JOURNAL OF RESEARCH of the National Bureau of Standards
Vol. 83, No. 5, September–October 1978

# Hashing with Linear Probing and Frequency Ordering

**Gordon Lyon**

**Institute for Computer Sciences and Technology, National Bureau of Standards, Washington, D.C. 20234**

**(June 5, 1978)**

A simple linear probing and exchanging method of Burkhard locally rearranges hash tables to account for reference frequencies. Examples demonstrate how frequency-sensitive rearrangements that depend upon linear probing can significantly enhance searches.

Key words: Hashing; linear probing; open addressing; optimal packing; retrieval improvement; nonuniform frequencies.

## 1. Linear Probing

*Linear probing* of a scatter (or hash) table interprets each key or item (these terms are interchangeable here) as a probe index into the table [1].[1] Typically, a key is divided by the table size and the remainder is used for indexing. If the selected slot is empty, the item is not present. Should the slot contain some other key, each next higher location is checked until the item is found, an empty slot is discovered, or the whole table has been examined. (Indexes that exceed table sizes wrap around.) Table slots thus searched define a key's collision resolution sequence.

Linear probing is generally not suitable for nearly filled tables, since as empty slots disappear, searches get very long. Nonetheless, linear probing can be improved by allowing for item frequency-of-reference.

### 1.1. Burkhard's Heuristic

Recently, Burkhard has suggested an heuristic method of reordering scatter tables that are accessed via linear probing [2]. Burkhard's scheme depends upon the item ensemble $E$ in a table having nonuniform access frequencies. Each reference to an item initiates an exchange of the item with its immediate predecessor on the collision resolution sequence, provided this is possible. Intuitively, one can see that more frequently accessed items remain near their original probes into the table, whereas unpopular table members migrate to poorer locations. The cost of exchanges can be reduced, for example, by performing them only on every tenth table access.

## 2. Theoretical Limitations

A number of recent studies examine limitations that exist on orderings of open addressing hash tables [3, 4, 5]. A typical limitation is that, given uniform frequencies, minimum average retrievals (in probes-per-item) are $A(0.8) \cong 1.49$, $A(0.9) \cong 1.61$, $A(0.95) \cong 1.69$, and $A(1.0) \cong 1.83$ for table loadings of 0.8, 0.9, 0.95 and 1.0. (Entries in table 1, $d = 0.0$, show how ordinary linear probing compares.) Such values of $A(\;)$ are determined by optimal solutions to "distribution" or "assignment" problems common in operations research. However, accounting for skewed frequencies can be as significant as using "assignment problem" solutions, which are also expensive for tables of several hundred entries. The potential advantages of rearrangements by frequencies include simplicity, low insertion costs, and

TABLE 1. *The influence of d.*

| $d$ | 0.0 | 0.25 | 0.50 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $A$ (0.50) | 1.50 | 1.44 | 1.39 | 1.33 | 1.27 |
| $A$ (0.90) | 5.50 | 4.68 | 3.87 | 3.05 | 2.23 |
| $A$ (0.95) | 10.50 | 8.57 | 6.63 | 4.70 | 2.77 |
| $A$ (0.99) | 50.50 | 38.92 | 27.30 | 15.77 | 4.10 |

[1] Figures in brackets indicate the literature references at the end of this paper.

adaptability. Population frequencies can even change with Burkhard's scheme. Weighting by frequency and "assignment" solutions can be combined, but the net improvement is often not nearly as great as each component might suggest. Given a sufficiently skewed population, it is most important to attend to the frequency weights [5].

### 2.1. Descending-Popularity Insertions

Peterson has proven that equiprobable items are insensitive to insertion order when linear probing is used [6]. A variation in argument shows that items inserted in order of decreasing frequencies achieve minimum average retrieval costs for their ensemble $E$ [1, 3]. Assume that a partially filled table is packed optimally. Consider a new item of frequency equal to or less than any other item already in the table. Let the new item be contending for a table slot that is filled. By the nature of linear probing, both the new and the resident will probe the same filled slots in a search for an empty alternative slot. Although the increases in search probes are identical, the extra probes should be assigned to the least frequent, which is the new item.

### 2.2. Large Tables

Continuous distributions and integrals provide good approximate results when tables are large. Let $f(x)$ be a distribution of ensemble items such that

$$\int_0^1 f(x)\,dx = 1 \quad \text{and} \quad f'(x) \leq 0$$

$f(x)$ represents a sorted order of descending frequencies. The expected probes-to-insert-an-item is approximated well for linear probing by [1]:

$$C'(\alpha) \cong (1/2)*(1 + (1/(1 - \alpha))**2)$$

"$\alpha$" indicates the table loading, i.e., the ratio of $|E|$ to the table size. Insertion of the full ensemble $E$ into a table gives an optimal retrieval average of

$$A(\alpha) = \int_0^1 f(x)*C'(\alpha x)\,dx.$$

The marginal insertion cost C′ is sensitive to the true occupied table fraction "$\alpha x$" rather than $x$ alone. Since the ensemble probability is unity, the retrieval expression can be simplified slightly to

$$A(\alpha) \cong (1/2)$$
$$+ (1/2)* \int_0^1 f(x)/((1 - \alpha x)**2)\,dx. \qquad (i)$$

*Example 1.* Imagine an ensemble with a sorted frequency distribution of $f(x) = 1 + d - (2dx)$. Such a population corresponds to a tilting or skewing of the usual uniform distribution ($d = 0$). Maximum skewedness ($d = 1$) gives a sawtooth distribution. Applying ($i$) and integrating directly,

$$A(\alpha) \cong (1/2)*[1 + (1 - d)/(\alpha*(1 - \alpha))$$
$$- (1 + d)/\alpha - (2*d)/(\alpha*\alpha)* \log(1 - \alpha)].$$

The effect of "$d$" is pronounced and useful, as demonstrated in table 1. Improvements are linear in $d$.

## 3. A Practical Estimation Technique

In many cases it is illuminating to apply ($i$) to estimate retrieval prospects for observed data. Empirical frequencies do not always fit common curves, e.g., data are discontinuous or derive from multimodal distributions. Nevertheless, ($i$) takes a very simple tabular form when a distribution $f(x)$ comprises segments of straight lines. Such lines are easily drawn on a data frequency plot, and the necessary end points and slopes read directly once the plot area has been normalized to unity. Expanding ($i$) by parts and noting that $f'(x)$ is constant for straight lines:

$$A(\alpha) \cong \frac{1}{2} + \frac{1}{2}\left[\frac{f(x)}{\alpha(1 - \alpha x)} + \frac{f'(x)}{\alpha^2}\log(1 - \alpha x)\right]_0^1.$$

Then breaking the interval [0, 1] into line segments $\{I\}$, each denoted $[I_-, I_+]$,

$$A(\alpha) \cong \frac{1}{2} + \frac{1}{2}\sum_{\{I\}}\left[\frac{f_I(x)}{\alpha(1 - \alpha x)} + \frac{f_I'(x)}{\alpha^2}\log(1 - \alpha x)\right]_{x=I-}^{x=I+}.$$

*Example 2 — (Linear segments).* The tabulation technique is easily applied to problems. Let $f(x) = 3$ when $0 \leq x \leq 0.25$, $f(x) = 0.50$ when $0.25 < x \leq 0.50$, and $f(x) = 1 - x$ otherwise. Note that $f(x)$ integrates to unity on the interval [0, 1] and that $f(x + \epsilon) \leq f(x)$. Applying the tabulation formula,

$$A(\alpha) \cong \frac{1}{2} + \frac{1}{2}\left[\frac{3}{\alpha\left(1 - \frac{\alpha}{4}\right)} - \frac{3}{\alpha} + 0\right.$$
$$- 0 + \frac{0.5}{\alpha\left(1 - \frac{\alpha}{2}\right)} - \frac{0.5}{\alpha\left(1 - \frac{\alpha}{4}\right)}$$
$$+ 0 - 0 + 0 - \frac{0.5}{\alpha\left(1 - \frac{\alpha}{2}\right)} + \frac{(-1)}{\alpha^2}\log(1 - \alpha)$$
$$\left. - \frac{(-1)}{\alpha^2}\log\left(1 - \frac{\alpha}{2}\right)\right].$$

Retrieval values for the example are $A$ (0.5) = 1.17, $A$ (0.9) = 1.68, $A$ (0.95) = 1.94, and $A$ (0.99) = 2.66.

## 4. Summary

Theoretical examples demonstrate that table reorderings make excellent improvements on linear probing, especially in everyday applications that more often than not have distinctly unequal reference frequencies. The simplicity and low costs of linear (exchange) probing make it an attractive possibility for practical applications.

## 5. References

[1] Knuth, D. E., The Art of Computer Programming, Vol **3**. (Addison-Wesley, Reading, Mass.), 518 ff.

[2] Burkhard, W. A., Self-organizing hash table search heuristics, Proc., 1978 Conf. on Inf. Sci. and Systems. (March, 1978), 378–399.

[3] Rivest, R. L., Optimal arrangement of keys in a hash table, Jour. ACM. **25**, 2 (April, 1978), 200–209.

[4] Gonnet, G. H. and Munro, J. I., The analysis of an improved hashing technique, Proc., Ninth Annual ACM Symp. on the Theory of Computing (May, 1977), 113–121.

[5] Lyon, G., Packed scatter tables, Comm. ACM. (to appear).

[6] Peterson, W. W., Addressing for random-access storage, IBM J. Res. Dev. **1**, 4 (April, 1957), 130–146.