Hash Codes

L.A. Bassalygo[1] †, M. Burmester ‡, A.Dyachkov, G.Kabatianski †[1]

†Institute for Problems of Information Transmission,
Bolshoi Karetniy, 19, Moscow, 101447, Russia; Email: kaba@ippi.ac.msk.su
‡ Dept. of Math., Royal Holloway, Univ. London
Egham, Surrey TW20 0EX, UK; Email: uhah205@vms.rhbnc.ac.uk;
Faculty of Mech.& Math., Moscow State Univ., Moscow, 119899, Russia.

Abstract — **We introduce and investigate the hash distance of codes. In the particular case of perfect hash functions we get new results on the size of such families.**

I. Introduction

Define the s-th hash distance $d_s = d_s(V)$ of a q-ary code V as the maximal number d such that, for any s different codewords there are at least d coordinates every one of which takes distinct values for these codewords. Note that $d_2(V)$ coincides with the usual minimal Hamming distance of the code V and $d_s(V) = 0$ for $s > q$. Enumerate the codewords $V = \{v_1, \dots, v_M\}$ and consider the n coordinates as functions from $\{1, \dots, M\}$ to $\{1, \dots, q\}$. Then a code of length n with $d_s \geq 1$ corresponds to an (M, q, s)-perfect set of n hash functions (see [1]) and vice versa.

Denote $M(n, d_s; q)$ the maximal cardinality of q-ary codes of length n and s-th hash distance at least d_s. In the particular case of perfect hash functions we denote $M(n, d_s = 1; q)$ as $M_s(n; q)$. As usual $R = n^{-1} \log_q M$ denotes the rate of a code. We denote by $k(n, d_s; q)$ the maximal dimension of q-ary linear codes of length n with s-th hash distance at least d_s. Our goal is to find upper and lower bounds for these values.

II. Lower bound

Theorem 1. There exist codes with

$$R \geq (s-1)^{-1} E(\delta_s, \pi_{s,q})(1 + o(1/n)),$$

where $\delta_s = d_s/n$, $\pi_{s,q} = (1 - \frac{1}{q}) \cdots (1 - \frac{s-1}{q})$,
$E(a, b) = g(a, b) - H(a)$, $g(a, b) = -a \log_q b - (1-a) \log_q (1-b)$ and $H_q(a) = g(a, a)$-entropy function.

Note that this bound gives the following "extremal" points: $(R = -(s-1)^{-1} \log_q (1 - \pi_{s,q}), \delta_s = 0)$ - for s-perfect hash functions [2] and $(R = 0, \delta_s = \pi_{s,q})$ - for maximal value of d_s for codes with "exponential size".

For linear codes we prove the following existence bound.

Theorem 2. There exist linear (n, k)-codes with

$$R \geq \min\{\tfrac{E(\delta_s, \pi_{s,q})}{s-1}, \tfrac{E(\delta_s, q^{-1}\binom{s}{2})}{s-2}\}(1 + o(1/n)).$$

This bound coincides with (1) for "sufficiently large" q. On the other hand, for "rather small" q even linear sets of hash functions (i.e. codes with $d_s \geq 1$) do not exists. For example for $q \leq 5$, linear (n, k)-codes with $k \geq 2$ and $d_4 \geq 1$ do not exist.

III. Upper bounds

Theorem 3. $M(n, d_s; q) \leq (s-1) q^{\frac{n - d_s + 1}{s-1}}$.

Corollary. $n^{-1} \log_q M_s(n; q) \leq (s-1)^{-1} + \frac{\log_q s}{n}$.

Hence for s fixed and $q >> s$, the rate R of the best families of hash functions asymptotically equals $(s-1)^{-1}$. Note that the best known upper bounds (see [3-5]) give for this case only $R \leq 1$.

Example. For $q = 7$ and $s = 4$ (i.e. 7-ary 4-perfect hash functions) the bounds of papers [3],[5] and of Th.3 give respectively: $R \leq 0.434$; 0.367; $1/3$. For $q = 100$ and $s = 10$: $R \leq 0.5$; 0.42; $1/9$.

Theorem 4. (Plotkin bound) $M(n, d_s; q) \leq d_s \binom{s}{2} (d_s - n\pi_{s,q})^{-1}$ provided that $d_t > n\pi_{t,q}$.

Together with the recursion $M(n, d_s; q) \leq \frac{q}{s-1} M(n-1, d_s; q)$ we get the following result:

Theorem 5. $M(n, d_s; q) \leq d_s \binom{s}{2} (\frac{q}{s-1})^{n - \frac{d_s - 1}{\pi_{s,q}}}$.

Note that this bound gives the same "end" point $(R = 0, \delta_s = \pi_{s,q})$ as the lower bound.

There is the following recursion for linear codes

$$d_{s-1}(V) - d_s(V) + 1 \geq k(V), \tag{1}$$

where $k(V) = \dim(V)$, which follows from another recursion $k(V) \leq k(n = d_{s'}(V), d_{s-s'+1} = d_s(V); q)$ by putting $s' = s - 1$ and applying the Singleton bound. Note that inequality (3) is not true for general (nonlinear) codes, but becomes valid if we replace $s - 1$ by $s - 2$.

This shows that the introduced notion of hash distance is natural because any set of s-perfect hash functions is automatically a hash code with rather large δ_{s-2} (or even δ_{s-1} - for linear codes). On the other hand, one can use linear hash codes with $\delta_s = const > 0$ just as a set of s-perfect hash functions for which it is simple to evaluate hash functions (because they are linear functionals) and to find a perfect hash function for given s points (because by random choice it takes only constant time to find such a function). The lower bounds (1) and (2) show that for "sufficiently large" q the number of functions increases only in constant time compared with ordinary perfect hash functions.

References

[1] K. Mehlhorn, *Data Structures and Algorithms*, vol. I, Springer-Verlag, 1984.

[2] J. Körner and M. Lucertini, "Compressing inconsistent data," *IEEE Transactions on Information Theory*, vol. 40, pp. 706–715, 1994.

[3] J. Körner and K. Marton, "New bounds for perfect hashing via information theory," *Eur. J. Comb.*, vol. 9, pp. 523–530, 1988.

[4] M.L. Friedman and J. Komlos, "On the size of separating systems and families of perfect hash functions," *SIAM J. Alg. Disc. Methods*, vol. 5, pp. 538–544, 1984.

[5] E. Arikan, "An improved graph-entropy bound for perfect hashing," *Proceedings of 1994 IEEE Int. Symp. Inform. Theory*, p. 314, Trondheim, Norway, 1994.

[1] Part of this work was supported by the Russian Fundamental Research Foundation (project No. 96-01-00884) and Visiting Fellowship grant of the EPSRC