UDC 004.652 (045)
DOI: 10.18372/1990-5548.56.12943

[1]M. G. Glava,
[2]E. V. Malakhov

COMPARISON OF NUMERIC PROPERTIES OF OBJECTS OF DIFFERENT DATA DOMAINS IN RELATIONAL DATABASES

[1]Department of Information systems, Odessa National Polytechnic University, Odessa, Ukraine
[2]Department of Mathematical Support of Computer Systems, Odessa I. I. Mechnikov National University, Odessa, Ukraine
E-mails: [1]glavamg@gmail.com, [2]opmev@mail.ru

***Abstract*—*The problem of association of models of data domains is considered. It is offered to compare objects of data domains on the basis of values of properties of copies of those objects. Methods of benchmarking properties differ depending on type of scales in which their values are measured. It is offered to compare properties of numerical data type by k-means methods, histograms, and also to check coincidence of the distribution law of values of properties. In conjunction of signs to make the decision on similarity of the compared properties.*

Index Terms—Database; subject domain; model of subject domain; information system; entity instance; numeric type properties; clustering; *k*-means; method of histograms.

I. INTRODUCTION

Today work of any organization of any field of activity is not possible without use of information technologies. The huge data flow is stored and processed by means of data bases and data storages that significantly simplify management and control of activity.

Considering an economic situation of country and analyzing the market subject to reorganization of enterprises, it is possible to draw conclusion on existence of problem of integration of information systems. Using as a part of enterprise information systems of diverse "components" can cause difficulties as at solution of tasks of enterprise management or information exchange, and at management of those components, their support and administration. All this forces to solve issues of compatibility of different systems, and also problems of obtaining up-to-date data at accomplishment of advanced analytical queries [1].

Change of the external environment involves need of business process reengineering of an enterprise that, in turn, demands integration of functioning information systems (IS). As any information system operates with a large number of important information which needs to be saved in the course of reengineering [2].

According to the ISO/IEC standard 2382:2015 Information technology, information system is a system intended for information storage, search and processing and appropriate organizational resources (human, technical, financial, etc.) which provide and distribute information. That is, modern information systems are a set of information which is contained in the databases (DB) and information technologies providing its processing.

Today ICs which cornerstone DBs on the basis of a relational model are most widespread in a business environment. Relational DB are attractive to a wide class of tasks as they provide simplicity and reliability of work and have the most worked mathematics of a manipulation of DB objects.

II. PROBLEM STATEMENT

Owing to the above in enterprise information systems there is a problem of interaction of DBs realized in different DBMSs. There is urgent problem of association of the existing heterogeneous DBs integrated into a common information space. Integration of a DB in this case is not so much urgent at the level of external schemes, i.e. user's representation, but at the level of data, i.e. the task comes down to association of information models of subject domains (SD), integrated heterogeneous DB.

According to the classical model of SD for creation of SD general model of higher level which will include SD models of integrated heterogeneous DBs, it is necessary to combine objects and connections between them. And for solution of the most important problems of integration of DBs, such as ensuring integrity of data and avoidance of their duplication, it is necessary to reveal similar objects of SDs and their property. In articles [3]–[4] the technology of search of projections of the same SDs (objects of SDs) in which it is offered to compare objects on the basis of values of properties of copies of these objects is offered. Algorithms of comparison differ depending on data type of specific property. In article [5] the benchmark method of properties of rated type is offered. In this work the benchmark method of properties of numerical type is offered.

According to the offered technology objects of compared SDs need to be prepared for analysis:

– to select the significant properties [6] characterizing a certain object of SD, having ranged them by the integral assessment based on statistical data of work of the functioning DBs;

– to range on the importance objects of each compared SD, based on the following criteria: the number of connections of a certain object with others in the same SD in which it takes part as parent and/or dependent, and amount of the significant properties measured by a certain scale (serial, rated, numerical);

– to sort copies of objects by values of rated and serial properties, observing the rank of properties received earlier;

– to select subsets of copies for analysis which will have equal power in objects of the corresponding ranks in each SD, at the same time not to admit for examination of NULL value of copies of objects of SD.

III. PROBLEM SOLUTION

Comparison of properties of numerical type is carried out parallel to the analysis of rated and serial properties that will allow reducing time expenditure by comparison of objects of SD.

It is possible to assume that the previous steps pulled together a rank of potentially similar objects and their properties. The rank is understood as the place of object/property in the sequence of the considered objects/properties defined by means of a serial scale.

A. k-means method

For comparison of properties of numerical type it is offered to break values of properties into groups on similarity that will simplify further processing and decision-making.

Selection from an initial great number of these groups of values with similar properties is called a clustering [7]. Methods of clustering are divided by a data handling method on hierarchical and not hierarchical as to the way of data processing. At a hierarchical clustering consecutive association of smaller clusters into big ones or separation of big clusters onto smaller is carried out [8]. But at large volume of data hierarchical methods are not effective. Not hierarchical methods work iterative, breaking an initial set, create clusters till the rule of a stop will be reached.

The *k*-means [9] method belongs to the simplest and effective not hierarchical methods of clustering. Its purpose is minimization of an Euclidean metrics between values of properties of one cluster. *k*-means consists of the following steps.

The number of clusters of *k* which has to be created from values of properties of initial selection is to be set.

K of records which will serve as the initial centers of clusters is in a random way selected.

For each value of initial selection the closest to it center of cluster is defined.

Calculation of the centers of gravity of clusters (centroids) is made. It is made by determination of an average for value each sign of all records in a cluster. Then the old centers of clusters are displaced in its centrodes. Thus, the centrodes become the new centers of clusters for the following iteration of an algorithm.

Stop of an algorithm is made when borders of clusters and arrangement of centroids cease to change, that is on each iteration in each cluster there is the same record set. The algorithm of *k*-means usually finds a set of stable clusters for several tens of iterations.

One of shortcomings of *k*-means is lack of clear criterion for choice of optimum number of clusters.

For comparison of potentially similar numerical properties the method of *k*-averages offers to break values of both properties separately a_i^n and a_i^m SD d_n and d_m, respectively, for equal quantity of clusters of q. The quantity of clusters has to be less or to equally minimum quantity of unique values of properties:

$$q <= \min\{|M_u^n|, |M_u^m|\}, \qquad (1)$$

where M_u^n and M_u^m are multitude of unique values of SDs of d_n and d_m properties, respectively.

B. Method of histograms

For verification of the made decision and deviation of accidental results it is offered to carry out comparison of properties by method of histograms and to compare the received results.

Histogram is a tool allowing to visually evaluate distribution of the statistical data grouped in the frequency of their hit in certain (preset) interval [10]. This method is one of graphic methods of data representation which allows perceiving well and easily received results. It effectively copes both with large volumes of selection, and with the characteristic of a small numerical row.

Histogram represents the column diagram constructed according to the analyzed data which are divided into a number of intervals located in ascending order on abscissa axis, and on ordinate axis there is frequency of hits of values of properties in a certain interval. The choice of the size of an interval is also important, as well as the choice of number of clusters in the analysis by *k*-means method. As a big interval can hide important information or push to false decision-making. A small interval can both reveal hidden characteristics, and, for solution of an objective, be not capable to generalize data to level suitable for comparison, i.e. groups of properties will turn out so small that similarity will not be revealed false (almost step-by-step comparison of values).

For solution of a set objective of comparison of numerical objects properties of different SDs on the basis of their values by method of histograms also assumes splitting values of properties a_i^n and a_i^m of SD d_n and d_m, respectively, for *q* of groups which has to match quantity of cluster in the analysis of properties by *k*-means method.

For reduction of groups to comparable measurements it is offered to calculate interval length as follows.

Step 1. To define the maximum value of multitude of values of m of both compared properties a_i^n and a_i^m

$$\max\{m^n, m^m\}. \qquad (2)$$

Step 2. To define the minimum value of multitude of values m of both compared properties a_i^n and a_i^m

$$\min\{m^n, m^m\}. \qquad (3)$$

Step 3. To deduct minimum from the maximum value and to divide the result into the number of groups *q*:

$$\left(\max\{m^n, m^m\} - \min\{m^n, m^m\}\right) / q, \qquad (4)$$

where m_n and m_m are values of SDs of d_n and d_m properties, respectively.

For analysis of each received group of different properties of an identical rank and decision-making it is offered to calculate and compare to the threshold set by experts in that SD, weighed deviation of each group. Weighed deviation is offered to be calculated as follows.

Step 1. To calculate amount of values in each group, i.e. the power of multitudes values in *j* group of $m_{q_j}^n$ and $m_{q_j}^m$.

Step 2. To calculate a deviation in each group as a difference of amount of values of properties a_i^n and a_i^m, received in the previous step.

Step 3. To calculate the weighed deviation dividing module of the result received in step 2 on quantity of values of a subset of copies which are selected for comparison of properties:

$$\left\| m_{q_j}^n \right| - \left| m_{q_j}^m \right| / m_{a_i} \Big|, \qquad (5)$$

where m_{a_i} are values of *i* property.

C. Coincidence of the distribution law

The last step before acceptance of the decision on similarity of properties of different SDs offers to check coincidence of the distribution law of a set of values of properties a_i^n and a_i^m.

It is offered to make a statistical hypothesis that the considered values of properties are taken from the same collection [11]–[12]. The statistical hypothesis is some assumption of properties of population which needs to be checked. Outputs received by check of statistical hypotheses have probabilistic character: they are accepted with some probability.

For check of a statistical hypothesis it is necessary to follow these steps [13].

Step 1. To formulate the main H_0 and alternative H_1 hypotheses.

H_0 is the assumption of properties of population which is logical and plausible, but demands check.

H_1 is the statement about properties of population which is accepted in case there is no opportunity to accept a main hypothesis.

Step 2. To select a statistical criterion by means of which hypothesis will be checked.

Statistical criterion is a statistical characteristic of selection calculated on some formula on the basis of the data which are available in selection. The statistical criterion is a random variable which distribution law is known. Than value of statistical criterion is closer to zero, especially it is probable that the main hypothesis is correct.

Step 3. To set value of significance value of α.

The significance value α is a probability of an error of first kind (rejection of a main hypothesis while it is right). Value of significance value is usually rather small and set by an analyst checking a hypothesis. Most often accepting values are 0.01, 0.05 and 0.1.

Step 4. To find borders of field of accepting a hypothesis.

The field of acceptance of a hypothesis is a subset of such values of criterion at which the main hypothesis cannot be rejected. The field of acceptance of a hypothesis always includes value 0.

Critical area is a subset of such values of criterion at which the main hypothesis cannot be accepted.

Step 5. To draw conclusion on acceptance or rejection of the main hypothesis of H_0.

If the value of criterion found on selective values of observations belongs to the field of acceptance of a hypothesis, conclusion that there is no opportunity to reject a main hypothesis is drawn.

If the criterion belongs to critical area, conclusion that there is no opportunity to accept a main hypothesis is drawn. In that case an alternate hypothesis is accepted.

There are two types of hypotheses of uniformity of selections. Uniformity of selections "in weak" can be checked: if their parameters, first of all, an average not significantly differs. Uniformity of selections "in strong" can be checked: if their distribution laws not significantly differ. By means of Student's criterion the hypothesis of uniformity of selections "in weak" is checked. By means of Kolmogorov–Smirnov's criterion the hypothesis of uniformity of selections "in strong" is checked, that is that distribution functions of selections not significantly differ from each other.

This criterion allows to find a point in which the amount of the saved-up frequencies of discrepancies between two distributions is the greatest and to evaluate reliability of this discrepancy.

Let there are two independent selections made from populations with unknown theoretical functions of distributing $F_1(x)$ and $F_2(x)$.

The checked null hypothesis has appearance of $H_0 : F_1(x) = F_2(x)$ against competing $H_1 : F_1(x) \neq F_2(x)$. Let's assume that functions $F_1(x)$ and $F_2(x)$ are continuous and for assessment we use Kolmogorov–Smirnov's statistics.

It is necessary to calculate relative frequencies for two sets of values of properties, division of frequencies into selection amount. Further to define the module of a difference of the corresponding relative frequencies for control and experimental selections. Among the received modules of differences of relative frequencies to select the greatest module. Experimental value of criterion of Kolmogorov–Smirnov has appearance of:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max \left| F_{n_1}(x) - F_{n_2}(x) \right|, \qquad (6)$$

where $F_{n_1}(x)$ and $F_{n_2}(x)$ are the empirical distribution functions constructed on two selections with volumes of n_1 and n_2.

To draw conclusion on similarity by the considered criterion between two sets of values of properties, it is necessary to compare experimental value of criterion to its critical value determined by the special table proceeding from significance value of α. The null hypothesis should be accepted if observed value of criterion does not surpass its critical value.

At the same time, the power of the considered sets has to be rather big: $n_1 \geq 50$ and $n_2 \geq 50$.

Decision on similarity of properties is made on set of the received signs, i.e. on proximity of the centers of clusters, dispersion and standard deviation in q clusters by method of k-averages, deviations in q groups on method of histograms and coincidence of the distribution law of values that will allow to reduce probability of emergence of errors. If at least in one of couples of signs there is distinction, properties are accepted as different. Others are moved for analysis to experts.

IV. Approbation of the Offered Method

Let's check operability of the offered steps on an example.

Original values for the analysis are presented in Table I.

Table I. Original Values

a_1^n	a_1^m
…	…
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
200.00	125.67
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
500.00	200.00
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
767.77	534.12
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
1057.75	767.77
…	…

Let's select quantity of clusters for separation of values of properties. The quantity of clusters is defined experimentally. In this example we will accept $q = 4$.

Let's analyze properties of a_i^n and a_i^m, having divided values by k-means method and having calculated the centers of clusters, standard deviation and dispersion for each cluster of both compared properties. Analysis results of values are presented by method of k-averages in Tables II and III.

TABLE II. RESULTS OF CLUSTERING a_i^n BY *K*-MEANS METHOD

Cluster No.	Center of Cluster	Standard Deviation	Dispersion
1	0.028	0.000	0.000
2	– 0.470	0.129	0.017
3	1.172	0.499	0.249
4	– 1.103	0.152	0.023

TABLE III. RESULTS OF CLUSTERING a_i^m BY *K*-MEANS METHOD

Cluster No.	Center of Cluster	Standard Deviation	Dispersion
1	0.166	0.000	0.000
2	– 0.502	0.204	0.042
3	1.188	0.341	0.116
4	– 1.136	0.079	0.006

The received values are similar, it is possible to speak about similarity of these properties. Let's check this assumption having compared values of the properties by method of histograms.

Analysis results of values are presented by method of histograms in Table IV.

TABLE IV. RESULTS OF COMPARISON OF PROPERTIES BY METHOD OF HISTOGRAMS

	1	2	3	4
Amount of values a_i^n	36.00	27.00	18.00	18.00
Amount of values a_1^m	45.00	18.00	18.00	18.00
The weighed deviation	0.091	0.091	0.000	0.000

The weighed deviation is equal in two clusters to zero, in two it makes only 9%, therefore, it is possible to draw conclusion on similarity of properties.

The last criterion for comparison of properties of numerical type of different SDs is check of coincidence of the distribution law of values by means of Kolmogorov–Smirnov's criterion. Results of calculation are presented in Table V.

TABLE V

CALCULATION OF CRITERION OF KOLMOGOROV-SMIRNOV

	n_1	n_2	$F_{n_1}(x)$	$F_{n_2}(x)$	$\lvert F_{n_1}(x)-F_{n_2}(x)\rvert$
871.798	36	45	0.364	0.455	0.091
1743.595	63	63	0.636	0.636	0.000
2615.393	81	81	0.818	0.818	0.000
3487.190	99	99	1.000	1.000	0.000

According to the item (6) and calculations received in Table V $\lambda'=0.64$, according to statistical tables at significance value of 5% critical value of criterion $\lambda_{0.05}=1.36$.

As empirical value of criterion is less than critical, $\lambda'<\lambda_{0.05}$, the null hypothesis H_0 is accepted, that is properties are described by the same distribution function.

Output and decision-making: set of signs of comparison of properties of numerical type confirm assumption of similarity of properties.

V. CONCLUSION

Proceeding from results of testing of the offered methods for comparison of properties of numerical type of objects of different data domains it is offered to use set of methods: k-medium, histograms and coincidence of the distribution law of values of properties in a complex with the approaches offered in the technology described in [3].

Application of the offered approach automates a set objective that considerably will reduce an operating time for a person making a decision and will reduce error probability of the 1st sort unlike the automated solution of a task.

REFERENCES

[1] M. G. Glava and T. P. Vasylieva, ”Major problems and methods of databases integration”, *First Independent Scientific Journal*, no. 1, pp. 28–32, 2015. (in Russian).

[2] M. Glava and V. Malakhov, ”Information Systems Reengineering Approach Based on the Model of Information Systems Domains”, *International Journal of Software Engineering and Computer Systems (IJSECS)*, vol. 4, no. 1, pp. 95–105, 2018, doi: 10.15282/ijsecs.4.1.2018.8.0041.

[3] M. Glava and E. Malakhov, “Searching Similar Entities in Models of Various Subject Domains Based on the Analysis of Their Tuples”, *2016 International Conference on Electronics and Information Technology (EIT'16)*, May 23–27, 2016, Odesa, Ukraine, pp. 97–100, 2016, doi: 10.1109/ICEAIT.2016.7501001, EID: 2-s2.0-84979503116.

[4] T. Filatova and M. Glava, “Mathematical Models of Information Manipulation in the Subject Field of Intellectual Production in Educational Institutions”, *2016 International Conference on Electronics and Information Technology (EIT'16)*, May 23–27, 2016, Odesa, Ukraine, pp. 92–96, 2016, doi: 10.1109/ICEAIT.2016.7501000; EID: 2-s2.0-84979554925.

[5] M. Glava, ”Comparison of the nominal type properties of objects of different subject subdomains in relational databases”, *Informatics and Mathematical Methods in Simulation,* vol. 6, no. 3, pp. 302-309, 2016. (in Russian).

[6] E. V. Malakhov, G. N. Vostrov and M. G. Mikulinska, ”Methods of subject domains objects properties importance definition”, *Refrigeration engineering and Technology*, no. 4 (126), pp. 73–77, Odessa, 2010. (in Russian).

[7] B. G. Mirkin, *Methods of cluster analysis for decision support: review,* Moskow, Publishing house of National Research University "Higher School of Economics", 2011. (in Russian).

[8] K. S. Yershov and T. N. Romanova, ”The analysis and classification of algorithms of clustering”, *New information technologies in automated systems*, pp. 274–279, 2016. (in Russian).

[9] I. A. Bessmertnyj, A. B. Nugumanova and A. V. Platonov, *Intellectual systems. Manual and practical work for SPO*, Moskow, Publishing House of Eurite, 2018. (in Russian).

[10] S. I. Solonin, *Method of histograms: Manual,* M.-Berlin: Direct-Media, 2015. (in Russian).

[11] A. I. Orlov, ”Well-founded criteria of check of absolute uniformity of independent selections”, *Factory Laboratory. Diagnostics of Materials*, vol. 78, no. 11, pp. 66–70, 2012. (in Russian).

[12] G. V. Rubleva, *Mathematical statistics: statistical criterions of check of hypotheses. The methodology for students of full-time courses of technical and engineering specialties*, Tyumen, Publishing House of the Tyumen State University, 2014. (in Russian).

[13] Yu. Subkov, ”*"Net" and applied mathematics*”, access mode: https://function-x.ru. (in Russian).

Received February 19, 2018.

Glava Maria. Assistant Professor.
Department of Information systems, Odessa National Polytechnic University, Odessa, Ukraine.
Education: Odessa National Polytechnic University, Odessa, Ukraine (2009).
Research area: relational databases, subject domains modeling.
Publications: 29.
E-mail: glavamg@gmail.com

Malakhov Eugene. Doctor of Engineering Science. Professor.
Department of Mathematical Support of Computer Systems, Odessa I. I. Mechnikov National University, Odessa, Ukraine.
Education: Odessa Polytechnic Institute, Odessa, Ukraine (1989).
Direction of scientific activity: information technology, database theory, subject domains modeling.
Publications: 137.
E-mail: opmev@mail.ru

М. Г. Глава, Є. В. Малахов. Порівняння числових властивостей об'єктів різних предметних областей в реляційних базах даних

Розглянуто проблему об’єднання моделей ПрО. Запропоновано зіставляти об'єкти ПрО на основі значень властивостей екземплярів цих об'єктів. Методи зіставлення властивостей розрізняються в залежності від типу шкал, в яких вимірюються їх значення. Пропонується порівнювати властивості числового типу даних методами k-means, гістограм, а також перевірити збіг закону розподілу значень властивостей. За сукупністю ознак прийняти рішення про подібність порівнюваних властивостей.

Ключові слова: база даних; предметна область; модель предметної області; інформаційна система; екземпляр об'єкта; властивості числового типу; кластеризація; метод k-середніх; метод гістограм.

Глава Марія Геннадіївна. Старший викладач.
Кафедра інформаційних систем, Одеський національний політехнічний університет, Одеса, Україна.
Освіта: Одеський національний політехнічний університет, Одеса, Україна (2009).
Напрямок наукової діяльності: реляційні бази даних, моделювання предметних областей.
Публікації: 29.
E-mail: glavamg@gmail.com

Малахов Євгеній Валерійович. Доктор технічних наук. Професор.
Кафедра математичного забезпечення комп'ютерних систем, Одеський національний університет імені І. І. Мечникова, Одеса, Україна.
Освіта: Одеський політехнічний інститут, Одеса, Україна (1989).
Напрямок наукової діяльності: інформаційні технології, теорія баз даних, моделювання предметних областей.
Публікації: 137.
E-mail: opmev@mail.ru

М. Г. Глава, Е. В. Малахов. Сравнение числовых свойств объектов различных предметных областей в реляционных базах данных

Рассмотрена проблема объединения моделей ПрО. Предложено сопоставлять объекты ПрО на основе значений свойств экземпляров этих объектов. Методы сопоставления свойств различаются в зависимости от типа шкал, в которых измеряются их значения. Предлагается сравнивать свойства числового типа данных методами k-means, гистограмм, а также проверить совпадение закона распределения значений свойств. По совокупности признаков принять решение о подобии сравниваемых свойств.

Ключевые слова: база данных; предметная область; модель предметной области; информационная система; экземпляр объекта; свойства числового типа; кластеризация; метод k-средних; метод гистограмм.

Глава Мария Геннадьевна. Старший преподаватель

Кафедра информационных систем, Одесский национальный политехнический университет, Одесса, Украина.

Образование: Одесский национальный политехнический университет, Одесса, Украина (2009).

Направление научной деятельности: реляционные базы данных, моделирование предметных областей.

Публикации: 29.

E-mail: glavamg@gmail.com

Малахов Евгений Валерьевич. Доктор технических наук. Профессор

Кафедра математического обеспечения компьютерных систем, Одесский национальный университет имени И. И. Мечникова, Одесса, Украина.

Образование: Одесский политехнический институт, Одесса, Украина (1989).

Направление научной деятельности: информационные технологии, теория баз данных, моделирование предметных областей.

Публикации: 137.

E-mail: opmev@mail.ru