Discrete Applied Mathematics 28 (1990) 171–176
North-Holland

NOTE

PERFECT ERROR-CORRECTING DATABASES

Zoltán FÜREDI*
Mathematical Institute of the Hungarian Academy of Sciences, 1364 Budapest, P.O. Box 127, Hungary

Received 3 January 1989

An $n\times m$ matrix is called a t-error-correcting database if after deleting any t columns one can still distinguish the rows. It is perfect if after omitting any $t+1$ columns two identical rows are obtained. (Stating with another terminology, the system of minimal keys induced by A is the system of all $(n-t)$-element subsets of an n-element set.)

Let $f_t(n)$ denote the minimum number of rows in a perfect t-error-correcting database of length n. We show that $f_2(n)=\Theta(n^2)$, and in general $\Omega(n^{(2t+1)/3})\le f_t(n)\le O(n^t)$ for $t\ge 3$, whenever $n\to\infty$.

1. Preliminaries

Let $n>t\ge 0$ be integers. A set V of sequences of length n (or the matrix A formed by these sequences, as rows) is called a *database*. A sequence $\alpha\in V$ can be considered as a function $\alpha: X\to Y$, where sometimes X is identified by the set of the first n integers, $X=[n]$. $D(\alpha,\beta)$ denotes the set of distinct coordinates, $D(\alpha,\beta)=\{i: \alpha(i)\ne\beta(i)\}$. The Hamming distance, $H(\alpha,\beta)$, of two sequences $\alpha,\beta\in V$ is the number of distinct coordinates, $H(\alpha,\beta)=|D(\alpha,\beta)|$. V is *t-error-correcting* if the Hamming distance between any two sequences is greater than t. In other words, after deleting any t of the columns of A, one can still distinguish the rows.

A is *perfect t-error-correcting* if the deletion of any $t+1$ columns leads to identical rows, i.e., for all $T\subset X$, $|T|\ge t+1$ one can find $\alpha,\beta\in V$, $\alpha\ne\beta$ such that $\alpha(i)=\beta(i)$ for all $i\in X\setminus T$. (Warning! This definition differs from the usual one concerning error-correcting *codes*.)

For $t\ge 1$ one can define a perfect t-error-correcting database as follows (see [1, 3]). Let $E_1, E_2, \ldots$ denote the t-element subsets of X, and

$$\alpha_j(i)=\begin{cases}0, & \text{if } i\notin E_j,\\ j, & \text{if } i\in E_j.\end{cases} \tag{1.1}$$

Then $V^{(t)}=\{\alpha_j: 1\le j\le \binom{n}{t}\}$ is a t-error-correcting database. Moreover, $V^{(1)}$ completed with the full 0-sequence forms a perfect 0-error-correcting database.

* Research supported partly by the Hungarian National Science Foundation grant No. 1812.

Denote by $f_t(n)$ the minimum number of sequences in a perfect t-error-correcting database of length n. It was proved in [5] that $f_0(n)=n+1$, $f_1(n)=n$ and the extremal databases are isomorphic to the above example. For $t\geq 1$ we have

$$f_t(n)\leq\binom{n}{t} \tag{1.2}$$

by (1.1). On the other hand, it is easy to see [5] that

$$f_t(n)>\sqrt{2\binom{n}{t+1}}=\Omega(n^{(t+1)/2}). \tag{1.3}$$

The aim of this paper is to give a better lower bound for $f_t(n)$, namely for $t=2$ we will show that $f_2(n)=\Theta(n^2)$.

2. A note on relational databases

Every database V determines a closure operation, a system of functional dependencies. (The definition of relational databases see [2], or the book of [8]. More definitions and notions see [9].) One of the basic notions is the following. A set $N\subset X$ is called a *nonkey* if it could not distinguish between the distinct rows, i.e., there are $\alpha, \beta\in V$, $\alpha\neq\beta$ such that $\alpha\mid N=\beta\mid N$. Otherwise the set is called a *key*. The family of maximal nonkeys is denoted by $\mathscr{N}$.

Let $\mathscr{K}_k^n$ be the hypergraph of the all k-subsets of an n-set. The determination of $f_t(n)$ is equivalent to the following problem. What is $s(\mathscr{K}_{n-t}^n)$, the minimum number of rows in a matrix A inducing $\mathscr{K}_{n-t}^n$ as a system of minimal keys? It is known [6, 4], that if k is fixed and n tends to infinity, then the order of magnitude of the lower bound in (1.3) is correct, namely $s(\mathscr{K}_k^n)=\Theta(n^{(k-1)/2})$.

3. Results

Theorem 3.1. $\frac{1}{12}n^2<f_2(n)<\frac{1}{2}n^2$.

Theorem 3.2. *For* $n>n_0(t)$, $t\geq 1$ *one has*

$$\frac{1}{(t+1)!}n^{(2t+1)/3}<f_t(n)<\frac{1}{t!}n^t.$$

Using the terminology of relational databases we can obtain the more general theorem:

Theorem 3.2′. *Suppose that V is a database inducing $\mathcal{N}$ as the system of maximal nonkeys,* $\max\{|N|: N\in\mathcal{N}\}=n-t+1$, $\mathcal{N}_i=\{N\in\mathcal{N}: |N|=i\}$. *Then*

$$|V|\geq\frac{1}{(t+1)!}|\mathcal{N}_{n-t+1}|^{2/3}n^{-1/3}.$$

4. Lemmas. The structure of the minimum distance graph

Let $\mathcal{G}$ be a graph with the vertex set V. (Usually we identify a graph with its edge set.) Suppose that $\mathcal{G}$ does not contain the complete bipartite graph $\mathcal{K}(2, p)$ as a subgraph. Then for the number of edges $e(\mathcal{G})$ we have that

$$e(\mathcal{G})<\sqrt{\frac{p-1}{2}}\,|V|^{3/2}+\frac{|V|}{2}, \tag{4.1}$$

as it was shown by Kővári, Sós and Turán [7].

Let V be a perfect t-error-correcting database of length n. Define the *minimum distance graph*, $\mathcal{G}$, with the vertex set V as follows. (α, β) is joint by an edge if and only if $H(\alpha, \beta)=t+1$. The edge $(\alpha, \beta)\in\mathcal{G}$ has color D (where $D\subset X$, $|D|=t+1$) if $D=D(\alpha, \beta)$. We will use the standard notations of graph theory, i.e., $\deg(\alpha, \mathcal{G})$ (or briefly $\deg(\alpha)$) stands for the *degree*. $\Gamma(\alpha, \mathcal{G})$ (or briefly $\Gamma(\alpha)$) denotes the *neighborhood* of α.

For every $(t+1)$-element set $D\subset X$ choose an edge of $\mathcal{G}$ of color D. These $\binom{n}{t+1}$ (or in the proof of Theorem 3.2′, these $|\mathcal{N}_{n-t+1}|$) pairs form the *reduced* minimum distance graph $\mathcal{G}_0$. This graph is not necessarily unique in general. The notations $\deg_0$, Γ_0 indicate that we are speaking about $\mathcal{G}_0$.

Lemma 4.1. *Suppose that $(\alpha, \beta)\notin\mathcal{G}$. Then $|\Gamma(\alpha)\cap\Gamma(\beta)|<3^{2t+2}$.*

Lemma 4.2. *For all α, β one has $|\Gamma_0(\alpha)\cap\Gamma_0(\beta)|<n3^t$.*

Proofs of Lemmas 4.1 and 4.2. We prove these lemmas simultaneously. Let $C=\{x\in X: \alpha(x)\neq\beta(x)\}$, i.e., $|C|=H(\alpha, \beta)$. We may suppose that $|C|\leq 2t+2$, otherwise both lemmas are trivial, there is no $\gamma\in V$ with $\{\alpha, \gamma\}\in\mathcal{G}$, $\{\beta, \gamma\}\in\mathcal{G}$.

Suppose that $\gamma, \gamma'\in\Gamma(\alpha)\cap\Gamma(\beta)$, $\gamma\neq\gamma'$. As $\gamma(x)$ differs from $\alpha(x)$ or $\beta(x)$ for all $x\in C$ we have

$$C\subset D(\alpha, \gamma)\cup D(\beta, \gamma). \tag{4.2}$$

Moreover $D(\alpha, \gamma)\setminus C=D(\beta, \gamma)\setminus C$.

Proposition 4.3. *Suppose that $D(\alpha, \gamma)\cap C=D(\alpha, \gamma')\cap C$ and $D(\beta, \gamma)\cap C=D(\beta, \gamma')\cap C$. Then $|C|=t+1$, and $(D(\alpha, \gamma)\setminus C)\cap(D(\alpha, \gamma')\setminus C)=\emptyset$.*

This proposition says that γ is (almost) determined by the traces of $D(\alpha,\gamma)$ and $D(\beta,\gamma)$ on C.

Proof. (4.2) gives

$$D(\gamma,\gamma')\subset D(\alpha,\gamma)\cup D(\alpha,\gamma').$$

γ and γ' agree on $C\setminus(D(\alpha,\gamma)\cap D(\beta,\gamma))$, hence we have

$$H(\gamma,\gamma')\le |D(\alpha,\gamma)\cap D(\beta,\gamma)\cap C|+|D(\alpha,\gamma)\setminus C|+|D(\alpha,\gamma')\setminus C|. \qquad (4.3)$$

As $|D(\alpha,\gamma')\setminus C|=|D(\beta,\gamma)\setminus C|$ we have that the right-hand side of (4.3) equals $2(t+1)-|C|$. If $|C|>t+1$, then this leads to the contradiction $H(\gamma,\gamma')<t+1$. So $|C|=t+1$, and equality holds in (4.3). Thus the sets $D(\alpha,\gamma)\setminus D$ and $D(\alpha,\gamma')\setminus D$ are disjoint. □

Proof of Lemma 4.1. Proposition 4.3 implies that the number of $\gamma\in\Gamma(\alpha)\cap\Gamma(\beta)$ is not more than the number of set pairs $A,B\subset C=D(\alpha,\beta)$ such that $|A|=|B|$, $A\cap B=\emptyset$ and $|A|\ge|C|-t-1$. (Here $A=D(\alpha,\gamma)\setminus D(\beta,\gamma)$ and $B=D(\beta,\gamma)\setminus D(\alpha,\gamma)$.) Hence

$$|\Gamma(\alpha)\cap\Gamma(\beta)|\le\sum_{i\ge|C|-t-1}\binom{|C|}{2i}\binom{2i}{i}<3^{|C|}\le 3^{2t+2}. \qquad \square$$

Note that in the case $t=2$ we obtain the following bounds

$$|\Gamma(\alpha)\cap\Gamma(\beta)|\le\begin{cases}18, & \text{if } |C|=4,\\ 30, & \text{if } |C|=5,\\ 20, & \text{if } |C|=6=2t+2.\end{cases} \qquad (4.4)$$

Proof of Lemma 4.2. Let again $C=D(\alpha,\beta)$. Now $|C|=t+1$. Choose the subsets $A,B\subset C$ such that $|A|=|B|=i$, $A\cap B=\emptyset$, and consider all $\gamma\in\Gamma_0(\alpha)\cap\Gamma_0(\beta)$ with $A=D(\alpha,\gamma)\setminus D(\beta,\gamma)$ and $B=D(\beta,\gamma)\setminus D(\alpha,\gamma)$. For an arbitrary $\gamma\in\Gamma_0(\alpha)$ we have $D(\alpha,\gamma)\setminus C\ne\emptyset$, by the definition of $\mathscr{G}_0$, so $i\ge1$. Proposition 4.3 implies that the number of such γ is at most $(n-|C|)/i$. Hence

$$|\Gamma(\alpha)\cap\Gamma(\beta)|\le\sum_{i\ge1}\binom{|C|}{2i}\binom{2i}{i}\frac{n-t-1}{i}$$

$$<n\sum_{i>1}\binom{t+1}{2i}\binom{2i}{i}\frac{1}{i}<n3^t. \qquad \square$$

5. The proof of Theorem 3.2

Consider the graph $\mathscr{G}_0$ defined in Section 4. Lemma 4.2 gives that $\mathscr{G}_0$ does not contain a complete bipartite graph $\mathscr{K}(2,n3^t)$. Then (4.1) yields that

$$\binom{n}{t+1} = |\mathcal{G}_0| \le \sqrt{\frac{n3^t}{2}}\,|V|^{3/2} + \frac{|V|}{2},$$

implying Theorem 3.2.

The proof of Theorem 3.2′ is similar.

6. The proof of Theorem 3.1

Consider the graph $\mathcal{G}_0$. We are going to estimate $\sum_{\alpha,\beta\in V}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)|$. By Jensen's inequality

$$\sum_{\alpha,\beta\in V}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)| = \sum_{\alpha\in V}\binom{\deg_0(\alpha)}{2}$$

$$\ge |V|\binom{2e(G_0)/|V|}{2} = 2\frac{e(G_0)^2}{|V|} - e(G_0). \tag{6.1}$$

To obtain an upper bound we split the sum into two parts. First, Lemma 4.1, more exactly (4.4) gives that

$$\sum_{\substack{\alpha,\beta\in V\\ \{\alpha,\beta\}\notin\mathcal{G}}}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)| \le \left(\binom{|V|}{2} - |\mathcal{G}|\right)30 < 15|V|^2. \tag{6.2}$$

Rearranging the rest of the sum we have

$$\sum_{\substack{\alpha,\beta\in V\\ \{\alpha,\beta\}\in\mathcal{G}}}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)| = \sum_{\substack{C\subset X\\ |C|=t+1}}\left[\sum_{\substack{\alpha,\beta\in V\\ D(\alpha,\beta)=C}}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)|\right]. \tag{6.3}$$

Proposition 6.1. *For any fixed $C\subset X$, $|C|=t+1$ one has*

$$\sum_{\substack{\alpha,\beta\in V\\ D(\alpha,\beta)=C}}|\Gamma_0(\alpha)\cap\Gamma_0(\beta)| < 3^t\binom{n}{\lfloor (t+1)/2\rfloor}.$$

Proof. The left-hand side is the number of triples α, β, γ such that $\{\alpha,\gamma\}$, $\{\beta,\gamma\}\in\mathcal{G}_0$ and $D(\alpha,\beta)=C$. Associate to this triple the sets $D(\alpha,\gamma)$, $D(\beta,\gamma)$. These sets determine α, β and γ, hence, by Lemma 4.1, the left-hand side in Proposition 6.1 is not more than the number of pairs $A,B\subset X$ such that $|A|=|B|=t+1$, $A\cup B\supset C$ and $A\setminus C=B\setminus C\ne\emptyset$. The number of such pairs is at most

$$\frac{1}{2}\sum_{j\ge 1}\binom{n-t-1}{j}\binom{t+1}{2j}\binom{2j}{j} < \binom{n}{\lfloor (t+1)/2\rfloor}3^t. \tag{6.4}$$

In the case $t=2$ the left-hand side of (6.4) is less than $3n$. □

Hence the right-hand side of (6.3) is at most $\binom{n}{3}3n$. Using this and (6.1), (6.2) we have that

$$2\frac{\binom{n}{3}^2}{|V|}-\binom{n}{3}\le 15|V|^2+\binom{n}{3}3n.$$

This inequality gives $|V|>0.089\ldots n^2$.

The combination of the proofs might give $f_t(n)\ge\Omega(n^{(2t+2)/3})$, but the real question is that whether the upper bound $O(n^t)$ can be decreased.

References

[1] W.W. Armstrong, Dependency structures of database relationship, Inform. Process. Lett. 74 (1974) 580–583.

[2] E.F. Codd, A relational model of data for large shared data banks, Comm. ACM 13 (1970) 377–387.

[3] J. Demetrovics, Candidate keys and antichains, SIAM J. Algebraic Discrete Methods 1 (1980) 92.

[4] J. Demetrovics, Z. Füredi and G.O.H. Katona, Minimum matrix representation of closure operations, Discrete Appl. Math. 11 (1985) 115–128.

[5] J. Demetrovics and G.O.H. Katona, Extremal combinatorial problems in relational data base, in: Lecture Notes in Computer Science 117 (Springer, Berlin, 1981) 110–119.

[6] Z. Füredi, Minimum relational databases, Alkalmaz. Mat. Lapok 9 (1983) 23–28 (in Hungarian).

[7] T. Kővári, V.T. Sós and P. Turán, On a problem of K. Zarankiewicz, Colloq. Math. 3 (1959) 50–57.

[8] D. Maier, The Theory of Relational Databases (Computer Science Press, Rockville, MD, 1987).

[9] R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, in: I. Rival, ed., Ordered Sets (Reidel, Dordrecht, 1982).