PUC

Série: Monografias em Ciência da Computação
Nº 21/77

THE INTERPOLATION-SEQUENTIAL SEARCH ALGORITHM

by

Gaston H. Gonnet
Lawrence D. Rogers

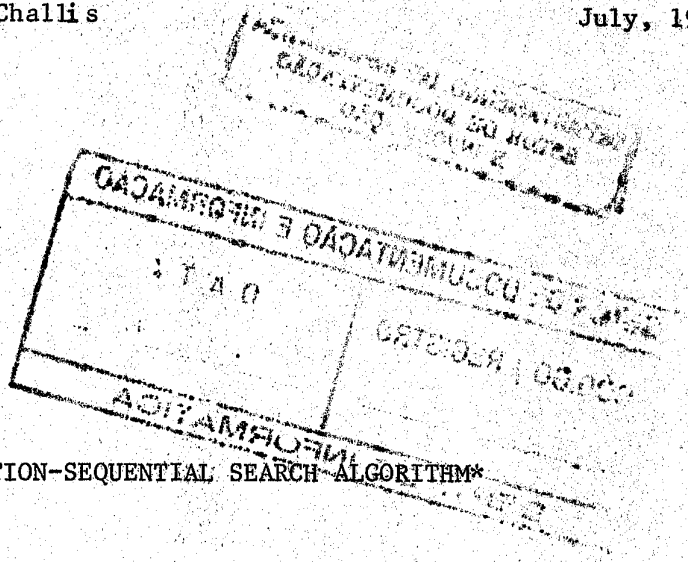
Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente 225 — ZC 19
Rio de Janeiro — Brasil

Serie: Monografias em Ciência da Computação
Nº 21/77

Series Editor: Michael F. Challis

July, 1977



THE INTERPOLATION-SEQUENTIAL SEARCH ALGORITHM*

by

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

and

Lawrence D. Rogers
Burroughs Corp. La Jolla, California

*This paper has been accepted for publication elsewhere. As a courtesy to the publishers it should not be widely distributed.

For copies contact:

Rosane T. L. Castilho
Head, Setor Doc. & Inf.
Depto. de Informática - PUC/RJ
Rua Marquês de São Vicente, 209 - Gávea
20000 - Rio de Janeiro, RJ - Brasil

Abstract. The interpolation sequential search algorithm is an algorithm to search ordered tables using implicit information about the distribution of the keys. It performs one probe based on an interpolation formula and then searches sequentially towards the appropriate end of the table. It is shown that under a model of random keys, the average number of accesses needed for the successful search is $1 + (n\pi/32)^{1/2} + o(1)$. The analysis is extended to the same search applied to blocked external files.

Keywords: Searching, Interpolation search, Sequential search, Binary search, Analysis of Algorithms, Asymptotic analysis.

Resumo. O algoritmo de pesquisa por interpolação sequencial é utilizado para pesquisa em tabelas ordenadas fazendo uso implícito da distribuição de probabilidade das chaves. Efetua um teste baseado em uma fórmula de interpolação e logo pesquisa sequencialmente até o correspondente fim. Demonstramos que com um modelo de chaves aleatórias o número médio de acessos é $1 + (n\pi/32)^{1/2} + o(1)$. Extendemos a análise para o caso de arquivos externos bloqueados.

Palavras chave: Pesquisa, Pesquisa por interpolação, Pesquisa sequencial, Pesquisa sequencial, Pesquisa binária, Análise de algoritmos, Análise assintótica.

The interpolation sequential search algorithm is an algorithm to search ordered tables. It was first described by Price (1971). The algorithm performs an initial probe in the table by interpolating the searched key value with the limiting values of the table. In case the element is not in the position searched, searches sequentially towards the appropriate end of the table. This algorithm is a variation from interpolation search [Gonnet, 77 and Yao & Yao, 76] and does not follow the $log_2(log_2(n))$ asymptotic behavior.

We will assume that we search a file of n, i.e., $X_1, X_2, \ldots, X_n$, random independent variables distributed $U(0,1,)$. Note that whenever we know the cumulative density function of the keys, e.g.

$$F(a) = Pr|Y \leq a|$$

and it is continuous, the transformation

$$X = F(Y)$$

allows us to work with $U(0,1)$ keys.

The interpolation sequential algorithm to search for α in the vector $X(1),\ldots, X(n)$ can be coded in a pseudo language as:

```
j:= ⌈nα⌉;
select (X(j))
        case  α: return(j);
        case <α: for j: = j + 1  to  n while X(j) ≥ α  do od;
        case >α: for j: = j - 1  step -1 to 1 while X(j) ≤ α  do od;
        end select;
if X(j) = α then return (j) else return (FAIL);
```

In the analysis of this algorithm we will first consider the successful search. We will consider the number of elements from X that are referenced our measure of complexity, which will be called the number of accesses.

If we consider the search of the j^{th} key, $X_j = \alpha$, the expected number of accesses needed to locate it is given by

$$E[\textit{accesses to find } X_j] = \int_0^1 [1 + |\lceil n\alpha \rceil - j|] I'_\alpha(j, n-j+1)\, d\alpha$$

$$= 1 + \int_0^1 |\lceil n\alpha \rceil - j| I'_\alpha(j, n-j+1)\, d\alpha \ ,$$

where $I_\alpha(j, n-j+1)$ is an incomplete Beta function [Abramowitz and Stegun, 1964] which is the distribution of the j^{th} ordered $U(0,1)$ random variable among n and $I'\alpha(j, n-j+1)$ denotes its derivative with respect to α.

The average number of accesses to find any X_j, assuming that each key is equally likely to be accessed, is

$$E[\textit{accesses}] = 1 + 1/n \sum_{j=1}^{n} \int_0^1 |\lceil n\alpha \rceil - j| I'_\alpha(j, n-j+1)\, d\alpha \ .$$

$$E[\textit{accesses}] = 1 + 1/n \sum_{n=1}^{n} [I_{1/n}(j, n-j+1) + I_{2/n}(j, n-j+1) + \ldots$$

$$+ I_{(j-1)/n}(j, n-j+1) + 1 - I_{j/n}(j, n-j+1) + 1 - \ldots +$$

$$+ 1 - I_{(n-1)/n}(j, n-j+1)] \ .$$

Using the symmetry relation $I_x(a,b) = 1 - I_{1-x}(b,a)$ we transform the above equation into

$$E[\textit{accesses}] = 1 + 2/n \sum_{j=1}^{n} \sum_{k=1}^{j-1} I_{k/n}(j, n-j+1) \ .$$

Reversing the order of summation and using the formula [Gonnet 1977]

$$\sum_{j=k}^{n} I_x(j, n-j+1) = nxI_x(k-1, n-k+1) - (k-1)I_x(k, n-k+1) \ ,$$

we derive

$$E[accesses] = 1 + 2/n \sum_{k=1}^{n-1} [kI_{k/n}(k,n-k) - kI_{k/n}(k+1,n-k)]$$

and finally using [Abramovitz and Stegun, 1964, e.q. 26.5.16]

$$E[accesses] = 1 + 2/n \sum_{k=1}^{n-1} \Gamma(n)/[\Gamma(k)\Gamma(n-k)][k/n]^{k}[(n-k)/n]^{n-k} .$$

So far the result is exact, but not very useful, so we will look for an asymptotic expansion in terms of n.

Using Stirling's approximation of the Gamma function we decompose each summand into

$$[2/(n^3\pi)]^{1/2} \times [1 + 1/12n + 1/(288n^2) + O(n^{-3})] \times [1 - n/[12k(n-k)] + \ldots] \times [k(n-k)]^{1/2}.$$

where the ... stand for a sum of terms of the form

$$\frac{b_{ij}n^{i-1}k}{[k(n-k)^j]} \quad \text{or} \quad \frac{c_{ij}n^{i}}{[k(n-k)^j]}$$

with b_{ij} and c_{ij} being constants independent of n or k, $j \geq 2$ and $i \leq j$.

This transformation is straight forward, except that we must collect the terms and simplify in the product $\Gamma(k)\Gamma(n-k)$.

The summation formula

$$\sum_{k=1}^{n-1} [k(n-k)]^{-s} = (n/2)^{1-2s}\, \pi^{1/2}\Gamma(1-s)/\Gamma(3/2-s) + 2n^{-s}[\zeta(s) + s\zeta(s-1)/n +$$

$$+ s(s+1)\zeta(s-2)/2n^2 + \ldots + \Gamma(s+i)\zeta(s-i)/\Gamma(s)i!n^i + \ldots] \quad [s \neq 2,3, \ldots]$$

is basic to the following derivation.

The first two terms of each summand are constants and are factored out of the summation. The sum of the terms denoted by ... is

$$0 < n^{i-1} \sum_{k=1}^{n-1} \frac{k}{[k(n-k)]^{j-1/2}} \le n^i \sum_{k=1}^{n-1} \frac{1}{[k(n-k)]^{j-1/2}} = \begin{cases} O(n^{1/2}) & \text{for } i = j \\ O(1) & \text{for } i < j \end{cases}$$

Hence, the sum of the last two terms, using the same summation formula, is

$$\sum_{k=1}^{n-1} \{[k(n-k)]^{1/2} - \frac{n}{12[k(n-k)]^{1/2}}\} + O(n^{1/2})$$

$$= \frac{n^2\pi}{8} + 2n^{1/2}\zeta(-1/2) + \ldots \frac{-n}{12}[\pi + 2n^{-1/2}\zeta(1/2) + \ldots + O(n^{1/2})$$

$$= \frac{n^2\pi}{8} - \frac{n\pi}{12} + O(n^{1/2})$$

multiplying we finally obtain

$$E[accesses] = 1 + [n\pi/32]^{1/2}[1 - 7/12n] + O(n^{-1})$$

$$\sim 1 + [(n-1)\pi/32]^{1/2} + O(n^{-1/2}) .$$

the latter being an equivalent asymptotic expansion with good numerical approximation to the exact result.

If we are searching in direct access secondary storage, we will be interested in the number of accesses to blocks of records. Let us assume that the n keys are accessible in blocks containing b keys each. Ignoring edge effects, if we need k sequential key accesses in any direction starting at a random location, the number of block accesses is

$$E[blocks\ accessed] = (k-1)/b + 1$$

The average number of block accesses for an interpolation-sequential search is

$$E[blocks\ accessed] = 1 + b^{-1}[(n-1)\pi/32]^{1/2} + O(n^{-1/2})$$

The following table compares interpolation-sequential search with binary search for some selected file sizes. These results are computed with the above formulas and those in [Knuth 1973].

file size	interpolation sequential	binary search	inter-seq (b=10)	binary (b=10)
10	1.92	2.9	1	1
50	3.19	4.86	1.22	2.2
100	4.11	5.8	1.31	2.9
500	8.00	8.00	1.70	4.86
1000	10.90	8.99	1.99	5.8
10000	32.33	12.36	4.13	8.99

The interpolation sequential algorithm is a reasonable alternative for internal searching in tables up to size 500. For external searches, it performs better than binary search over a wide range of sizes, without taking into account the advantage of reduced seek time.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1964), Handbook of Mathematical Functions, Dover Publications Inc., New York

Gonnet, G. H. (1977), Interpolation and Interpolation Hash Searching, Research Report CS 77-02, Department of Computer Science, University of Waterloo, Waterloo, Ontario

Knuth, D. E. (1973), The Art of Computer Programming, Volume 3, Addison Wesley Publishing Company, Inc., Massachusetts

Price, C. E. (1971), Table Look-Up Techniques, Computing Surveys, Vol. 3, No. 2, June 1971, p. 56-58

Yao, A. C. and Yao, F. F. (1976), The Complexity of Searching on Ordered Random Table, Proceedings of the Symposium on Foundations of Computer Science, Houston, October 1976, p. 173-176