\text{low-bound}] = \ln(n) + O(1).$$

Numerical computation of the above expected value suggests that

$$E[\text{low-bound}] = \ln(n) + 1.07\ldots + o(1).$$

When we do not have a full table, we define T_k to be the number (a random variable) of different table positions that appeared in the first k probe positions of the n keys. The distribution of T_k is also an occupancy distribution and [Stevens 37, Johnson 69]

$$E[T_k] = m[1-(1-1/m)^{nk}],$$

$$\text{var}(T_k) = m(1-1/m)^{nk} + m(m-1)(1-2/m)^{kn} - m^2(1-1/m)^{2kn}.$$

For $\alpha = n/m$ and $n,m \to \infty$ we derive

$$E[T_k] = m(1-e^{-k\alpha}) + O(1),$$

$$\text{var}(T_k) = m(e^{-k\alpha} - (1+\alpha^2)e^{-2k\alpha}) + O(1),$$

and

$$\text{coef-var}(T_k) = O(m^{-1/2}).$$

Consequently for $n,m \to \infty$

$$\Pr\{T_k \geq n\} \to 0 \quad \text{iff} \quad E[T_k] < n$$

$$\to 1 \quad \text{iff} \quad E[T_k] > n,$$

(note that $E[T_k]$ cannot be an integer for $nk > 1$) and the expected value of the lower bound on the minimax, is k such that

$$m(1-e^{-k\alpha}) > n$$

or

Theorem:

$$k = \lceil -\alpha^{-1} ln(1-\alpha) \rceil,$$

with variance 0, i.e. the distribution is single valued when $m \to \infty$.

Lower bound for the average number of accesses.

The second possible optimum is the average number of accesses. To solve this optimality problem we have to find an order in which the keys have to be inserted so that the average number of accesses is minimized. This is a special case of the assignment problem, and it is discussed in [Gonnet & Munro 77, Rivest 77]. We will find a lower bound on the number of accesses needed to search a full table. In this case our model of the hashing function is one that produces, for each key, a random permutation of the locations 1 through m, and furthermore we will consider full tables.

Since the table is full, $n=m$, each location must contain a key. Consequently, whatever the order of insertion is, the location i will contain a key that needed a number of accesses greater or equal to the smallest order of appearance of i in any hash probe sequence, i.e. in any of the n permutations of 1 through m. The probability that location i does not appear in position $1,2,..,x$ of any random permutation is

$$\Pr\{\text{first appearance of i is after x probes}\} = [(m-x)/m]^m.$$

If location i has its first appearance after x probes, any key that is finally located in i will need more than x accesses. Consequently the expected value of the "first appearances" is a lower bound of the expected value of the number of accesses and is equal to

Theorem:

$$E[\text{first appearance}] = \sum_{k=0}^{m-1}[(m-k)/m]^m = e/(e-1) + O(m^{-1}) \sim 1.5819...$$

This lower bound can be further improved if we consider the collision of unique "first appearances". More precisely, we define a "first appearance" of position i in x probes as *critical* when only one key probes to i in x probes. For each collision (two or more critical appearances for one key) we will need at least on more access to fill both critical locations. Counting collisions we obtain

$$E[\text{collisions}] = \Pr\{2 \text{ critical for each key}\} + 2\times\Pr\{3 \text{ critical}\} + 3\times...$$

$$= E[\text{critical}] - 1 + \Pr\{\text{no critical}\}$$

$$= (e-1)^{-1} - 1 + \prod_{i=1}^{\infty}(1-e^{-i}) + O(m^{-1}).$$

The lower bound when $m\to\infty$ is now

Theorem:

$$2/(e-1) + \sum_{i=-\infty}^{\infty}(-1)^i e^{-(3i^2+i)/2} = 1.66838...$$

Note that when we write e.g. Pr{2 critical}, Pr{no critical}, etc. we are considering the probability of the joint events. The above method can be extended to consider two keys at a time, and count 3 collisions or more etc., but this extension does not seem to provide significant improvements.

Unsuccessful search.

The unsuccessful search case depends only on the load factor of the table. Under the natural assumption that probe positions do not repeat, the average number of accesses needed to complete the unsuccessful search is equal to the number of access needed to find an empty position, i.e. [Knuth 73]

$$E[\text{accesses}] = \sum_{k=1}^{n} k(m-n)n!(m-k)!/[(n-k+1)!m!]$$

$$= (m+1)/(m-n+1) \sim (1-\alpha)^{-1} \quad \text{if } m>n,$$

$$= n \quad \text{if } m=n.$$

Conclusions.

In this paper we derived average lower bounds for the, worst case and average, number of accesses in open-addressing hash tables. We find that for full tables the average worst case is $ln(n)+O(1)$. The following tables show simulation results for the minimax allocation algorithm [Gonnet & Munro 77]. Notice that the bounds apply very tightly.

Simulation results for Minimax Optimal Hashing
Size of table = 499 Sample size = 100

occup. factor	number of records	average accesses	average max. acc.	average p.q.o.
80%	399	1.49378±0.00670	3±0	4464.±198.
90%	449	1.64829±0.00785	3.05±0.0429	22120.±1744.
95%	474	1.69945±0.00704	3.99±0.0196	41644.±2787.
99%	494	1.78824±0.00774	5.12±0.0893	77304.±4815.

Table I

Simulation of Minimax Optimal Hashing for full tables.

file size	sample files	average accesses	average max. acc.	average p.q.o
19	1000	1.74858±0.0111	3.929±0.0622	241.04±7.35
41	600	1.79638±0.0102	4.665±0.0877	938.2±31.2
101	250	1.80737±0.0102	5.528±0.140	4851.±231.
499	100	1.82998±0.00807	7.38±0.287	91915.±3396.

Table II

Only in 5%-10% of the files simulated there were more accesses needed in the worst case, than those predicted for the lower bound. It is certainly surprising that the lower bound for the worst case in non-full tables is a constant integer. It is conjectured, and verified by simulation, that this bound is not only tight, but when $n \to \infty$ the minimax coincides with the lower bound.

The lower bound on the average number of accesses indicates that separate overflow chaining methods will always beat any "smart" rearrangement of keys in open-addressing hash. The following simulation results on the optimal average number of accesses suggest a value of 1.83 for the average which leaves little margin for improving this bound.

Simulation of Optimal Hashing for full tables.

file size	sample files	average accesses	average max. acc.	average p.q.o
19	1000	1.72895±0.0107	4.385±0.0710	224.13±5.46
41	500	1.78283±0.0111	5.296±0.105	888.3±26.2
101	200	1.79837±0.0105	6.3±0.175	4611.±157.
499	50	1.82381±0.0110	7.92±0.358	89937.±4334.
997	50	1.82794±0.00639	8.98±0.382	332365.±12373.

Table III

References.

[Arfwedson 51] Arfwedson, G. A Probability Distribution Connected with Stirling's Second Class-Numbers. Skandinavisk Aktuarietidskrift. 34-3 (1951), 121-132.

[David & Barton 62] David, F.N. and Barton, D.E. *Combinatorial Chance*. Hafner, New York, (1962).

[Gonnet & Munro 77] Gonnet, G.H. and Munro, J.I. The Analysis of an Improved Hashing Technique. Proceedings of the Ninth Annual ACM Symposium on the Theory of Computing. (May 1977).

[Johnson 69] Johnson, N.L., and Kotz, S. *Distributions in Statistics*. Houghton Mifflin, Boston, 1969, Vol. I, (Discrete Distributions).

[Knuth 73] Knuth, D.E. *The Art of Computer Programming*. Addison-Wesley, Reading, 1973.

[Moser & Wyman 58] Moser, L. and Wyman, M. Stirling Numbers of the Second Kind. Duke Mathematical Journal. 25 (1958), 29–43.

[Rivest 77] Rivest, R. Optimal Arrangement of Keys in a Hash Table. to appear JACM.

[Stevens 37] Stevens, W.L. Significance of Grouping. Annals of Eugenics, 8 (1937), 57–69.