technical correspondence

On Cichelli's Minimal Perfect Hash Functions Method

☐ In his paper on determining perfect hashing functions Cichelli [1] has not given an overview of those cases for which his method is not applicable. More precisely, he only mentioned that no two keys should agree in their lengths and first and last characters. There are many other conditions under which no perfect hash functions of Cichelli's kind exist, and further, we shall prove the existence of infinitely many key sets for which the above mentioned trivial conditions are not satisfied and for which Cichelli's method is not applicable. In order to find these counterexamples we give a formal description of his approach.

Given is a finite set W of words over an alphabet Σ. Let $\alpha(x)$ denote the first, and $\omega(x)$ the last character of a word $x \in \Sigma^*$. Let $\lambda(x)$ be the length of x, and N the set of nonnegative integers. If $g : \Sigma \to N$ then we define a mapping $\varphi_g : W \to N$ as follows

$$\varphi_g(x) = \lambda(x) + g(\alpha(x)) + g(\omega(x)).$$

Cichelli's approach now seeks a g such that φ_g is a bijective mapping from W onto a subset of N of the form

$$A_S^{(n)} = \{s, s+1, s+2, \ldots, s+n-1\}$$

with $s \in N \wedge n = \text{card } W$.

Trivial conditions for the nonexistence of functions g of the desired kind (admissible functions) are, for instance,

CD1 (Cichelli's condition). There are two words $x, y \in W$ such that

$$\lambda(x) = \lambda(y) \wedge \alpha(x) = \alpha(y) \wedge \omega(x) = \omega(y).$$

CD2 There are two words $x, y, \in W$ such that

$$\lambda(x) = \lambda(y) \wedge \alpha(x) = \omega(y) \wedge \alpha(y) = \omega(x).$$

CD3 $\lambda(x) \equiv \lambda(y) \bmod 2$ for all $x, y, \in W$ and card $\{x \mid x \in W \wedge \alpha(x) = \omega(x)\} > (n+1)/2$ with $n = \text{card } W$.

A small example of English words of constant length where CD1, CD2, CD3 are not satisfied and for which no admissible g exists is the following one

W = {AND, ARE, AIM, ANN, DUE, DIM, DUN, ELM, EON, MAN}.

The smallest counterexample consists of four words

W = {AT, IT, AETHAN, IDENTIFICATION}.

We show now that for any integer n with $5 \le n \le \text{card } \Sigma$ we can find a set W of $\binom{n}{2}$ words for which CD1, CD2, CD3 are not satisfied and where no admissible g exists.

Let $\Sigma = \{\zeta_1, \zeta_2, \ldots, \zeta_6\}$ and $5 \le n \le 6$. Then define a set $W_1^{(n)}$ by

$$W_1^{(n)} = \{\zeta_i\zeta_1\zeta_j \mid 1 \le i < j \le n\}.$$

Clearly, CD1, CD2, CD3 do not hold. We have then for $x \equiv \zeta_i\zeta_1\zeta_j \in W_1^{(n)}$

$$\varphi_g(x) = 3 + g(\zeta_i) + g(\zeta_j).$$

Abbreviate $g(\zeta_i)$ by a_i and assume now that

$$\varphi_g : W_1^{(n)} \to \{s, s+1, s+2, \ldots, s+\binom{n}{2}-1\}$$

for some integer $s \in N$. ζ_j then is bijective and therefore $a_i + a_j \neq a_i + a_k$ for all $j, k \in \{1, \ldots, n\}$, i.e., $a_j \neq a_k$ and we may assume without loss of generality

(1) $a_1 < a_2 < \cdots < a_n$.

Further, we have

(2) $a_i + a_j \neq a_k + a_l$ for all $1 \le i < j \le n$, $1 \le k < l \le n$.

Then there are numbers $j \le n$ such that for all i with $2 \le i \le j$

$$a_i = a_2 + (i-2);$$

for instance $j = 2$ and $j = 3$ (since $a_1 + a_3 = a_1 + a_2 + 1 = s + 1$). Let t be the greatest integer j of this kind, i.e.

$$a_i = a_2 + (i-2) \quad \text{for} \quad 2 \le i \le t$$

and

(3) $a_{t+1} \neq a_2 + t - 1$.

As already pointed out we have $t \ge 3$. If $t \ge 5$ then $a_4 = a_2 + 2$, $a_5 = a_2 + 3$ and $a_2 + a_5 = 2a_2 + 3 = a_3 + a_4$, which contradicts (2). Therefore we have $t = 3$ or $t = 4$.

We demonstrate that both situations ($t = 3$, $t = 4$) are impossible which implies that φ_g cannot be of the supposed type.

a) Let, at first, $t = 3$. Then (1), (2), and (3) imply that

$$\begin{aligned} s &= a_1 + a_2 = 2a_1 + 1 \\ a_2 &= a_1 + 1 \\ a_3 &= a_1 + 2 \\ a_4 &= a_1 + 4 \\ a_5 &= a_1 + 7. \end{aligned}$$

For $n = 5$ we have then

$$s + 9 = 2a_1 + 10 \in A_S^{(10)} - \text{range } (\varphi_g).$$

For $n \ge 6$ we have

$a_6 \neq a_1 + 8$ since otherwise $a_1 + a_6 = a_2 + a_5$

$a_6 \neq a_1 + 9$ since otherwise $a_3 + a_6 = a_4 + a_5$

$a_6 \neq a_1 + 10$ since otherwise $a_2 + a_6 = a_4 + a_5$

and $a_6 \ge a_1 + 11$ which implies $a_i + a_6 > 2a_7 + 10 = s + 9$, i.e., again

$s + 9 \in A_S^{(\eta)} - \text{range } (\varphi_g)$ and so we obtain $t \neq 3$.

b) Let, finally, $t = 4$. Then (1), (2), and (3) imply that

$$\begin{aligned} a_2 &= a_1 + 2 \\ a_3 &= a_1 + 3 \\ a_4 &= a_1 + 4 \\ a_5 &= a_1 + 8. \end{aligned}$$

Thus, for $n = 5$, $s + 7 = 2a_7 + 9 \in A_S^{(10)} - \text{range } (\varphi_g)$. For $n \ge 6$ we have $a_6 \neq a_1 + 9$ since otherwise a_2

$+ a_6 = a_3 + a_5$, and therefore $a_6 \geq a_1 + 10$ which implies $a_i + a_6 > s + 7 = 2a_1 + 9$, i.e. again $s + 7 \in A_S^{(g)} -$ range (φ_g). q.e.d.

By defining generally for every natural number m

$$W_m^{(n)} = \{\zeta_i \xi_1^m \zeta_j / 1 \leq i < j \leq n\}, \quad 5 \leq n \leq \text{card } \Sigma,$$

we observe that also for these infinitely many $W_m^{(n)} (m = 1, 2, \ldots)$ no admissible g which does not satisfy CD1, CD2, CD3 can be found.

Remark:

In order to guarantee the existence of a function g of the desired kind in any case it doesn't help to change α or ω (to denote, for instance, second or fourth letter, respectively) as is easy to be seen. Take for example the set

W = {AABB, AACC, . . . , DDEE} having 10 words.

Conclusion:

Though Cichelli's method often works, it should be realized that it is not foolproof.

G. Jaeschke
G. Osterburg
IBM Scientific Center
Heidelberg, West Germany

1. Cichelli, R. Minimal perfect hash functions made simple. *Comm ACM 23*, 1 (Jan. 1980), 17–19.

Author's Response:

Jaeschke and Osterburg rightly emphasize that my method, though gratifyingly effective, carries no guarantee of success. Perhaps the title should be "Minimal Perfect Hash Functions Made Fairly Simple." [Note: In practice, near minimal perfect hash functions are almost as good as minimal ones.]

In the first column on page 18 of my article, in the second paragraph from the bottom, one might change "Two disadvantages . . . " to "Three disadvantages . . . " and add: "(3) for certain contrived lists of keys, it may be impossible to determine in advance whether this method will yield a minimal solution."

Incidentally, in providing minimal perfect hash functions to several correspondents, I discovered a key set where my heuristics for presearch ordering were clearly suboptimal. The set was the ASCII control code names. Below is the function which was devised by doing the previously recommended orderings and then moving LF and VT to the first and second positions of the list and reapplying the second ordering.

A = 15, B = 3, C = 14, D = 11, E = 16, F = 1, G = 16, H = 10, I = 13, J = 0, K = 17, L = 0, M = 14, N = 6, O = 12, P = 17, Q = 15, R = 9, S = 2, T = 0, U = 14, V = 0, W = 0, X = 5, Y = 0, Z = 0, 1 = 13, 2 = 14, 3 = 17, 4 = 15

VT, LF, FF, FS, BEL, BS, SUB, NUL, STX, SYN, HT, RS, DEL, SOH, SO, SI, US, EOT, GS, SP, ETB, CAN, ETX, CR, NAK, DC1, DC2, DC4, DLE, DC3, EM, ESC, ENQ, ACK

To discover the better order, it was necessary to monitor the search through progressive deepening to locate and move forward the tough words. The program modifications to do this were, of course, trivial.

Richard J. Cichelli
Software Consulting Services
Allentown, PA 18103

Algol-W Approach to Line Number Maintenance

☐ I believe that Paul Klint's article [1] on "Line Numbers Made Cheap" dismisses the mechanism used in an implementation of Algol-W [2] far too lightly, and in doing so, passes up substantially improved diagnostics. The only disadvantage he cites to reject the table, which associates addresses with program sequence numbers, is that it is somehow independent of the code file and, thus, difficult to manage. This is a trivial objection, since the table can easily be stored with the object program itself. In doing this, only a little additional secondary storage is required; there is no impact on hardware design, and the software necessary can be implemented on existing machines.

Burroughs' current facility, which has been in use at least since 1969 when I was a customer, is similar to the Algol-W approach. It displays the sequence number in the symbolic whenever a program fault or abnormal termination occurs. In addition, it displays the sequence number of each procedure call still active in the stack. Thus, it shows not only the error location, but the sequence of calls leading up to this error. There is no execution penalty whatsoever if such a fault does not occur.

Given the trivial drawbacks of the Algol-W approach and its substantial advantages, I fail to see the reason for a more complicated solution to the problem.

John McClintock
Burroughs Corporation
Mission Viejo, CA 92675

1. Klint, P. Line numbers made cheap. *Comm. ACM 22*, (Oct. 1979), 557–559.
2. Satterthwaith, E. Debugging tools for high level languages. *Software Practice and Experience 2*,3 (July–Sept. 1972), 197–217.

Author's Response:

I agree with John McClintock that there are circumstances under which the Algol-W method for the maintenance of line numbers is viable. The reasons for dismissing that method in my paper were:

(a) If the table with line number information is kept in a separate file, then the files with object program and table have to be kept together. Moreover, if the technique is used in a portable system, additional operating system facilities have to be used for creating and accessing the table.

(b) If the table is included in the object program itself, then the size of the object program is increased and additional operations are needed for creating and decoding the contents of the table. Again, in a portable system this may lead to the need for additional abstract machine operations.

Depending on implementation objectives, the above disadvantages of the Algol-W method may or may not be significant.

The method proposed in my paper hardly increases the size of the object program, but it does require (trivial) extensions of the abstract machine.

Paul Klint
Mathematisch Centrum
1098 SJ Amsterdam
The Netherlands