The Third International Conference "Problems of Cybernetics and Informatics"
September 6-8, 2010, Baku, Azerbaijan. Section #1 "Information and Communication Technologies"
www.pci2010.science.az/1/32.pdf

IDENTIFICATION OF DATA INTERRELATION PATTERN AND FORMATION OF DATABASE ORGANIZATION STRUCTURE

Geray Kengerlinsky

Cybernetics Institute of ANAS, Baku, Azerbaijan.
kengerli@mail.ru

A database (DB) is a core of any integrated information control system. It represents an electronic storage of data [1] and has homogeneous properties access to which is provided by a computer. A database is classified by many signs the principal of these being a structure of data storage organization on which depends selection of the majority of other characteristics and parameters. Though the number of the most employed structures is practically limited to network, hierarchical and relaxation ones, principles of data storage in them are different. For this reason a possible mistake in selecting an organization structure at an early stage of DB development shows up later as the appearance of duplicate information and a decrease in efficiency of storage resources use.

Meanwhile, the selection of DB structure is up to the present completely dictated by the nature of interrelation of content records. And the right selection of the structure depends on the extent to which it is possible to make the correct determination of this nature. Unfortunately, a sufficiently universal method excluding personal errors in the determination of interrelation pattern and , hence, in the selection of DB storage structure, was never worked out. It is in this that the aim of the given research consists.

As is known, in any DB information is initially structurized and stored as individual records or two-dimension tables, which, in their turn, are formed by a group of logically related data elements. In the paper [2] the nature of formation of functional statistic connections of starting data is considered in detail. Also proposed is a suitable measure – K.Shannon's mutual information [3] – for its determination. On this basis one can immediately engage in analysis of the pattern of DB record interrelations. Actually, assume that all records of $A=A_1,\ldots,A_n$, data stored in DB are enumerated. On a set $A_i=(a_1,\ldots,a_n)$, $i=\overline{1,n}$ of data elements are prescribed probabilities (including joint ones) of using both these data and all kinds of their combinations during realization of queries.

Then mutual information of any record combination is expressed as [2,3]

$$I(A_1;\ldots;A_n) = \sum_{(a)} p(a_i,\ldots,a_n) I(a_i,\ldots,a_n) \tag{1}$$

But mutual information of any record combination is expressed through the difference of mutual information bits for combinations of lesser number of the same records with the help of the recurrent formula

$$I(A_i;\ldots;A_k;A_l) = I(A_i;A_k) - I(A_i;\ldots;A_k / A_l) \tag{2}$$

where A_l .is any record. By analyzing relation among the values in the right-hand part (2) it is possible to establish not only the fact itself but also the pattern of average level of interconnection of arbitrary record combination.

In point of fact, equality and their $(A_i;\ldots; A_k) = I(A_i;\ldots;A_k/A_l)$ is an information analogue of a generalized independence condition of random events of probability theory and is in line with a case when all records and all their possible combinations are not interrelated. As for two other possible relationships each of them characterizes different pattern of interrelations of records and their combinations. Now let

$$I(A_i;\ldots;A_k) \geq I(A_i;\ldots;A_k/A_l). \tag{3}$$

be fulfilled.

Then the fact of interrelation of record in any combination is responsible for the presence of interrelation in combinations with lesser number of the same records. If

$$\left| I(A_i;\ldots;A_k) \right| \leq \left| I(A_i;\ldots;A_k/A_l) \right| \tag{4}$$

is valid and the number of the records being analyzed exceeds two then the fact of interrelation of any record combination is invariant to interrelations in all combinations of lesser number of the same records. For example, in spite of pair wise non-relation of three records of combinations comprising all the three records they may turn out to be non-interrelated.

So, all possible types of interrelation peculiar to any set of data are confined to two ones of principally different nature. The first is formally described by the relation (3) and by all means envisages the presence of mutual relation among individual records (data elements in DB), i.e. element–by-element interconnection. The second one following from (4) can be called group because in this case the presence of interconnection among combinations (groups) and not among individual records is determining. It remains only to find out with which organization structures of DB each of them is connected.

To do this one can make use of entropy formula $H(A_1,...,A_n)$, which displays the average value of indetermination of all records in DB in relation to queries that is numerically equal to inherent information $I(A_1,...,A_n)$, contained in all records of DB concerning query implementation.

This expression is written in canonical form and considers all possible connections among the records.

$$H(A1,\ldots,An) = I(A1,\ldots,An) = \sum_i (A_i) - \sum_{i,j} (A_i;A_j) + \sum_{i,j,k} (A_i;A_j;A_k) + \ldots + (-1)^{n-1} I(A_i;\ldots A_n) \qquad (5)$$

Here and in subsequent expansions of inherent information the summation extends in all possible combinations of ordered subscripts $i < j < k < \ldots < n$.

If by DB structure is implied a collection of records with information connection among them any expansion of information $I(A_1,...,A_n)$, .is, in fail a certain structure. It should only be borne in mind that only next expansions are of interest for the formation of DB organization structures. Firstly, the expansions obtained with consideration for element-by-element or group pattern of interrelations. Secondly, those expansions which are formed of constituents having the least absolute value-i.e. elementary ones.

Because only this ensures analysis of all possible variants of expansion (5) and hence all principally possible structures corresponding to them.

Set's assume that interrelation of records revealed in DB is of element-by-element pattern. Then with the use of the recurrent formula (2) with consideration for inequality (3) mutual information of any record combination (5) at every step of transformation is broken down into two of lesser value until $l \le n$ as a result of which the canonical formula (5) becomes

$$I(A) = I(A_1;\ldots;A_n) = \sum_i I(A_i/A^*) + \sum_{i,j} I(A_i;A_j/A^*) + \sum_{i.j.k} I(A_i;A_j;A_k/A^*) + \ldots + I(A_1;\ldots;A_n). \qquad (6)$$

where A^* is located under the sign of sum stands for addition of corresponding records or their combinations up to the full set of records. The number of components making up every sum is equal to the quantity C_n^v, where v is number of summation indices (the amount of records in a combination) $i=\overline{1,n}$ and $v = n$ corresponds to $I(A_1;\ldots;A_n)$ because $C_n^n = 1$.

So, the expansion (6) displays the inherent information $I(A)$ of records in DB through the sum of components having the least value each of them being conditional or un conditional information, and expresses one of all kinds of combinations differing from all the rest by one record at least. From this immediately follow two important structural features of DB. Firstly, every record must be connected to all combinations in which it is contained. Secondly, there is no need in information connections among inherent record combinations, i.e. in the course of query implementation appropriate combinations are formed by all necessary records in parallel and independently of one another.

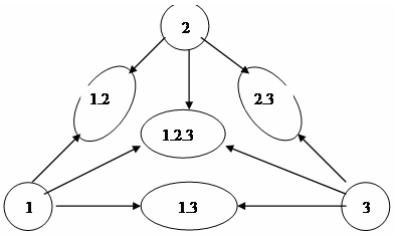
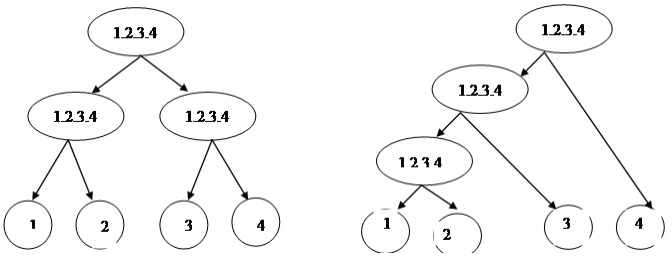


Fig. 1 Network organization of DB



а б

Fig. 2 Fragment of pyramidal (a) and stepwise structure (b) structure of DB

Thus, all revealed peculiarities of the expansions obtained with element-by-element pattern of data interrelation in fact, are in line with properties of simple network structures. The latter represent various records connected among them selves through relations (information relation) of the "one-to-one" and "one-to-many" type. As an explanation to the above said Fig.1 shows an organization structure synthesized for expansion

$$I(A_i,A_j,A_k) = \sum_i I\ (A_i/A_j,A_k)\ + \sum_{i,j} I\ \ (A_i;A_j/A_k)\ + I(A_1;A_2;A_3)$$

having element-by-element pattern of interrelation of three records.

Now let records interrelation in DB be of group pattern. It is not difficult to detect that as distinct from the previous element-by-element pattern, here the terms only increase at every step of expansion of the canonical formula (5) with the use of the recurrent relation (2). This means that in case of group pattern of record interrelation elementary components, starting from the second sum, will be only mutual unconditional bits of information, i.e

$I(A)= I(A_1,\ldots,A_n) =$

$$X_1 \sum_i I\ (A_i/A^* + X_2 \sum_{i,j} I\ (A_i;\ A_j) + X_3 \sum_{i.j.k} I\ (A_i;\ A_j;\ A_k) + \ldots + X_n\ I(A_1;\ldots;A_n). \qquad (7)$$

Here, as in the expansion (6) the number of components making up every sum still equals $\overset{v}{C}_n$ and values of coefficients X_v ($v = \overline{1,n}$) are determined from the formula $X_v = (-1)^n(v\text{-}1)$, with the assumption that $X_v = 1$ when $v = 1$, as the coefficient $X_1 = 1$.

By correlating the expansion (6) with (7) one can make sure that here any elementary component also contains all information about appropriate combinations of records and does not contain any information about all the rest. However, unlike the expansion (6), the expansion (7) contains components not only of different but also of the same kind, i.e. equivalent. This is attested by coefficients X_v at every term of the expansion beginning from the third any only the first two comprise components of different kind as $X_v = 1$ when $v = \overline{1,2}$.

Now one can turn to determination of information relations in structures synthesized on the basis of the expansion (7). When $v = \overline{1,2}$ .the coefficient $X_v = 1$ and all elementary components of the first two terms of the expansion (7) have different form. That is the same information relations as in structures with element-by-element pattern of interrelations are typical of them.

As for the remaining terms of (7) each of them contains (v -1)-th component of the same form due to group pattern of record interrelation. They all display interrelation of combinations of the same records, i.e. they are equivalent and unambiguously determine one another. For this reason in the given case every record, firstly, must be information ally connected not with one but immediately with (v -1)–th combinations in which it participates. Secondly, all record combinations must be interconnected. Yet, attempts at developing a DB structure only on this basis run into serious precisely every of (v -1)-th combinations must be information ally connected and in what way these combinations are interconnected.

Full solution of this problem is obtained in [2] where it is proved that any combination consisting of v records must be informationally connected- either with two other combinations of the same records or with one of (v -1)-th combinations of the same records and one of v records of the same combination; or with two records from the same combination. Moreover, every of v records can be connected with only one combination.

So, properties of all kinds of expansions obtained in case of group pattern of record interrelation on the whole conform to the principle of data ordering into hierarchical sequence in which retrieval is performed by step-by-step going down. Similar relations are usually structurized as a set of trees with information connection of the "one-to-one" or "one-to-many" type.

By way of example Fig.2 demonstrates two variants of forming information connections for a fragment of hierarchical structure of DB when v =4 and (v -1)=3. It is easy to make sure that both pyramidal (a) and stepwise (b) variants equally conform to the above stated principles of hierarchical structural organization of records in DB.

At present reserves of intensive widening storage capacity are virtually exhausted, all possibilities are believed to be reached through increasing affectivity of storage resources. A search for principally new structures and ways of organizing data storage in DB is sure to become a constituent of this program. And here it is necessary to be guided by the use of non-traditional approaches and methods allowing to reveal all aspects of organizing storage and control of interrelated data. The potentialities of one of these based on methods of K.Shannon's information theory were described in this report.

References

1. I.A. Ibragimov, G.A. Kengerlinsky. Conception of State Information Resources Storage of Azerbaijan/ The paper of the Second International Conference: "Problems of Cybernetics and Informatics", v. 1, Baku, 2008, pp. 23-26.
2. G.A. Kengerlinsky. Minimizing Information Duplikation in Relational Databases / Proceedings of the Third International Conference: "Problems of Cybernetics and Informatics", v_, Baku, 2010.
3. V.D. Kolesnik, G.Sh. Poltyrev. "A Course in Information Theory" (in Russian), Moscow, "Nauka" Publishing House, Moscow 1982, 416 p.
4. G.A. Kengerlinsky. Principles of Building Structures of Decentralized Information Processing Systems (in Russian)//Transactions of USSR Academy of Sciences, "Technical Cybernetics", № 5, Moscow, 1980, pp. 31-40.