MINIMIZING INFORMATION DUPLICATION IN RELATIONAL DATABASES

Geray Kengerlinsky

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

Relational databases (DB) are well tried out and generally accepted solutions at a level of "de facto" standard in electronic storages of information [1]. A set of data making up DB content is structurized in the form of two-dimension tables or relation, i.e. elementary information unit containing lines and columns of data.

Without going into details of relational DB functioning organization attention should be focused on its important feature: both network and hierarchical models can be reduced to relational one and will be supported by DB relational control system if special measures for preventing contradictions in information fund are envisaged while working out a specific information control system. In this sense a relational model can be thought to be a certain generalization. But the main thing is that such reduction of the models is accompanied by inevitable data duplication. But then a reasonable question arises: what are the mechanisms of duplication and is it possible at least to minimize information being duplicated.

Meanwhile, it is proved in the most general way in [2] that information duplication accompanies decomposition of any complicated system on partial or complete rupture of interrelations objectively characteristics of its elements. To put it in another way, duplication represents an unavoidable consequence of artificial breakdown of interrelated elements and their display as formally non-related group. And the price to be paid for it are additional resources employed for retaining predetermined functionality of a system through processing this information being duplicated.

Before passing to formal aspect of the problem attention must be paid to the fact that information in DB is initially structurized in the form of individual records or tables which, in their turn, are made up by a group of logically interconnected data elements.

It is just these data that are used in servicing queries the customers of which are both DB users and various applications of information control system. For example, they are used for computing some indices and preparing documents on their basis in information systems of corporative management; in analytical systems they are applied for constructing tables, diagrams and graphics relationships; in geoinformation systems they are used for making clients strata of cartographic materials on topographic base of digital cards etc.

However, for all functional diversity of queries they have the following in common. Firstly, it is always known apriori what precisely set of data will be needed in the performance of one or another query. Secondly, all data stored in DB are employed jointly, i.e. in the implementation of various queries one and the same data or their combinations can be used. It is because of this reason they turn out to be functionally connected among themselves. More than that, this interrelation is of probabilistic nature as the frequency of usage of some or other data as well as their combinations in the general case is determined by statistics of realizing appropriate queries. Then formalization of all the above said is reduced to the following in the terms of information theory.

Set all records of $A=A_1,\ldots,A_n$, stored in DB be 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 in the realization of queries.

The stated sufficiently general conditions allow to employ an amount of K.Shannon's mutual information as a measure of data interrelation. This measure characterizes the average level of statistical dependence among individual elements (or their arbitrary set) and all the rest of elements of the set $A_i, i=\overline{1,n}$ in case of intersecting and non-intersecting sets.

An expression for the average mutual information numerically determines the degree of data interrelation in content averages by all records and is represented as the following symmetrical value [3]:

$$I(A_1;...;A_n) = \sum_{(A)} p(a_i;...;a_n) I(a_i;...;a_n),$$

where: $I(a_i;...;a_n) = \log \dfrac{[\prod p(a_i,a_j)][\prod p(a_i,a_j,a_k,a_l)]...}{[\prod p(a_i)][\prod p(a_i,a_j,a_k)]...}$,

and *i, j ,k, l, . .* $\in \{N\} = \overline{1,n}$, while $i < j < k < l < \ldots < n$, and the products denoted by Π are computed from all possible combinations of ordered subscripts; expressions in the form $p(a_{/i}a_j)$, $p(a_i,a_j,a_k),\ldots$ - are joint probabilities of using corresponding combinations of data elements.

Then the inherent information's $I(A)$ of a set of DB records with network organization structure numerically equal to entropy $H(A)$ of content will be written [4] in the form of the following expansion

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

$$+ I(A_1;...;A_n), \qquad (1)$$

where: $i = \overline{1,n}, A^*$ is addition of suitable records or their combinations up to the full set of records.

The components of every sum of the given expression beginning from the second one, are elementary constituents which are the least by their value-mutual information bits characterizing the degree of record [4] interrelation in combinations of different number of records. Meanwhile, it is known that the value (1) varies within the limits $I(A_1,...,A_n) \le I(A) = H(A) \le \sum_i I(A_i)$.

If DB content is regarded to be a unit, i.e. with consideration for interrelation of all records, then the quantity $I(A) = H(A)$ takes on the lower boundary value equal to $I(A_1,...,A_n)$.- which is possible if under the sign of every sum in (1) appears a minimum number of all possible combinations of corresponding record number.

This condition is formally realized in two ways. The first of hem requires that the summation of elementary components in every sum in the expression (1) be performed with respect to all possible combinations of ordered subscripts $i < j < k < l < \ldots < k$. The second way is connected with the counting of elementary components of the v -th sum (v -quantity of subscripts) as the number of combinations C_n^v, a $v = n$, while υ =n matches up the combination $I(I_1;...;I_n)$ as $C_n^n = 1$.

In another extreme case when DB content is considered as a set of formally unrelated records $I(A) = H(A)$ takes on the upper boundary value equal to $\sum_i I(A_i)$ It is clear that for this it is necessary that every component in the expansion (1) comprise a maximum possible number of elementary constituents.

To achieve this it is required that under the sign of every sum in (1) appear not one but all possible combinations of corresponding records. It is clear that they all differ from one another only by the order of records in a combination and for this reason are equivalent. Accordingly, the summation in the expansion (1) must be carried out not only with respect to all ordered subscripts $i < j < k < l < \ldots < n$ but also with respect to all $i > j < k < l < \ldots < n$ In other words,

the number of elementary constituents of every component including equivalent ones, will be equal to $v\,C_n^v$.

The same mechanism is observed in the expansion of the inherent information *I(A)* when DB organization structure is hierarchical [4].

$$I(A) = (H) = X_1 \sum_i I(A_i) + X_2 \sum_{i,j} I(A_i; A_j) + X_3 \sum_{i,j,k} I(A_i; A_j; A_k) + \ldots + A_n I(A_1; \ldots; A_n)\,. \quad (2)$$

Really, if $I(A) = H(A) = I(A_1, \ldots, A_n)$, then here, as in the previous expansion the number of elementary constituents under the sign of every sum is still equal to C_n^v, where *v*, $v = \overline{1,n}$, is the number of subscripts of the corresponding sum. But unlike (1), the coefficient X_v at every term of the expansion (2) points to the appearance of equivalent components in it. The total number of elementary constituents in every term of the expansion (2) will become equal to $X_v = (-1^v)(v\text{-}1)$ and it is supposed that $X_v = 1$ when $v = 1$ as the coefficient $X_1 = 1$. Thus, in this case, the number of elementary constituents of every component will be equal to $v\text{-}1\ C_n^v$. If $I(A) = H(A) = \sum_i I(A_i)$ then the coefficient X_v takes on the value *v* and the number of elementary constituents of every term of the expansion (2) will be the same as in the expansion (1), i.e.

$v\ C_n^v$. which corresponds to the equality (2).

So, any content decomposition including the reduction of network and hierarchical DB models to relational one as well, is always accompanied by the breaking of record interrelation and the emergence of equivalent constituents in the corresponding terms of an expansion. It should be noted that it is just they that form duplicates when content is decomposed. An exception is presented only by equivalent constituents of the expansion (2) in case of $I(A) = H(A)$ which are not duplicates. It is explained by two reasons. Firstly, the equality (2) is the lower boundary value of the inherent information $I(A_1, \ldots, A_n)$, i.e. taking into consideration the interrelation of all content records. Secondly, it is explained by a group pattern displaying not the interrelation of individual records in combinations but the interrelation of combinations on the whole [4].

But if it is principally impossible to avoid duplicating data in the reduction of network and hierarchical structure to relational one attempts should be made at least to minimize it. This problem is solved only in one way: it is necessary to do one's best to retain the strongest interrelation and ignore weak ones when content is decomposed. In other words, it is required to form such kind of group that would contain strongly interrelated records, but the interrelation of the group themselves would be minimal. Then, still using the average amount of K.Shannon's mutual information as a measure of the record interrelation and keeping all introduced symbols and condition, one can formalize this problem in the following way.

The interrelated records $A = A_1, \ldots, A_n$, stored in relational DB, must be broken into some group $B_1, \ldots, B_m$ so that $m < n$, $U_{g=1}^m B_g = A_1, \ldots, A_n$ $_g = A_1, \ldots, A_n$ and $B_g \cap B_d = \varnothing$ if $g = d$ for all g, $d = \overline{1,m}$, and that every formed B_g –th group would contain the most strongly interrelated records. In this case, with consideration for all the above said the expression (1) for entropy or inherent information of the interrelated records stored in DB will be rewritten as

$$I(A) = \left[H(A) - \sum_{g=1}^m H(B_g / B_*)\right] + \sum_{g=1}^m \left[H(B_g / B_*) - \sum_{i \in G_g} H(A_i / A_j, \ldots, A_r)\right] + +$$
$$\sum_{i=1}^n H(A_i / A_{j,\ldots}, A_r),$$

where $H(B_g / B_*)$ is entropy of the q-th group given, that values of initial data of all the remaining group are known; G_g -a set of indices of initial data included in the g-th group.

Here the expressions in every square bracket stand for mutual information. But if in the first bracket it characterizes the degree of the formed . B_g –th group interrelation, in the second one it characterizes the interrelation degree of the records included in the g -th group. It is clear that a value of the third component in the expression being considered does not depend on the way of grouping the records and always remains constant. As for the two others, different values of these components correspond to a set of grouping variants and their total value being stable, i.e. maximization of intragroup record interrelation involves minimization of the corresponding intergroup interrelation and vice versa. It remains only to add that a value of the first square bracket equals a value of all duplicating information when content is decomposed.

Thus, solution of a similar optimization problem with the aim to minimize data duplication can serve as the basis for reorganizing traditional DB designing procedures and, first of all, such as the reduction of network and hierarchical models to relational ones and content decomposition when compiling two-dimension tables.

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. Information Approach to Decomposition of Complicated Systems (in Russian)// Transaction of USSR Academy of Sciences, "Technical Cybernetics", №1, Moscow, 1978, pp. 121-128.
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. Identification of Data Interrelation Pattern and Formation of Database Organization Structure/Proceedings of the Third International Conference: "Problems of Cybernetics and Informatics", v. 1, Baku, 2010.