Aplikace matematiky

Jaroslav Král

# Some very effective methods of searching in tables

*Aplikace matematiky*, Vol. 14 (1969), No. 1, 1–14

Persistent URL: http://dml.cz/dmlcz/103203

# SOME VERY EFFECTIVE METHODS OF SEARCHING IN TABLES

JAROSLAV KRÁL

(Received February 23, 1967)

## 1. INTRODUCTION

In many areas of automatic programming the following problem must be solved. Some source A generates a sequence of items $x_1, x_2, x_3, \ldots$ which will be called keys. It is assumed that on the set of keys $x_i$ a relation of equality $(=)$ is given. We have to construct a table $T$ in the following manner: For every $m \geqq 1$ $T$ contains just all distinct keys from the set $\{x_1, x_2, \ldots, x_m\}$.

**1.1. Example.** During the translation of an Algol program a table of constants in the program without repetition must be constructed. In this case the source is the scanning part of the translator, the keys being scanned numbers.

**1.2. Example.** In mechanical translation of natural languages, keys are the words appearing during the reading.

**1.3. Example.** Automatic stock administration. In this case the customers, giving their demands, form the source in question. The keys are the names of items in the stock.

Every key is a head (or key) of further information. For example in 1.1. the information is the key itself, in 1.2. an equivalent of a given word, in 1.3. an information needed for the stock administration. The question of formation of this information or obtaining some will not be discussed further.

The usual examination of a table $T$ is a successive examination of all keys in $T$. Then the mean value of examination under the condition that $x \neq x_i$ for all keys $x_i$ in $T$ is just $n$. In the opposite case the mean value is equal (under certain conditions) to $(n + 1)/2$. In [1] an algorithm was proposed needing approx. $\log_2 (n)$ $(1/2 \log_2 n)$ examinations. But in [2], [3], [5] and [6] there were introduced and studied methods which, as we shall see below, need a bounded number of examinations if the number

$n$ of items in the table $T$ tends to infinity. This fact is proved in section 3. In section 2 three types of algorithms for constructing tables are given. In section 4 exact formulae for the mean value of a number of examinations are given. In chapter 5 the situation, when a backing store is used, is discussed. Some numerical results will be given and discussed in a subsequent paper.

## 2. DEFINITION OF ALGORITHMS

**2.1. Method.** *A* (see [1]). This method assumes that on the set of keys a relation $\prec$ of full ordering is defined. During the forming of the tabel T the so called admissible tree is constructed. The admissible tree is a rooted finite binary tree (i.e. if an vertex of a tree is not an end-vertex, then it has just two successors: the left and the right one) the vertices of which are labelled by keys placed in the table. The tree fulfils two additional conditions:

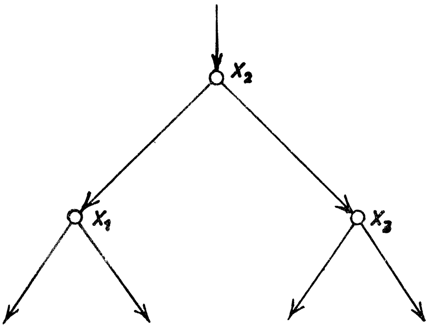

Fig. 1

a) If the keys $x_1$, $x_2$, $x_3$ denote labelling of three vertices from figure 1 and $x_L = x_1$ ($x_R = x_3$) or $x_L$ ($x_R$) is accessible from $x_1$ (from $x_3$) then $x_L \prec x_2 \prec x_R$. The set of labels of all vertices is just the set of all items in $T$.

b) If $l(x_i, x_{i1}, \ldots)$ denotes the length of the longest path in the tree, the first two vertices of which are labelled by $x_i$, $x_{i1}$, then for every vertex $V$ in the tree labelled by $x$

$$|l(x, x_l, \ldots) - l(x, x_r, \ldots)| \leqq 1\,, \tag{1}$$

$x_l$ or $x_r$ denotes the labelling of the left or of the right successor of $V$ respectively.

It can be shown that for the length $l_0$ of the longest path in an admissible tree $l_0 \leqq 3/2 \log_2 n$ where $n$ is a number of vertices in the tree and that the fact that the key $x$ does not label any vertex in the tree can be discovered by examining just one path in the tree. In [7] it was proved that the mean value of lengths of paths leading from the root to the end vertices in admissible trees is $\log_2 n$. If $x$ is not found, the vertex labelled $x$ is added and the tree is reconstructed by a very simple manner in order to save the property of admissibility. In the computer an admissible tree is coded as a table each member of which contains two keys pointing to the left and right successor.

**2.3. Definition.** Key function $f(a, n)$ is a single valued function defined on the set $K$ of all keys, with values from $1, 2, \ldots, n$ i.e. a function which to every key $a$ assigns just one positive integer not greater than $n$.

**2.4. Definition.** Consider a table $T$. We shall assume that the table $T$ is an array of length $n$, $T = \{T_1, \ldots, T_n\}$. $T_i$ has either an undefined value or its value is a key. It is understood than $T$ is formed by means of a key function $f$ by method $B$ if at the beginning of the formation of $T$ the value of every member of $T$ is undefined and if for every key $x$ generated by the source the following operations are carried out:

(i) The value $t = f(x, n)$ is evaluated.

(ii) The value of $T_t$ is examined. If the value of $T_t$ is undefined, we put the value of $T_t$ equal to $x$, these operations being thereby completed. The operations also end if the value $T_t$ is $x$. If the value of $T_t$ is not $x$, $t$ is put equal to $t \dotplus 1$ and (ii) is repeated. Here, as well as below, $\dotplus$ ($\dot{-}$) denotes addition (substraction) modulo $n$. $T_i$ will further be called the $i$-th member of $T$ and $n$ — the number of $T_i$ in $T$ — the length of $T$.

**2.5. Example.** The method of searching in $T$, as was stated in the introduction, is method $B$ with $f(a, n) \equiv 1$.

**2.6. Definition.** A key function $f(x, n)$ is, for the given source $A$ of keys, a random key function is and only if for the sequences $x_1, x_2, \ldots$ generated by the source $A$ the sequence $f(x_1, n), f(x_2, n), \ldots$ is a sequence of independent random variables uniformly distributed on $1, 2, \ldots, n$ i.e. $P(f(x, n) = i) = 1/n$.

**2.7. Definition.** The table $T$ is created by the method $C$ if

(v) at the beginning of the formation of $T$ the table $T$ contains members $T_1, \ldots, T_n$, the values of which are undefined, and $r = n$, each member $T_i$ of $T$ contains besides the value of a key, which can be undefined, a pointer part, the value of which can be undefined or is an integer $s > n$.

(iv) The value $t = f(x, n)$ is evaluated

(iiv) The value of the key part in $T_t$ is examined. If the value of the key in $T_t$ (value of $T_t$) is not $x$, then the pointer part of $T_t$ is examined. If its value is not defined, a new member $T_{r+1}$ is added to $T$, the value of the key part in $T_{r+1}$ is put equal to $x$, then $r$, which is the length of $T$, is increased to $r + 1$ and the value of the pointer part in $T_i$ is put equal to $r + 1$. If the value of the pointer part in $T_t$ is $s$, the member $T_s$ is examined according to (iiv).

**2.8. Proposition.** *If a sequence $x_1, x_2, \ldots, x_s$ of keys does not contain more different keys than $n$, then all the methods $A$, $B$, $C$ form a table $T$ without repetitions containing all different keys in $x_1, x_2, \ldots, x_s$.*

Proof can be easily carried out by induction.

**2.9. Terminology.** The table $T$ contains $k$ keys if just $k$ members of $T$ have defined values. The table $T$ has parameters $(n, k)$ if its length is equal to $n$ and it contains

$k$ keys. A key $x$ is introduced into the table $T$ if the value of some $T_i$ is put equal to $x$. $T$ does not contain $x$ if no value of $T_i$ is equal to $x$; in the opposite case $T$ contains $x$. In the remaining part of the paper we shall assume that a random key function is used.

## 3. SOME THEOREMS FOR MEAN VALUES OF THE NUMBER OF EXAMINATIONS

**3.1. Theorem.** *Let $T$ have parameters $(n, k)$. Let $k \leqq n \, . \, d$ (where $0 < d < 1$) for every $n$. Then the mean value of the number of examinations before finding that a key $x$ must be introduced into the table $T$ is bounded as $n \to \infty$ for methods $B$ and $C$.*

Proof. Denote the dependence of events and its probabilities on $k$ and $n$ by the upper subscripts. Let $P(A)$ denote a probability of $A$ and let $A_{i,m}^{k,n}$ be the event that, beginning with the examination of $T_i$, just $m$ examinations are done.

$$(1) \qquad P(A_{i,m}^{k,n}) = \sum_{j=0}^{n-m} P(B_{i-j,m+j}^{k,n}) = \sum_{j=0}^{n-m} P(B_{i,m+j}^{k,n}) = \sum_{j=0}^{n-m} P_{m+j}^{k,n}$$

where $P(B_{i,q}^{k,n})$ is the probability of the event $B_{i,q}^{k,n}$ that values of $T_{i-1}$, $T_{i+q}$ are undefined and values of $T_i, \ldots, T_{i \dotplus q \dotminus 1}$ are defined. The probability $P(B_{i,g}^{k,n})$ obviously does not depend on $i$, so $P(B_{i,q}^{k,n})$ is equal to some number $p_q^{k,n}$. But

$$(2) \qquad P(B_{i,q}^{k,n}) \leqq p(C_{i,q}^{k,n})$$

$C_{i,q}^{k,n}$ is the event that during the formation of $T$ the value of the key function $f$ equal to $i \dotminus 1$ and $i \dotplus q$ was not obtained and for keys which appear in the sequence for the first time just $q$-times the value of the key function equal to $i, i \dotplus 1, \ldots, i \dotplus q \dotminus 1$.

We have

$$(3) \qquad P(C_{i,q}^{k,n}) = \frac{k!}{0!\,0!\,q!\,(k-q)!} \left(\frac{1}{n}\right)^0 \left(\frac{1}{n}\right)^0 \left(\frac{q}{n}\right)^q \left(1 - \frac{q+2}{n}\right)^{k-q}$$

If $q$ remains constant for $b \to \infty$ and $\lim\limits_{n \to \infty} k/n = d < 1$ we obtain

$$(4) \qquad \lim_{n \to \infty} P(C_{i,q}^{k,n}) = \lim_{n \to \infty} \frac{q^q}{q!} \left(1 - \frac{q+2}{n}\right)^{k-q} \frac{k}{n} \left(\frac{k}{n} - \frac{1}{n}\right) \ldots \left(\frac{k}{n} - \frac{q-1}{n}\right) = \\ = e^{-d(q+2)} \, . \, \frac{(qd)^q}{q!}$$

A generating function for probabilities (4) is

$$(5) \qquad G(x) = e^{-2d} \sum_{m=0}^{\infty} \frac{(d \, . \, m \, . \, e^{-d} \, . \, x)^m}{m!}$$

where we put

$$P(C^d_{i,0}) = e^{-2d}$$

Now, using the independence of probabilities on $i$, we obtain from (1)

$$(6) \qquad P(A^{k,n}_{i,m}) = P_m = \sum_{j=0}^{n-m} p^{k,n}_{m+j}$$

The mean value $E(n, k)$ of examinations in $(ii)$ in 2.4. under the condition that $x$ is not in $T$ is equal to the number of examinations before the $T_i$ of an undefined value is found plus one, i.e.

$$(7) \qquad E(n, k) - 1 = \sum_{j=1}^{k} jP_j = \sum_{j=1}^{k} j \sum_{t=j}^{k} p^{k,n}_t = \sum_{j=1}^{k} \frac{j(j+1)}{2} p^{k,n}_j$$

From (5) for great $n$ using Stirling's formula together with (2) and (4) we get

$$(8) \qquad p^{k,n}_m \leqq \frac{(de^{-d+1})^m}{\sqrt{(2\pi m)}} e^{-2d} < \frac{(de^{-d+1})^m}{\sqrt{(2\pi)}} \,.\, e^{-d2}$$

so for $n \to \infty$, $k/n \to d < 1$ and $L \geqq 4$ denoting $S_L = \sum_{i=1}^{L} p^{k,n}_i \,.\, i$

$$(9) \qquad E(d) = \lim_{\substack{n\to\infty \\ n/k \to d}} E(n, k) \leqq S_L + \left( \sum_{j=L}^{\infty} e^{-2d} \frac{j(j+1)}{2} \frac{(de^{-d+1})^j}{\sqrt{(2\pi j)}} \right) \times 1{\cdot}02$$

$$\leqq S_L + \frac{1{\cdot}02e^{-3d+1}}{\sqrt{L} \,.\, 2 \,.\, \sqrt{(2\pi)}} \left( \sum_{j=1}^{\infty} x^{j+1} \right)''_{x=de^{1-d}}$$

and the theorem is proved. For method $C$ the conclusion of the theorem follows from proposition 3.3.

**3.2. Remark.** It is clear that (9) does not estimate $E(d)$ too well. For example for $d = 0{\cdot}5$ we obtain $E(d) \leqq 11{,}5$. Taking into consideration, however, that the number of cases when the value of a key function $f$ is equal to $i, i + 2, \ldots, i + l/2$ cannot be less than the number of cases when $f$ is equal to some number from $i + l/2, \ldots$ $\ldots, i + l \dot{-} 1$ plus one we can find out that $E(1/2) \leqq 6{\cdot}0$.

**3.3. Proposition.** *For the mean value $E_C$ and the dispersion $D_C$ of the number of examinations according to method C before finding $T_i$ of an undefined value it is true*

$$(1) \qquad E_C(d) = D_C(d) = d$$

Proof. The number of examinations for $n \to \infty$, $k/n \to d$ has the Poisson's distribution with parameter $d$. The proposition remains true also in the case that $d \geqq 1$ (see [6]).

**3.4. Lemma.** *Let T be a table with parameters $(n, k)$ formed by method B. Then*

$$(1) \qquad P(B^{k,n}_{i,q}) = p^{k,n}_q = \frac{k!}{n^k}\left(\sum_{Q_{q,q}} \frac{1}{\prod\limits_{i=1}^{q} a_i!}\right)\left(\sum_{Q_{n-q-2,k-q}} \frac{1}{\prod\limits_{i=1}^{n-q-2} a_i!}\right)$$

*where*

$$(2) \qquad Q_{r,s} = \{(a_1, a_2, \ldots, a_r) |\ a_i \geqq 0,\ a_i \text{ is an integer for } i = 1, \ldots, r,\ \sum_{i=1}^{r} a_i = s,$$
$$\sum_{i=r-h+1}^{r} a_i \leqq h,\ h = 1, 2, \ldots, r-1\}$$

*$\{x \mid \mathscr{P}(x)\}$ denotes the set of all x for which a proposition $\mathscr{P}(x)$ is true.*

Proof. We note that the considered probabilities are given by the polynomial law, i.e.

$$(3) \qquad P(B^{k,n}_{i,q}) = \sum_{c} \frac{k!}{a_1!\, a_2! \ldots a_n!} \frac{1}{n^k} = \sum_{Q_1} (\prod_{j=i}^{i+q} a_j!)^{-1} \sum_{Q_2} (\prod_{\substack{j<i\\ j>i+q}} a_j!)^{-1} \frac{k!}{n^k}$$

It can be shown by simple combinatorial considerations that $Q_2$, $Q_1$ are just the sets given in (1), (2).

**3.5. Lemma.** *If we denote*

$$(1) \qquad D(k, q) = \sum_{Q_{k,q}} \frac{1}{\prod\limits_{i=1}^{k} a_i!}$$

*then*

$$(2) \qquad p^{k,n}_q = (k!/n^k)\, D(q, q)\, D(n - q - 2, k - q)$$

*where $D(i, j)$ are defined for $i, j \geqq 0$, $D(i, 0) = 1$ for all $i \geqq 0$, $D(i, j) = 0$ for $j > i$ and for $0 \leqq k \leqq q$ it holds*

$$(3) \qquad D(k, q) = \sum_{j=0}^{q} \frac{1}{j!} D(k - 1, q - 1)$$

Proof.

$$(4) \qquad \sum_{Q_{k,q}} \frac{1}{\prod\limits_{i=1}^{k} a_i!} = \sum_{a_1=0}^{q} \frac{1}{a_1!} \sum_{Q'_{k,q}} \frac{1}{\prod\limits_{i=2}^{k} a_i!}$$

where

$$Q'_{k,q} = \{(a_2, \ldots, a_k) \mid a_i \geqq 0, a_i \text{ is an integer for } i = 2, 3, \ldots, k\,;$$
$$\sum_{i=2}^{k} a_i = q - a_1,\ \sum_{i=k-h+1}^{k} a_i \leqq h\}$$

so

$$Q'_{k,q} = Q_{k-1,q-a_1} \text{ and (1) is proved .}$$

**3.6. Theorem.** *The mean value $E(n, k)$ of examinations in table $T$ for method $B$, before a member of $T$ of an undefined value is found fulfils the equality*

(1) $$E(n, k) = \sum_{j=1}^{k} \frac{k!}{n^k} \frac{j(j+1)}{2} D(n-j-2, k-j) D(j, j)$$

Proof. Immediately from 3.1.7. and 3.5.2.

**3.7. Theorem.** *For the dispersion $S(n, k)$ of the number $N$ of examinations in the table $T$ before finding $T_i$ of an undefined value it holds*

(1) $$S(n, k) = \sum_{j=1}^{k} \frac{j(j+1)(2j+1)}{6} \frac{k!}{n^k} D(n-j-2, k-j) D(j, j) - (E(n,k))^2$$

Proof. the same as in 3.6. but instead of 3.1.7. we use

$$\sum_{j=1}^{k} j^2 P_j = \sum_{j=1}^{k} j^2 \sum_{t=j}^{k} p_t^{k,n} = \sum_{j=1}^{k} \frac{j(j+1)(2j+1)}{6} p_j^{k,n}$$

**3.8. Definition.** The average price $Q(n, k)$ of forming a table $T$ with parameters $(n, k)$ is the mean value of the random variable

(1) $$\frac{1}{k} \sum_{x \in T} N(x)$$

where $N(x)$ is the number of examinations before finding that the key $x$ is not in $T$ yet.

**3.9. Corollary.** *For methods $A$, $B$ and $C$ (note that the using of a random key function is assumed).*

(1) $$Q(n, k) = \sum_{i=0}^{k} E(n, i) \,.\, 1/k + 1$$

**3.10. Corollary.** *For the method $C$ and for a random key function*

(1) $$Q_C(n, k) = \frac{k(k+1)}{2n \,.\, k} + 1$$

*therefore*

(2) $$\lim_{n \to \infty} Q_C = d/2 + 1 \,, \quad d = \lim_{n \to \infty} k/n < \infty$$

**3.11. Remark.** Approximate values of $Q_B(n, k)$ for the method $B$ will be given in a subsequent paper.

**3.12. Lemma.** *If a source $A$ generates a sequence of mutually independent random variables $X_1, X_2, \ldots, X_n, \ldots$ with the same discrete distribution and for a table $T$ with parameters $(n, k)$ $M(n, k)$ denotes the mean value of examinations before a key $x$ in $T$ is found (i.e. it is assumed that the key $x$ was already placed in $T$) then for methods $B$ and $C$ (and a random key function)*

$$(1) \qquad M(n, k) = \sum_{x} \sum_{i=1} p_i(x)\, E(n, i)\, P(x \mid x \in T)$$

*where $P(x/x \in T)$ is the probability for $x$ to be generated by the source under the condition that $x$ was already placed in $T$. $p_i(x)$ is the probability that the key $x$ was generated by the source $A$ under the assumption that $i$ different keys have already been produced before.*

Proof. The value of the number of examinations before the key $x$ is found is equal to the number of examinations carried out when the key $x$ was placed into table $T$. Then, however, the same is true for mean values and (1) follows from independence of $x$, because the random key function is used.

**3.13. Remark.** In many situations the table $T$ is previously formed and then used. In this case 3.12.1 remains true, but $P(x/x \in T)$ has different meanings at the time of forming and using.

**3.14. Theorem.** *If for a source the assumptions of* 3.12. *are valid,* $P(x/x \in T)$ *has the same meaning as in* 3.12, *and a random key function is used, then*

$$(1) \qquad M(n, k) \leqq Q(n, k)$$

*and the equality holds for the probability distribution of $X_j$ for which $P(X_j = x) = $ $= c$ where $c$ is a constant independent from $X$ (we say that produced keys are "uniformly" distributed). We assume that a random key function can be constructed for the given probability "distribution" of $X_j$.*

Proof. Let the keys are uniformly distributed and let $T$ contain $k$ keys. Then the probability $p_i(x)$ for $x$ from $T$ is independent on $x$, therefore

$$p_i(x) = 1/k$$

and equality (1) follows from 3.12.1 as $\sum_{x} P(x \mid x \in T) = 1$. In case of the general probability "distribution" of $X_j$ we note that a key $x$ with greater probability is placed into $T$ earlier and is used more often. Consequently, the mean value of examinations cannot be greater than in case of the uniform probability distribution.

**3.15. Remark.** Theorem 3.14. remains true if, during the use of table $T$ (see 3.13), the probability of the generation of a key $x$ is just $P(x \mid x \in T)$ while forming $T$.

**3.16. Theorem.** *For method A the mean value $M(n)$ of the number of examinations, before a vertex labelled x is found, is for a source with uniform probability distribution equal to* $1/2(\log_2 n + 1)$.

Proof. In [7] it was proved that the mean length of path in an admissible tree is $\log_2 n$. But the probability that $x$ labels the $i$-th vertex in a path cannot depend on $x$ so the mean value of examinations is $1/2(\log_2 n + 1)$

**3.17. Remark.** A variant of inequality 3.14.1. for method $A$, i.e. that the uniform distribution of keys is the worst one for searching in the table $T$, is not true. In fact, let the keys which are extreme in ordering have great probabilities. In an admissible tree they would be near to the ends of the paths, so variant of $M(n, k)$ for the method $A$ is near to $\log_2 n$.

## 4. SOME MODIFICATIONS OF THE DESCRIBED METHODS

The main advantage of the method $A$ is the fact that it allows variable length of the table $T$. We shall suggest a variant of methods $B$ and $C$ allowing variable length of table.

**4.1. Remark.** Let us assume that we have a key function $f(x, n)$, where $n = 2^m$ is a sufficiently large power of two and let its values be expressed in binary system, i.e. values of $f(x, n)$ are given by sequences $(d_1, d_2, \ldots, d_m)$ of zeros and ones. Then all the functions $f_i(x, 2^i)$, $1 \leqq i \leqq m$, the values of which in the binary system are given by sequences $(d_{m-i+1}, \ldots, d_m)$ are also random. This fact follows directly from the assumption that $f(x, n)$ is random.

**4.2. Definition.** Let us choose some $d$, $0 < d < 1$. We shall say that a table of length $n$ is overcrowded if it contains $k$ keys and $k > d \,.\, n$. A table $T$ of the length $2^i$ is extended to the table $T'$ of the length $2^{i+1}$ by the following operations:

(i) At the beginning of the algorithm the values of all members of $T'$ are undefined

(ii) The members of $T$ are successively scanned. If a member $T_j$ contains a key $x$ then $f_{i+1}(x, 2^{i+1})$ is evaluated and $x$ is put into $T'$ according to the definition 2.4. or 2.7.

It can be easily shown that $T'$ has the same structure as in case when unmodified methods $B$ or $C$ with the key function $f_{i+1}$ are used for the construction of $T'$ and that the extension of $T$ needs the same number of examinations as forming $T'$ without extension.

**4.3. Remark.** The operation on $T$ described in 4.2. can be modified in the following manner. Let $T = \{T_1, T_2, \ldots, T_n\}$, $T' = \{T'_1, \ldots, T'_{2n}\}$, $T'_i$ be identical with $T_i$ for $i \leqq n$. At the beginning the extension the values of $T'_i$ are identical with the values of $T_i$ for $i \leqq n$ and the values of $T'_i$, $i > n$ are undefined. We now use the fact that, if $f_{i+1}(x, 2^{i+1}) = (0, d_i, \ldots, d_n)$ and if $x$ is the value of $T'_j$ and $T_k$ then $k \geqq j$ so that we can proceed in the following way:

(i) Examine successively $T_1, T_2, \ldots, T_n$.

(ii) If the value of $T_i$ is undefined, examine $T_{i+1}$, else make the value of $T_i$ undefined and a key $x$ which was the value of $T_i$ put into $T'$ according to 2.4. (or 2.7. if method $C$ is used).

It can be shown that this modification preserves all properties of the original algorithm 4.2. We shall call the operation 4.2. (4.3.) the operation of extension.

**4.4. Remark.** Probably any "better" method of extension preserving the properties of structure of nonextended table does not exist.

**4.5. Corollary.** *If — during the construction of table $T$ — the operation of extension was used once and the extension was realized immediately after the table was overcrowded, then for the average price $Q_1^*$ of creation of $T$, i.e. for the mean value per one key in $T$ of examinations made during forming $T$ with parameters $(2^{i+1}, k)$ we have*

$$(1) \qquad Q_1^*(2^{i+1}, k) = \frac{1 + [d \,.\, 2^i]}{k}\, Q(2^i, [d \,.\, 2^i] + 1) + Q(2^{i+1}, k)$$

*where $[\ ]$ denotes the integer part. In case that the operation of extension was used $s$-times we have similary*

$$(2) \qquad Q_s^*(2^{i+s}, k) = \frac{1}{k} \sum_{j=i}^{i+s-1} k_j Q(2^j, k_j) + Q(2^{i+s}, k)$$

*where $k_j = [d \,.\, 2^j] + 1$.*

**4.6. Theorem.** *For great $j$ and a table $T$ with parameters $(2^j, k)$ created with $s$ extensions by methods $B$ or $C$, we have*

$$(1) \qquad (1 + \varepsilon)\, 2(1 - 2^{-s})\, Q(2^j, [d \,.\, 2^j] + 1) + Q(2^j, [d \,.\, 2^j + 1]) \geqq Q_s(2^j, k) \geqq$$
$$\geqq Q(2^j, [d \,.\, 2^{j-1}])$$

*where $[d \,.\, 2^{j-1}] < k \leqq [d \,.\, 2^j]$, $\varepsilon > 0$, $\varepsilon \to 0$ for $j \to \infty$, $s$ is a constant.*

Proof. From 3.10.2. we know that

$$(2) \qquad \lim_{n \to \infty} Q_C(n, [d \,.\, n]) = 1 + d/2$$

The existence of the limit for $Q_B(n, [dn])$ follows the fact that there exists the limit $E_B(n, [dn])$ as the sequence $\{E_B(n)\}$ is bounded and asymptotically not decreasing.

Now we use the fact that

$$(3) \qquad \lim_{j \to \infty} \frac{[d \,.\, 2^{j-r}]}{[d \,.\, 2^j] - 1} = \lim_{j \to \infty} \frac{[d \,.\, 2^{j-r}] + 1}{[d \,.\, 2^j] - 1} = 2^{-r}$$

But $O \leqq Q(n, k) \leqq Q(n, k + 1)$ so if $s$ extensions were carried out then

$$(4) \quad Q(2^j, [d \,.\, 2^{j-1}]) \leqq Q_s^*(2^j, k) = \sum_{i=1}^{s} \frac{[d \,.\, 2^{j-i}] + 1}{k} Q(2^{j-i}, 1 + [d \,.\, 2^{j-i}]) + \\ + Q(2^j, k)$$

and the theorem is proved.

**4.7. Corollary.** For great $j$ the operation of the extension of $T$ increases the price of the construction of table $T$ no more than three times. $M(n, k)$ (see 3.12) remains unchanged.

## 5. SITUATION WHEN A BACKING STORE IS USED

In many situations a table $T$ must be placed on a backing store. Let us assume that an information from a backing store can be called in tracks of the length $L$.

We shall assume that tracks have fixed bounds, i.e. every track always contains the same "location" of a backing store. One call of a track is equivalent to several hundreds or thousands of examinations according to the time needed. Therefore, it is important to find the mean value of the number of the tracks called before an item $x$ in $T$ is found, or before it is established that $x$ is not in $T$ yet.

**5.1. Theorem.** *Let a table $T$ be placed on a backing store and let the tracks have the length $L$. Then the mean value $E^0$ of the number of track calls, before finding that a key $x$ is not value of any member of $T$, is for method $B$ equal to*

$$(1) \qquad 1 + \frac{1}{L} E(n, k) - \frac{1}{L} \sum_{i=1}^{k} i p_i^{k,n}$$

*where $p_i^{k,n}$ has the same meaning as in* 3.1.7. *and $n = s \,.\, L$, $s > 1$, $T$ is in $s$ succeeding tracks.*

Proof. We prove the following lemma:

**5.2. Lemma.** *Let $i$ be uniformly distributed on $\{1 : n\}$ and let $n = s \,.\, L$. Then the mean value $R$ of numbers of $j$ which fulfils the conditions $i \leqq j \leqq i + (q - 1)$,*

*$j = r . L$, where $\dotplus$ denotes addition* mod $n$ *and $\dot{\leqq}$ denotes that $j \leqq i \dotplus q \dot{-} 1$ if $i + q - 1 \leqq n$, nad $j < n$ or $1 \leqq j \leqq i \dotplus q \dot{-} 1$ if $i + q - 1 > n$, fulfils the following equality*

$$R = \frac{q}{L} \tag{1}$$

Proof.

$$R = \frac{1}{n} \sum_{i=1}^{n} \left( \left[ \frac{i + q - 1}{L} \right] - \left[ \frac{i - 1}{L} \right] \right) = \frac{q}{n} . \frac{n}{L} = \frac{q}{L},$$

[ ] denotes the integer part

Proof of 5.1. We remark that if, on carrying out the examinations, some locations $T_i, T_{i \dotplus 1}, \ldots, T_{i \dotplus (q-1)}$ are examined then $i$ is uniformly distributed on $\{1 : n\}$. Using 5.2. in a similar way as in 3.1. we obtain

$$E^0 = 1 + \sum_{r=1}^{k} \sum_{i=r}^{k} \frac{i - 1}{L} p_i^{k,n} = \frac{1}{L} E(n, k) - \frac{1}{L} \sum_{i=1}^{k} i p_i^{k,n} + 1$$

Here $1/L(E(n, k)) - \sum_{i=1}^{k} i p_i^{k,n}$ is the mean value of the number calls of the second track and of the following ones.

**5.3. Remark.** Theorem 5.1. is not true when $n \neq s . L$ or when $T$ is placed so as to cover more tracks than necessary, But. 5.1.1. remains true asymptotically.

**5.4. Corollary.** *For the mean value $E^*$ of track calls before $x$ which is already a member $T$ is found, it is true*

$$E^* = \frac{1}{L} Q(n, k) - \frac{1}{L} \frac{1}{k} \sum_{t=1}^{k} \sum_{i=1}^{t} i p_i^{t,n} \tag{1}$$

Proof. is similar as for 5.1. if we use 3.9.

**5.5. Theorem.** *If $E^0$ or $E^*$ has the same meaning as in* 5.1. *or* 5.4. *and* $\lim_{n \to \infty} (k/n) =$ $= d$ *then for method C and great n*

$$E^0 \doteq 1 + d(1 - e^{-d})$$

*and if the keys are uniformly distributed then*

$$E^* \doteq 1 + \frac{d}{2}(1 - e^{-d/2})$$

*For the method A we have for great n*

$$E^0 \doteq \log_2 n \tag{2}$$

$$E^* \doteq \tfrac{1}{2} \log_2 n \tag{3}$$

Proof. For method $C$ the mean value of the number of keys, having the same value of a key function as another key in $T$ is $k/n$. On the other hand, the number $Q$ of segments is not less than $n/L$ so $\lim_{n\to\infty} Q = \infty$ and $(k/n)/Q \to 0$. The probability that a generated key is in the given track is independent on the key. Therefore the mean value of track calls is asymptotically equal to the mean value of the number of keys having the same value of the key function. Conclusion for $E_B$ results from the fact that the mean value of the number of examinations is $d/2(1 - e^{-d/2})$. (2) and (3) is obtained similarly, because $\lim_{n\to\infty} \log_2 (n)/n = 0$.

**5.6. Remark.** The following variant of using the backing store for the method $B$ or $C$ is possible. A table $T$ with parameters $(n, k)$ is created in the core store. When it is overcrowded $T$ is left unchanged and the new keys are placed into the table $T^0$ which is on a backing store. A key $x$ is found in the following manner. Members of $T$ are examined as decribed on 2.4. or 2.7. by using a random key function $f(x, n)$. If $x$ is not found in $T$, the table $T^0$ is examined by using a random key function $f^0(x, n_0)$. Then the number of examinations before discovering that the key $x$ is not in $T$ and $T_0$ is given by the following expression

$$E(n, k) + E(n_0, k_0) \tag{1}$$

and the number of calls of tracks is less than

$$1 + \frac{1}{L} E(n_0, k_0) \tag{2}$$

but the mean number of examinations before a key $x$ is found is

$$\sum(p(x) (Q(n, k) P_k(x) + (1 - P_k(x)) Q(n_0, k_0)) \tag{3}$$

and the number of track calls is less than

$$\sum_x (1 - P_k(x)) \frac{1}{L} Q(n_0, k_0) p(x) . \tag{4}$$

$P_k(x)$ is the probability that a key $x$ is placed in $T$ under the condition that $T$ contains, $k$ keys, $p(x)$ the probability that $x$ is produced. This modification can give great profit if some $x$ are frequently produced.

**5.7. Remark.** It can be easily verified that one operation of extension needs $N_1$ calls of tracks from a backing store and $N_2$ writings of tracks on a backing store. $N_1$ denotes the number of tracks in which $T$ is placed before extension and $N_2$ has the same meaning for the table after the extension.

**6. Conclusion.** If a random or almost random key function can be constructed, then the best method for a table $T$ in the core store is method $C$. Method $B$, taking into account that for method $C$ a rendundant information "the pointers" is placed in $T$ so the ration (numbers of keys) (number of locations occupied by $T$) is smaller has for $k/n \leqq 0{,}75$ also reasonable properties. For table $T$ on the backing store, the best method is method $B$ which needs for great $n$ and for a great length of tracks (hundreds of locations) practically one call of track per one call of table. The main advantage of method $A$ is that a random key function need not be constructed. For great $n$, however, this method has worse properties than methods $B$ or $C$.

## *References*

[1] *Г. М. Адельсон-Вельский, Е. М. Ландис:* Один алгоритм организации информации, ДАН 146, № 2, (1962).

[2] *А. П. Ершов, Г. И. Кожухин, И. В. Поттосин:* Обзор особенностей альфа-транслятора, Альфа система автоматизации программирования под редакцией А. П. Ершова, Новосибирск 1965 (the English translation of this paper is in *J. of ACM*, Jan. 1966).

[3] *W. W. Peterson:* Adressing for random-access storage, IBM J. Res. and Devel. *4*, No 4, (1957).

[4] Э. *К. Иванова:* О выборе функции расстоновки для организации табличных просмотров, Отчет ВЦ СО АН СССР, Новосибирск 1961.

[5] *К. И. Курбаков:* Способ адресации, использующий сжатые коды слов в качестве адресов памяти, ДАН 163, № 4, 841—844 (1965).

[6] *Shay, G., Raver, N.:* A method for key-to-address transformation, IBM J. Res. and Devel. 7, 121—132 No 2, (1963).

[7] *Л. А. Хиздер:* Некоторые свойства диадических деревев, Ж. В. М. М. Ф. 6, 389—394, № 2 (1966).

[8] *W. Feller:* An Introduction to Probability Theory and its Applications, Vol. 1, J. Willey, New York 1950.

[9] *R. Morris:* Scatter storage techniques, Comm. of ACM 11, 38—44, No. 1, (1968).

## Souhrn

## NĚKTERÉ VELMI RYCHLÉ METHODY HLEDÁNÍ V TABULKÁCH

JAROSLAV KRÁL

V článku jsou studovány tři methody rychlého hledání v tabulkách, jedna založená na konstruování binárního stromu a dvě používající tzv. „náhodnou klíčovou funkci" umožňující nalezení informace v tabulce libovolné délky s dobou vyhledávání shora omezenou číslem nezávislým na délce tabulky. Jsou dány přesné výrazy pro střední hodnoty délky vyhledávání a jsou studovány vlastnosti method při použití vedlejších pamětí a při zvětšování délky tabulky.

*Author's address: Jaroslav Král*, Ústav výpočtové techniky ČSAV a ČVUT, Praha 2, Horská 3.