Error-correcting keys in relational databases

János Demetrovics
Comp. and Autom. Institute
Hungarian Academy of Science
Kende u. 13-17, H-1111, Hungary
dj@ilab.sztaki.hu

Gyula O.H. Katona and Dezső Miklós
Alfréd Rényi Institute of Mathematics
Hungarian Academy of Sciences
Budapest P.O.B. 127 H-1364 Hungary
ohkatona@renyi-inst.hu, dezso@renyi-inst.hu

February 10, 2012

Abstract

Suppose that the entries of a relational database are collected in an unreliable way, that is the actual database may differ from the true database in at most one data of each individual. An error-correcting key is such a set of attributes, that the knowledge of the actual data of an individual in this set of attributes uniquely determines the individual. It is showed that if the minimal keys are of size at most k, then the smallest sizes of the minimal error-correcting keys can be ck^3 and this is the best possible, all minimal error-correcting keys have size at most $3k^3$.

1 Introduction

A database can be considered as an $m \times n$ matrix M, where the rows are the data of one individual, the data of the same sort (*attributes*) are in the same column. Denote the set of attributes (equivalently, the set of columns of the matrix) by Ω, its size is $|\Omega| = n$. It will be supposed that the data of two distinct individuals are different, that is, the rows of the matrix are different. A subset K of Ω is called a *key* if the data in K determine the individual (row) uniquely. In other words, there are no two distinct rows of the matrix which are equal in K. A key is a *minimal key* if its no proper subset is a key. Denote the family of all minimal keys by $\mathcal{K}$.

Suppose that the data are collected in a non-reliable way, that is, at most e of the data of each individual can be incorrect. Let M denote the matrix of the real data and M^* ($m \times n$, again) the collected ones. We know that M and M^* differ in at most e entries in each row. Although it is here also supposed that the real data of two distinct individuals are different, that is the rows of M are different, this cannot be stated about M^*. Moreover a key K cannot determine the row if the entries of the unreliable matrix M^* are given in the columns belonging to K. In the present paper we will investigate such sets of attributes (columns) which uniquely determine the individual from M^*. We say that C is an *e-error-correcting key* if it has this property, that is, knowing the entries of M^* in the columns belonging to C, the individual (and its row in M) can be uniquely determined.

The number of different entries in two rows is called the *Hamming distance* of these two rows. The $m \times |C|$ submatrix of M determined by the set C of its columns is denoted by $M(C)$. If the Hamming distance of any two rows of $M(C)$ is at least $2e + 1$ then the Hamming distance of any two rows of $M^*(C)$ is at least 1, that is, knowing the entries of the unreliable matrix in C it determines the row uniquely, C is an e-error correcting key. The converse is true, too: if the Hamming distance of two rows of $M(C)$ is at most $2e$ then it may happen that the rows are equal in $M^*(C)$, that is, C is not an e-error-correcting key. We obtained the following proposition.

Proposition 1.1 *$C \subset \Omega$ is an e-error-correcting key iff the pairwise Hamming distance of the rows of $M(C)$ is at least $2e + 1$.* □

It is easy to see that if the pairwise Hamming distance of the rows of $M(C)$ is at least $2e$ then the knowledge of $M^*(C)$ detects the error, but does

not determine the row uniqely. This case is less interesting, but it makes worth introducing the more general definition: $C \subset \Omega$ is called a *d-distance key* iff the pairwise Hamming distance of the rows of $M(C)$ is at least d.

The main aim of the present investigations is to find connections between the family of keys and the family of d-distance keys. The next proposition is the first step along this line.

Proposition 1.2 *$C \subset \Omega$ is a d-distance key iff for any choice $a_1, \ldots, a_{d-1} \in C$ one can find a $K \in \mathcal{K}$ such that $K \subset C - \{a_1, \ldots, a_{d-1}\}$.*

Proof. The necessity will be proved in an indirect way. Suppose that there exist $a_1, \ldots, a_{d-1} \in \Omega$ such that $C - \{a_1, \ldots, a_{d-1}\}$ contains no member of $\mathcal{K}$, that is, $C - \{a_1, \ldots, a_{d-1}\}$ is not a key. Therefore there are two distinct rows of M which are equal in $M(C - \{a_1, \ldots, a_{d-1}\})$. The Hamming distance of these two rows in $M(C)$ is less than d. This contradiction completes this part of the proof.

To prove the sufficiency suppose, again in an indirect way, that $M(C)$ contains two distinct rows with Hamming distance $< d$. Delete those columns where these columns are different. We found a set $C - \{a_1, \ldots, a_{d-1}\}$ satisfying the condition that $M(C - \{a_1, \ldots, a_{d-1}\})$ contains two distinct rows which are equal everywhere, therefore $C - \{a_1, \ldots, a_{d-1}\}$ is not a key in M, it cannot contain a member of $\mathcal{K}$. □

It is easy to see that the family $\mathcal{K}$ of minimal keys is non-empty and *inclusion-free*, that is, $K_1, K_2 \in \mathcal{K}, K_1 \neq K_2$ implies $K_1 \not\subset K_2$. On the other hand, it is known ([1], [2]) that there is a database for any non-empty inclusion-free family $\mathcal{K}$ in which this is the family of all minimal keys. This is why it is sufficient to give a non-empty inclusion-free family rather than constructing the complete database or matrix. Note that, by Proposition 1.2, $\mathcal{K}$ and d determine $\mathcal{C}_d$, therefore the notation $\mathcal{C}_d(\mathcal{K})$ will be used, if it is necessary to emphasize that $\mathcal{C}_d$ is generated by $\mathcal{K}$.

Our first observation is that it may happen that there is no d-distance key at all. Fix an element $a \in \Omega$ (that is, a column) and an integer $2 \leq k$. Define $\mathcal{K}$ as the family of all k-element sets ($\subset \Omega$) containing a. Then $C - \{a\}$ cannot contain any key, so the condition of Proposition 2 does not hold for any C if $2 \leq d$: there is no d-distance key in this database for $2 \leq d$.

On the other hand, if $\mathcal{K}$ consists of all k-element subsets of Ω then all sets C with at least $k + d - 1$ elements are d-distance keys. In the case when

there are d-distance keys, it is enough to consider the minimal ones. Let $\mathcal{C}_d$ denote the family of all minimal d-distance keys. Our last example suggests that the sizes of the members of $\mathcal{C}_d$ do not exceed the sizes of the members of $\mathcal{K}$ by too much. We will show that this is not really true.

Now we introduce some notations. Let $\binom{\Omega}{\leq k}$ denote the family of all subsets of Ω with size not exceeding k. Furthermore

$$f_1(\mathcal{K}, d) = \min\{|C| : C \in \mathcal{C}_d(\mathcal{K})\},$$

$$f_2(\mathcal{K}, d) = \max\{|C| : C \in \mathcal{C}_d(\mathcal{K})\},$$

$$f_i(n, k, d) = \max_{\mathcal{K} \subset \binom{\Omega}{\leq k}} f_i(\mathcal{K}, d).$$

We will prove the following theorem in Section 2.

Theorem 1.3

$$c_1 k^d \leq f_1(n, k, d) \leq f_2(n, k, d) \leq c_2 k^d$$

holds for $n_0(k, d) \leq n$ where c_1 and c_2 depend only on d.

Section 3 contains suggestions how to continue this research.

2 The proof

Let $\mathcal{K}$ be a family of subsets of Ω. We say that the elements $a_1, \ldots a_{d-1} \in \Omega$ *represent* $\mathcal{K}$ if each $K \in \mathcal{K}$ contains one of the as. Proposition 1.2 can be said in the form that $\mathcal{C} \subset \Omega$ is a d-distance key iff no $d-1$ elements can represent the family $\{K : \ K \in \mathcal{K}, K \subset C\}$. If C is minimal with respect to this property then no proper subset of C has the above property, that is, for all $a \in C$ the family $\{K : \ K \in \mathcal{K}, K \subset C - \{a\}\}$ can be represented by $d-1$ elements. This gives a new variant of Proposition 1.2:

Proposition 2.1 *$C \in \mathcal{C}_d$ iff $\{K : \ K \in \mathcal{K}, K \subset C\}$ cannot be represented by $d-1$ elements, but it can be represented by d elements $a, a_1, \ldots, a_{d-1}$ where a can be given arbitrarily in C, in advance.*

□

Lower estimate. We give a non-empty, inclusion-free family $\mathcal{K}$ consisting of k-element sets which generates a $\mathcal{C}_d$ consisting of one member having size at least ck^d.

Fix an integer $1 \leq i$ and take a subset $A \subset \Omega$ of size $i+d-1$. Let $A_1, A_2, \ldots$ be all the $\binom{i+d-1}{i}$ i-element subsets of A and

$$\mathcal{K}(i) = \{A_1 \cup B_1, A_2 \cup B_2, \ldots\},$$

where $A, B_1, B_2, \ldots$ are pairwise disjoint and $|B_1| = |B_2| = \cdots = k-i$. This can be carried out if

$$i + d - 1 + \binom{i+d-1}{i}(k-i) \leq n. \tag{2.1}$$

Show that the only member of $\mathcal{C}_d(\mathcal{K}(i))$ is $C = A \cup \cup_i B_i$. It is easy to see that $\mathcal{K}(i)$ cannot be represented by $d-1$ elements. On the other hand, if $a \in B_j$ for some j then the d-element $\{a\} \cup (A - A_j)$ represents $\mathcal{K}$. If, however, $a \in A$ then any d-element $D \subset A$ containing a represents $\mathcal{K}$, therefore C is really a member of $\mathcal{C}_d(\mathcal{K}(i))$. It is easy to see that there is no other member.

Choose $i = \left\lfloor k(1 - \frac{1}{d}) \right\rfloor$. Then the size of C, given by the left hand side of (2.1) asymptotically becomes

$$\frac{(d-1)^{d-1}}{d^d(d-1)!}k^d.$$

□

Upper estimate. Let $C \in \mathcal{C}_d(\mathcal{K})$ where $\mathcal{K} \subset \binom{\Omega}{\leq k}$. We will prove that $|C| \leq dk^d$. Since we have to consider only the subsets of C, so it can be supposed that all members of $\mathcal{K}$ are subsets of C.

Proposition 2.1 defines d-element subsets D of C each of them is representing $\mathcal{K}$. Moreover, still by Proposition 2.1, their union is C. Denote this family by $\mathcal{D}$. We know

$$\cup_{K \in \mathcal{K}} = \cup_{D \in \mathcal{D}} = C, \tag{2.2}$$

$$D \cap K \neq \emptyset \text{ for all } D \in \mathcal{D}, K \in \mathcal{K} \tag{2.3}$$

and $\mathcal{K}$ cannot be represented by a set with less than d element.

Let $I \subset C$. Define the I-degree of $\mathcal{D}$ as the number of members of $\mathcal{D}$ containing I, that is,

$$\deg_I(\mathcal{D}) = |\{D \in \mathcal{D} : I \subset D\}|.$$

Lemma 2.2

$$\deg_I(\mathcal{D}) \leq k^{d-|I|}.$$

Proof. We use induction on $j = d - |I|$. Suppose that $j = d - |I| = 1$, that is, $|I| = d - 1$. If all members of $\mathcal{K}$ meet I then $\mathcal{K}$ can be represented by $d - 1$ elements, a contradiction. Therefore there is a $K \in \mathcal{K}$ which is disjoint to I. By (2.3) all the sets D satisfying $I \subset D$ must intersect this K, therefore their number is $\leq |K| \leq k$. This case is settled.

Now suppose that the statement is true for $j = d - |I| \geq 1$ and prove it for $j + 1 = d - |I|$. Let $|I^*| = d - j - 1$. There must exist a $K \in \mathcal{K}, K \cap I^* = \emptyset$ otherwise $\mathcal{K}$ is represented by less than d elements, a contradiction. Let $K = \{x_1, \ldots, x_l\}$ where $l \leq k$. By (2.3) we have

$$\{D \in \mathcal{D} : \ I^* \subset D\} = \cup_{i=1}^{l} \{D \in \mathcal{D} : \ (I^* \cup \{x_i\}) \subset D\}. \tag{2.4}$$

The sizes of the sets on the right hand side are $\deg_{I^* \cup \{x_i\}}(\mathcal{D})$ which are at most k^{d-j} by the induction hypothesis. Using (2.4)

$$\deg_{I^*}(\mathcal{D}) \leq lk^{d-j} \leq k^{d-j+1}$$

is obtained, proving the lemma. □

Finally, consider any $K = \{y_1, \ldots, y_r\} \in \mathcal{K}$ where $r \leq k$. By (2.3), the families $\{D \in \mathcal{D} : y_i \in D\}$ cover $\mathcal{D}$. Apply the lemma for $I = \{y_i\}$:

$$\{D \in \mathcal{D} : y_i \in D\} \leq k^{d-1}.$$

This implies $|\mathcal{D}| \leq k^d$ and

$$|\cup_{D \in \mathcal{D}} D| \leq |\mathcal{D}| d \leq dk^d.$$

Application of (2.2) completes the proof: $|C| \leq dk^d$. □

Let us emphasize the simplest case when the probability of an incorrect data is so small that practically at most one data of an individual can be incorrect. In this case $e = 1, d = 3$, therefore, if the minimal keys have at most k elements, then the 1-error-correcting keys have at most $3k^3$ elements, and this is sharp up to a constant factor. So even in this simple case, the error-correcting keys may be much larger then the keys.

3 Further problems

1. Although Theorem 2.1 determines the order of magnitude of $f_1(n,k,d)$, it does give the exact value. We believe that the lower estimate is sharp.

Conjecture 3.1

$$f_1(n,k,d) = \max_i \{i + d - 1 + \binom{i+d-1}{i}(k-i)\}.$$

holds for $n_0(k,d) \leq n$.

2. Knowing $\mathcal{K}$, can we effectively compute $\mathcal{C}_d$ (for a given d)? If k is fixed, then Theorem 1.3 shows that the problem can be decided in polynomial time. If the size of $\mathcal{K}$ is exponential, then this is trivial. We cannot answer the question, e.g. when $\mathcal{K}$ consist polynomially many sets with unbounded sizes.

3. It is very easy to characterize the families which can be the family of minimal keys of a database. Can it be done for $\mathcal{C}_d$?

4. The investigations of the paper should be extended for the dependency structure of the databases.

5. The questions analogous to the results of the present paper can be asked for any other database model, replacing the relational one.

6. The following problem sounds similar to the problem treated here, but it is actually very different. Suppose that the data go through a noisy channel, where each data can be distorted with a small probability. Try to add new attributes to make the effective keys for the transmitted database small.

References

[1] Armstrong, W.W., Dependency structures of data base relationship, in: *Information Processing 74*, North-Holland, Amsterdam, pp. 580-583.

[2] Demetrovics, J., On the equivalence of candidate keys with Sperner systems, *Acta Cybernet.* **4**(1979) 247-252.