Nicholas Pippenger

On the Application of Coding Theory to Hashing

Quick proofs are given for the characterization (due to Schay, Raver, Hanan, and Palermo) of the collision distance of a linear hashing function and for a dual notion (called the restriction distance), which relates to the accessibility of addresses by sets of keys and the uniform distribution of sets of keys over addresses.

Schay and Raver [1], and also Hanan and Palermo [2], have associated hashing functions with linear algebraic codes. (For a general account of hashing, see Knuth [3]; for algebraic codes, see Berlekamp [4].) They observe that the collision distance of a hashing function (the Hamming distance between the closest pair of keys that hash to the same address) is just the minimum distance of the associated code. The purpose of this note is to point out a dual relationship: the restriction distance of a hashing function, which we define to be the cardinality of the smallest set of key components whose prescription renders some address inaccessible (or, equivalently, causes the resulting set of keys to fail to distribute uniformly over the addresses), is just the minimum distance of the dual code.

We demonstrate these relationships in a way that allows most of the argument to be supplied by well-known facts from coding theory. Consider an n-dimensional vector space $\mathbf{F}^n$ over a finite field $\mathbf{F}$. An *(n,k)-code* $\mathbf{K}$ is simply a k-dimensional subspace of $\mathbf{F}^n$; the elements of $\mathbf{K}$ are called *codewords*. An (n,k)-code can be represented as the image of a linear mapping from $\mathbf{F}^k$ into $\mathbf{F}^n$; the transpose $\mathbf{G}$ of the matrix $\mathbf{G}^T$ of such a mapping is called a *generator matrix* for the code (and the elements of $\mathbf{F}^k$ are called *messages*). Alternatively, an (n,k)-code can be represented as the kernel of a linear mapping from $\mathbf{F}^n$ into $\mathbf{F}^\ell$, where $\ell = n - k$; the matrix $\mathbf{H}$ of such a mapping is called a *parity-check* matrix for the code (and the elements of $\mathbf{F}^\ell$ are called *syndromes*). To summarize, the code $\mathbf{K} = \mathrm{Im}(\mathbf{G}^T) = \mathrm{Ker}(\mathbf{H})$ is the ham in the exact sandwich

$$\mathbf{0} \to \mathbf{F}^k \xrightarrow{\mathbf{G}^T} \mathbf{F}^n \xrightarrow{\mathbf{H}} \mathbf{F}^\ell \to \mathbf{0}.$$

The *minimum distance* $d(\mathbf{K})$ of a code $\mathbf{K}$ is the Hamming distance between the closest pair of codewords. It is the cardinality of the smallest set of columns of $\mathbf{G}$ whose deletion renders the rows of the resulting matrix linearly dependent. Alternatively, it is the cardinality of the smallest set of linearly dependent columns of $\mathbf{H}$.

If the elements of $\mathbf{F}^n$ are regarded as *keys* and the elements of $\mathbf{F}^\ell$ are regarded as *addresses*, the linear mapping determined by $\mathbf{H}$ can be used as a *hashing function*. The *collision distance* of such a hashing function is defined as the Hamming distance between the closest pair of keys that hash to the same address. If two addresses hash to the same address, the set of key components in which they differ must correspond to a linearly dependent set of columns of $\mathbf{H}$. Thus the collision distance is just $d(\mathbf{K})$, the minimum distance of the associated code.

Corresponding to each (n,k)-code $\mathbf{K}$ is an (n,ℓ)-code $\mathbf{L}$ called the *dual* code. As a subspace of $\mathbf{F}^n$, it is the orthogonal complement of $\mathbf{K}$. It is obtained by dualizing everything in sight: $\mathbf{L} = \mathrm{Ker}(\mathbf{G}) = \mathrm{Im}(\mathbf{H}^T)$ in the exact sequence

$$\mathbf{0} \leftarrow \mathbf{F}^k \xleftarrow{\mathbf{G}} \mathbf{F}^n \xleftarrow{\mathbf{H}^T} \mathbf{F}^\ell \leftarrow \mathbf{0}.$$

If $\mathbf{H}$ is used as a hashing function, its *restriction distance* is defined as the cardinality of the smallest set of key components whose specification renders some address inaccessible. But such a set of key components must correspond to a set of columns of $\mathbf{H}$ whose deletion renders the rows of the resulting matrix linearly dependent. Thus, the restriction distance is just $d(\mathbf{L})$, the minimum distance of the dual code.

Finally, we observe that the restriction distance is also the cardinality of the smallest set of key components whose specification causes the resulting set of keys to fail to distribute uniformly over the addresses. This is because the fiber of every element in the image of a linear mapping is a translate of the kernel, so the cardinalities of all these fibers are equal. Thus, if every address is accessible, the keys distribute uniformly over the addresses.

References

1. G. Schay and N. Raver, "A Method for Key-to-Address Transformation," *IBM J. Res. Develop.* **7**, 121 (1963).
2. M. Hanan and F. P. Palermo, "An Application of Coding Theory to a File Address Problem," *IBM J. Res. Develop.* **7**, 127 (1963).
3. D. E. Knuth, *The Art of Computer Programming*, Vol. 3 *(Sorting and Searching)*, Addison-Wesley Publishing Co., Reading, MA, 1973, Section 6.4.
4. E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill Book Co., New York, 1968.

Received September 1, 1978

The author is located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.