Andrej Belyakov
Central Technological Research Institute, Moscow

Non-Parametrical Data Analysis in the Organization of Integrated Information Systems

Andrej Belyakov (b.1964) is a post-graduate student. His professional interests lie in the organization of information interaction of computer systems and the storage and transfer of data.

Belyakov, A.: **Non-parametrical data analysis in the organization of integrated information systems.**
Knowl.Org. 20(1993)No.4, p. 205
The transfer of data in information complexes encounters the problem of coordinated distribution of the information flows among specialized databases. It is possible to realize such a harmonization on the level of the users' interface with the help of methods of cluster and statistical analysis (Author)

The use of computers in information industry substantially expands the access of specialists to knowledge in the course of the solution of scientific-technical problems. Today thousands of automated systems, functioning on the basis of computers of different types, are being created.

The spontaneous appearance of computerized systems for the solving of specific, specialized problems stimulates the appearance of artificial barriers on the way of the information flows and the breaking-up of technological information chains. And if before information was processed manually and was, in its form, comprehensive to any specialist, now, with the use of computerized systems, the access to machine data is accomplished on the level of the user interface, which takes into account the specific features of only a particular problem. This leads to the slowdown, and sometimes even to the complete cessation of the movement of information in cases where for the solving of a problem, complex and highly diverse data are needed. The existing standards of data exchange do not completely solve the problem, since they cover only the most widely spread and recognized developments, the potentials of which, though great, are not boundless, and they also represent separate cases, which can be used only on certain conditions. This is particularly manifest in the attempt to make joint use of heterogeneous data bases that differ in their models and conceptual schemes.

A new approach to the solution of the problem of transfer and harmonization of data in the integration of databases of computerized systems can be gained through the application of non-parametrical methods of analysis of information held in the data bank, which do not attempt to structure the information according to some standard model. In this case all information, contained in the database is viewed as non-structured data which lack conceptual and semantic definiteness. Accordingly at the first state of such an analysis of a certain database there appears the problem of identification and classification in the data file of the information unit - the symbol. Moreover, any symbol belonging to the database is classified in accordance with the following features: 1) Relation to the functional field (systemic, service or information), 2) Relation to the type of data (symbol, digital).

At the initial state, the reference of this or that symbol to a definite class can be realized only with a certain degree of probability. This probabilistic classification makes it possible to establish a large number of stable combinations of symbols of information units of the database. These units are of indefinite character since the symbols of which they are made up have been classified with a certain degree of probability and have not yet gained a definite meaning. At this stage, using the mathematical aggregate of pragmatic statistical methods of information processing, such as cluster analysis and multi-dimensional measurement, the periodogram of the sequence of all the information units is constructed. As a result of this, it is only at the second state of the analysis that it will be possible to re-create a rough, and in some cases even a precise structure of the file of the analyzed information. In order to raise the reliability of identification, the next stage of the work, connected with the lending of definiteness to the information units through the checking of the correspondence of the structure of each unit to the standard structure, which has been created in the course of the statistical processing of information at the second stage, is realized. As a result of this final comparison, a model of the structure of the database, which is very close to the original one, is generated.