MTA Számítástechnikai és Automatizálási Kutató Intézete, Közlemények 27/1982

LOGICAL AND STRUCTURAL INVESTIGATIONS OF THE RELATIONAL DATA MODEL

J. DEMETROVICS

The concepts data base and data base management system have a central role in the computer aided information service and retrieval. Large masses of data representing complicated relationships are impossible to view on the machine level. The user may have but a generalized view of the wars of data stored by the computer and of its structure. The database management system is the software tool that establishes the link between the machine and the user's "general" level and thus permits to generate and conduct the storage, updating and retrieval of data on a logically higher level.

There are several ways to classify database management systems, the most generally accepted going by the way data and links between them are represented for the user. From the several data models proposed up to now three have had a relatively general acceptance and practical use: the hierarchical, the net and the relational models. With respect to its possibilities in future use the relational one (introduced by E.F.Codd) looks like one of the most promising data management tools.

In the relational approach data links are represented in the n-tuples of data. The relational model's advantage is not making prefound use of the machine representation ways, but representing data in a user conceivable form. It is a means apt to describe the logical structure of data bases as well. Data in it are stored in matrix form. A matrix stands for a unit of a Codd's third normal form relational data base, with its rows being the data records and its columns the data attributes. So this unit of the relational data base can be stored as a flat file.

The exact definition is the following:
Let Ω be a nonvoid finite set ($\Omega=\{a_1,a_2,\ldots,a_n\}$). A finite set of unary functions over this set is called a relation. These are depicted as two

dimensional arrays: if R is a relation over Ω and $R=\{h_1,h_2,\ldots,h_k\}$ where every h_i has arity n, the table of this relation is

	a_1	a_2	...	a_n
h_1	$h_1(a_1)$	$h_1(a_2)$	...	$h_1(a_n)$
h_2	$h_2(a_1)$	$h_2(a_2)$	...	$h_2(a_n)$
..	...	...	...	...
..	...	...	...	...
h_k	$h_k(a_1)$	$h_k(a_2)$	...	$h_k(a_n)$

Ω in this table is the set of attributes: the elements of R (i.e. the rows of the table) are essentially the records of data. These can have no repetitions in the relation, as the latter is defined as a set.

The relational data model has two main theoretical aspects. One of them is finding ways to discover and maintain links between data which are adequate to the structure of the relational data model. The two principal methods to describe links of data in an abstract way that these investigations have yielded use the concept of functional dependences and that of intersection dependences resp. The other theoretical aspect concerns the investigation of query methods obtainable to data bases constructed according to the relational data model. It is well known, that (commercially available) relational data bases are incomparably slower in queries than hierarchical or net data bases, especially if the forms of future queries can be obtained previously. Still, query forms cannot always be preassigned to data bases which gives to the relational data base good future chances if an effective and to the user demands well compatible query language is defined and the relational data base management system is properly organised which are the two basic tasks for the immediate future.

In the present paper the functional dependency and three analogous concepts are treated then the problems of the query operations [1], [5], [24] are considered.

For an effective data retrieval interconnections among data have to be dealt with properly. Functional dependencies (as introduced by E.F. Codd [9]) is an important tool for taking interconnections among data into consideration in relational data bases [3].

Definition 1: Let Ω be a set of attributes and R a relation over it.
$A, A \subseteq \Omega$ functionally depends on $B, B \subseteq \Omega$ iff

$$(\forall h,g \in R)((\forall a \in A)\ (h(a) = g(a)) \Rightarrow (\forall b \in B)\ (h(b) = g(b))),$$

This is denoted by $A \overset{f}{\underset{R}{\rightarrow}} B$ and heuristically means that determining the attribute values on A leaves no choice as to the attribute values on B.

Let R be a relation over Ω. We denote by F_R the set of its functional dependencies, i.e.

$$F_R = \{(A,B) : A \subseteq \Omega,\ B \subseteq \Omega,\ A \overset{f}{\underset{R}{\rightarrow}} B\}.$$

The set F_R is called the full family of functional dependencies in the relation R and much investigated because knowing only this of a relation permits to design its structuring in a relational data base in a concise and (in memory requirements) economical way.

Other types of data dependencies in a relation R are quite possible, of which we mention but three the dual, strong and weak dependencies, [10] of which the full families will be denoted by D_R, S_R and W_R

Definition 2: Let R be a relation over the attribute set let A and B be subsets of Ω.
B depends on A dually in R iff

$$(\forall h,g \in R)\ ((\exists a \in A)\ (h(a) = g(a)) \Rightarrow (\exists b \in B)\ (h(b) = g(b))),$$

B depends on A strongly in R iff

$$(\forall h,g \in R)\ ((\exists a \in A)\ (h(a) = g(a)) \Rightarrow (\forall b \in B)\ (h(b) = g(b))).$$

B depends on A weakly in R iff

$$(\forall h,g\in R)\ ((\forall a\in A)\ (h(a) = g(a)) \Rightarrow (\exists b\in B)\ (h(b) = g(b))).$$

By the logical structure of a relation we shall mean the full families F_R, D_R, S_R, W_R.

Knowing its dual and weak dependencies may greatly increase the efficiency of data retrieval from a relational data base when only partial information is required or when the user doesn't know all the attribute values necessary for the retrieval.

The knowledge of the strong dependencies is useful in the decomposition of big relations into smaller ones and thus helps in the reduction of the space requirements of the storage. The way of decomposition is analogous to how in the case of functional dependencies it is made [9], [12].

A basic problem in the theory of relational data bases is to characterize functional dependencies in a self contained way, i.e. to axiomatize them. A basic result of W.W. Armstrong [3] is to give an axiomatization of full f-families. This system of axioms is however, a little unpractical in handling important combinatorical problems of full f-families ([5], [6], [7]).

In [6] one can find a modification of Armstrong's axioms which faciliates handling certain types of combinatorical problems ([5], [19]). One of the purposes of the present paper is to give an axiomatization of full f-families based on their combinatorial properties. Next two theoretical problems of full families are discussed [15] and linear relations described [14]. A description of a factual system is given by the relational data model. A duality principle is stated between the functional and dual full families, the systems of axioms (with similar structures) are given for the full f-,d- and s-families and full w-families with no subset of attributes dependent on the void set (Theorem 1).

For the sake of completeness we give the axiom systems given for full f-, d- and s-families in papers [3] and [10]: let $Z \subseteq 2^{\Omega} \times 2^{\Omega}$ and $A,B,C,D \subseteq \Omega$.

The φ axioms are:

(F1) $(A,A)\in Z$;

(F2) if $(A,B)\in Z$ and $(B,C)\in Z$ then $(A,C)\in Z$;

(F3) if $(A,B)\in Z$ and $C \supseteq A$, $D \subseteq B$ then $(C,D)\in Z$;

(F4) if $(A,B)\in Z$ and $(C,D)\in Z$ then $(A\cup C, B\cup D)\in Z$.

The ν axioms are:

(D1) $(A,A)\in Z$;

(D2) if $(A,B)\in Z$ and $(B,C)\in Z$ then $(A,C)\in Z$;

(D3) if $(A,B)\in Z$ and $C \subseteq A$, $B \subseteq D$ then $(C,D)\in Z$;

(D4) if $(A,B)\in Z$ and $(C,D)\in Z$ then $(A\cup C, B\cup D)\in Z$;

(D5) if $(A,\emptyset)\in Z$ then $A\neq\emptyset$.

The γ axioms are:

(S1) $(\forall a\in\Omega)\ (\{a\}, \{a\})\in Z$;

(S2) if $(A,B)\in Z$ and $(B,C)\in Z$ and $B\neq\emptyset$ then $(A,C)\in Z$;

(S3) if $(A,B)\in Z$ and $C \subseteq A$, $D \subseteq B$ then $(C,D)\in Z$;

(S4) if $(A,B)\in Z$ and $(C,D)\in Z$ then $(A\cap C, B\cup D)\in Z$;

(S5) if $(A,B)\in Z$ and $(C,D)\in Z$ then $(A\cup C, B\cap D)\in Z$;

Next we give new systems of axioms of a new pattern for the full s-, d-, and f-families, then their analogous for weakly dependent families containing no subset of the attributes dependent on the void set (see Theorem 3). Let $Z \subseteq 2^{\Omega} \times 2^{\Omega}$ then Z satisfies the corresponding systems of axioms iff the following conditions hold:

The F axioms are:

$\forall(X,Y)\in(P(\Omega) \times P(\Omega)\setminus Z)$ $\exists E \subseteq \Omega$ such that

(i) $X \subseteq E$ and $Y \nsubseteq E$;

(ii) if $(A,B)\in Z$ and $A \subseteq E$ then $B \subseteq E$ holds.

The D axioms are:

$\forall(X,Y)\in(P(\Omega) \times P(\Omega)\setminus Z)$ $\exists E \subseteq \Omega$ such that

(i) $X \cap E \neq \emptyset$ and $Y \cap E = \emptyset$;

(ii) if $(A,B)\in Z$ and $A \cap E \neq \emptyset$ then $B \cap E \neq \emptyset$ holds.

The S axioms are:

$\forall(X,Y)\in(P(\Omega) \times P(\Omega)\setminus Z)$ $\exists E \subseteq \Omega$ such that

(i) $X \cap E \neq \emptyset$ $Y \nsubseteq E$;

(ii) $(A,B)\in Z$ $A \cap E \neq \emptyset$, $B \subseteq E$

The W axioms are:

$\forall(X,Y) \subseteq (P(\Omega) \times P(\Omega)\setminus Z)$ $\exists E \subseteq \Omega$ such that

(i) $X \subseteq E$ and $Y \cap E = \emptyset$;

(ii) $(A,B)\in Z$ and $A \subseteq E$ then $B \cap E \neq \emptyset$ holds.

Theorem 1: The φ, ν and γ systems of axioms are equivalent with the F, D and S systems, respectively.

Nest we shall define equality sets of matrices and show them to be characterized by the property that the 3 equality sets determined by 3 rows are a Δ system. (Theorem 2). Then we give some reason why the F,D,S and W axiom systems have such similar forms.

Definition 3: Let g,h be two rows of a relation R over Ω. The equality set of the rows g and h is

$$E(h,g) = \{a\in\Omega : h(a) = g(a)\}.$$

The equality set of the relation R (i.e. the matrix R) is defined as

$$\varepsilon(h,g) = \{a\in\Omega : h(a) = g(a)\}.$$

Definition 4: A class of sets is said to be a Δ-system if for any $A\neq B$, $C\neq D$ of its sets $A\cap B=C\cap D$.

Theorem 2: Let f,g,h be rows of the relation R. Then the class of sets $\{E(f,g), E(g,h), E(f,h)\}$ is a Δ-system.

Let $\varepsilon = \{E_{i,j} : 1 \leq i < j \leq k\}$ a class of subsets of Ω for which $\{E_{i,j}, E_{i,l}, E_{j,l}\}$ is a Δ-system for any $1 \leq i < j < l \leq k$. Then a relation R over Ω can be constructed with $\varepsilon_R=\varepsilon$. Theorem 2 permits a new formulation of the F,D,S and W-axioms which are equivalent with the old ones except for the W case (Theorem 3).

Let $Z \subseteq 2^{\Omega} \times 2^{\Omega}$ and E be an arbitrary class of sets

$$\{E_{i,j} : 1 \leq i < j \leq k, E_{i,j} \subseteq \Omega\}.$$

The F' axiom is:

for Z there are such k and E that

(i) if $(X,Y)\in P(\Omega)\times P(\Omega)\setminus Z$ then there are such $i,j\ (1 \leq i < j \leq k)$ that $X \subseteq E_{i,j}$ and $Y \nsubseteq E_{i,j}$;

(ii) if $(A,B)\in Z$ and $A \subseteq E_{i,j}$ then $B \subseteq E_{i,j}$ with $1 \leq i < j \leq k$;

(iii) if for any i,j,ℓ $1 \leq i < j < \ell \leq k$ $\{E_{i,j},E_{i,l},E_{j,l}\}$ is Δ-system.

The D' axiom is:

for Z there are such k and E that

(i) if $(X,Y)\in P(\Omega) \times P(\Omega)\setminus Z$ then there are such i,j $(1 \leq i < j \leq k)$, if $X \cap E_{i,j} \neq \emptyset$ and $Y \cap E_{i,j} = \emptyset$;

(ii) if $(A,B)\in Z$ and $A \cap E_{i,j} \neq \emptyset$ then $B \cap E_{i,j} \neq \emptyset$ with $1\leq i<j\leq k$;

(iii) if for any i,j,ℓ $(1 \leq i < j < \ell \leq k)$ $\{E_{i,j},E_{i,l},E_{j,l}\}$ is Δ-system.

The S' axiom is:

for Z there are such k and E that

(i) if $(X,Y)\in P(\Omega) \times P(\Omega)\setminus Z$ then there are such i,j $(1 \leq i < j \leq k)$ that $X \cap E_{i,j} \neq \emptyset$ and $Y \not\subseteq E_{i,j}$;

(ii) if $(A,B)\in Z$ and $A \cap E_{i,j} \neq \emptyset$ then $B \subseteq E_{i,j}$ with $1 \leq i < j \leq k$;

(iii) for any i,j,ℓ $(1 \leq i < j < \ell \leq k)$ $\{E_{i,j},E_{i,l},E_{j,l}\}$ is Δ-system.

The W' axiom is:

for Z there are such k and E that

(i) if $(X,Y)\in P(\Omega) \times P(\Omega)\setminus Z$ then there are such i,j $(1 \leq i < j \leq k)$ that $X \subseteq E_{i,j}$ and $Y \cap E_{i,j} = \emptyset$;

(ii) if $(A,B)\in Z$ and $A \subseteq E_{i,j}$ then $B \cap E_{i,j} \neq \emptyset$ with $1 \leq i < j \leq k$;

(iii) for any i,j,ℓ $(1 \leq i < j < \ell \leq k)$ $\{E_{i,j},E_{i,l},E_{j,l}\}$ is Δ-system.

Theorem 3: The F', D', S' axioms are equivalents to the F, D and S axioms respectively. The W' axioms are definitely stronger then the W axioms.

The cause of the last statement is that W-axioms are meaningless for set pairs of the form $(\emptyset,B)$.

Theorem 4 states, that F',D',S' and W'-axioms characterize f-, d-, s- and w-families.

Theorem 4: Let $Z \in 2^{\Omega} \times 2^{\Omega}$ have one of the properties F,D,S,W and let Y' denote the corresponding set of axioms F',D',S' or W'. Then Z statisfies the Y'-axioms iff a relation R over Ω exists for which $Y_R = Z$ holds. (Y_R is the set of the dependencies of the kind in point in the relation.)

Next we mention two combinatorial problems.

The first is to give the minimal number of rows in a relation which represents any given full f-family (or antichain) of an n-element set as its set of functional dependencies (or minimal keys, respectively) ([15], [18]). These minimal numbers of rows are denoted by s(n) and S(n) for the two problems above. In [15] the bounds

$$\sqrt{2\binom{n}{[n/2]}} \leq s(n) \leq 2\binom{n}{[n/2]}$$

were proved; a more recent result of the author is

Theorem 5:

$$\frac{1}{n^2}\binom{n}{[n/2]} \leq s(n) \leq \binom{n}{[n/2]} + 1 \quad \text{and}$$

$$\frac{1}{n^2}\binom{n}{[n/2]} \leq S(n) \leq \frac{3}{2}\binom{n}{[n/2]} .$$

The upper bound can be found as a byproduct of characterising generator sets of full f-families by their maximal dependent attribute subsets' intersection irreducible sets.

Next we metion theorems about linear relations. These are relations with rational attribute values, the rows of which are a linearly closed subset of the vector field $Q^{|\Omega|}$.
Of course a linear relation is always characterized by a finite subset of its rows.

Let $R \subset Q^{|\Omega|}$ be a linear relation; then

Theorem 6: Every dependence $(A,B) \in F_R$ in R is linear, i.e. is given by a linear operator in $Q^{|\Omega|}$.

Theorem 7: All the minimal keys in R have the same cardinality k where k can take an arbitrary value between 1 and $|\Omega|$.

The problem of axiomatizing full f-families of linear relations is equivalent with the (internal) characterization of coordinatable matroids over

Q and is therefore open.

A practical example for illustrating the efficiency of the relational data base structure is the two major parts of the Maze and Industial Plants Producing Branch of the Nádudvar Red Star Co-op data base which deals with stock and demand registering with special tasks for most of the subplants.

The investigation of but the functional dependencies proved sufficient to reduce storage area requirements of the system by about 40%. (Dual and weak dependencies were not found in this system.)

This task put forth the problem of efficient queries in the system which were interpreted as special tables (see [2], [24]). In [1] an efficient simplification algorithm for a large class of queries is given which improves the response time of the system which can in some cases quite long due to the difficulties that lie in the execution of the relation join operation.

REFERENCES

[1] Aho, A.V., Saigv, Y., Ullman, J.D.: Efficient Optimization of a Class of Relational Expressions, ACM Trans. Database Systems 4. (1979) 4, 435-454.

[2] Aho, A.V., Sagiv, V., Ullman, J.D.: Equivalences among relational expressions, SIAM J. Comput., 8 (1979), 218-246.

[3] Armstrong, W.W.: Dependency structures of data base relationship, Information Processing 74, North-Holland Publ. Co. (1974), 580-585.

[4] Armstrong, W.W.: On the generation of dependency structures of relational data bases, Publication 272, Universite de Montreal (1977)

[5] Beeri, C., Berstein, P.A.: Computational Problems Related to the Design of Normal Form Relational Schemas, ACM Trans. on Database Systems, 4, (1979) 1, 30-59.

[6] Békéssy, A., Demetrovics, J.: Contribution to the theory of data base relations, Discrete Math. 27 (1979), 1-10.

[7] Békéssy, A., Demetrovics, J., Hannák, L., Frank, P., Katona, Gy.: On the number of maximal dependencies in a data base relation of fixed order, Discrete Math. 30 (1980) 83-88.

[8] Codd, E.F.: A relational model for large shared data banks, Comm. ACM 13 (1970) 377-387.

[9] Codd, E.F.: Further normalization of the data base relational model, Data Base Systems, R. Rustin, ed. Prentice-Hall, Englewood Cliffs, NJ, (1972) 33-64.

[10] Czédli, G.: Függőségek relációs adatbázis modellben, Alk. Mat. Lapok (1980)

[11] Demetrovics, J.: On the number of candidate keys, Informations Processing Letters 7 (1978) 6, 266-269.

[12] Demetrovics, J.: Relációs adatbázis modell, MTA SZTAKI Közlemények 20 (1978) 21-23.

[13] Demetrovics, J.: Homogén file kulcsairól, Alkalmazott Matematikai Lapok 3 (1977) 185-191.

[14] Demetrovics, J.: On the equivalence of candidate keys with Sperner systems, Acta Cybernetica 4 (1979) 3, 247-252.

[15] Demetrovics, J.: Candidate keys and antichains, SIAM J. Alg. Disc. Meth. 1 (1980) 1,92.

[16] Demetrovics, J.: O klucsah odnonarodnüh fajlov, Kibernetika (Kiev) (to appear).

[17] Demetrovics, J., Gyepesi, Gy.: On the Functional Dependency and Some Generalization of it, Acta Cybernetica (to appear)

[18] Demetrovics, J., Gyepesi, Gy.: A note on candidate keys and full families, SIAM J. Alg. Disc. Meth. (to appear)

[19] Fagin, R.: Functional dependencies in a relational database and propositional logic, IBM J. Res. and Develop. 21 (1977) 6, 534-544.

[20] Kleitman, D.: On a combinatorial problem of Erdős, Proc. AMS (1966) 139-141.

[21] Maier, D., Mendelzon, A.O., Sagiv, Y.: Testing Implications of Data Dependencies, ACM Trans. Database Systems 4 (1974) 4, 455-469.

[22] Rissanen, J.: Independent components of relations, ACM Trans. Database Syst. 2 (1977) 4, 317-325.

[23] Sperner, E.: Ein Satz über Untermengen einer endlichen Menge, Mathematische Zeitschrift 27 (1928) 544-548.

[24] Zloof, M.M.: Query-byexample: A data base language, IBM Syst.J. 16 (1977) 4 324-343.

Összefoglaló

A relációs adatmodel logikai és strukturális vizsgálata

Demetrovics János

Ebben a cikkben a relációs adatmodellben definiálható függésekkel foglalkozunk, pontosabban a funkcinális függéssel és három analogonjával: a duális, erős és gyenge függésekkel. Adott tipusú függés vizsgálatakor az első feladatot az ún. teljes családok axiomatizálásá – a cikk első részében ezt végezzük. Kétféle axiómasémát adunk; az első csak a funkcionális, duális és erős fuggések teljes családjainak axiomatizálására alkalmas, mig a második séma a gyenge függésekére is.

Vizsgálunk még két, funkcionális függőségek teljes családjainak generálására, illetve kulcsrendszerekre vonatkozó probémát.

Végül megemlitjük azokat a lineáris relációkra vonatkozó tételeket, melyekből kiderül, hogy milyen következményei vannak a linearitásnak a reláció funkcionális függéseire.

РЕЗЮМЕ

Логическое и структуральное исследование в реальной базе данных

Я. Деметрович

В настоящей работе мы изучаем обобщения функциональных зависимостей. Кроме этого занимаемся линейными функциональными зависимостями и изучаем некоторые комбинаторные вопросы, связанные с реляционными базами данных.