Comparison of the nominal type properties of objects of different subject domains

Maria G. Glava[1] and Eugene V. Malakhov[2]

[1] Odessa National Polytechnic University, 1, Shevchenko Ave., 65044 Odessa, Ukraine (E-mail: glavamaria@mail.ru)
[2] I.I. Mechnikov Odessa National University, str. Dvoryanskaya 2, 65026 Odessa, Ukraine (E-mail: opmev@mail.ru)

This article discusses the problem of the SD models merge. It is proposed to compare SD objects on the basis of the properties values of the tuples of these objects. The methods for comparing the properties of objects differ depending on the type of scales in which their values are measured. The nominal type properties we suggest to compare by constructing contingency tables and Pearson's chi-squared test.

These days, the work of any organization in any industry is not possible without the use of information technologies. Databases and data storages can store and process the huge data stream, which simplifies the management and control activities to a considerable extent.

Taking into account the economic situation of the country and analyzing the market, which is subject to reorganization of the enterprises, we can conclude that there are problems of the integration of information systems (IS).

The any IS design starts with a description of the subject domain (SD) and with a construction of its model. The solving this problem is the merging of SD models of the analyzed IS. To excluding redundancy and inconsistency of data, it needs to determine the similar objects in different SD.

In [1] the search technology of the same SD projection (SD objects) is proposed, which is proposed to compare objects on the basis of the properties values of the tuples of these objects. The comparison algorithms differ depending on the data type of the specific property. This paper we propose the comparison algorithm of the nominal type properties.

Under the proposed technology, it needs to prepare the objects of the potentially similar subject domain for the comparison, having allocated significant properties, having ranked, having grouped by data types and having sorted the tuples [1]. It carries out manipulation of the properties of the nominal type after analysis of serial properties.

We can assume that the preceding steps brought together the rank of potentially similar objects and their properties. Implementation the comparison algorithm of the serial type properties brought together the tuples compared objects. Accordingly, values of the nominal properties compare for aligned capacities of the tuple sets.

The comparison of the values symbol-by-symbol rejected because the information on the same object or action we can represent by different nominal values.

To compare the nominal type properties we suggest creating a base of signs F, which characterize any nominal properties. These characteristics include, for example, the number of spaces in property values; the number of capital letters; the part of speech; the presence of punctuation; the presence of abbreviations, etc.

The next step is to fill in the sets of signs F, having processed the nominal values of each property of the compared SD objects.

To determine the measure of the properties similarity, we form the contingency table (or the cross-table) on the following signs:

The number of spaces = 0; 0 < The number of spaces < 3; The number of spaces $\geq$ 3; The number of capital letters = 0; 0 < The number of capital letters < 4; The number of capital letters $\geq$ 4; The presence of abbreviations = "yes"; The presence of abbreviations = "no"; The presence

of punctuation = ”yes”; The presence of punctuation = ”no”; The part of speech = ”noun”; The part of speech = ”adjective”; The part of speech = ”verb”.

During next steps, we suggest to compare the nominal properties pairwise in the order that we defined in step a ranking of objects and their properties. When detecting a low level of compliance, it exclude these properties from consideration. When selection of properties with high and average similarity measure, it analyze them with help of the experts, because the software method does not guarantee that the error will be excluded in the analysis of nominal properties. Nevertheless, this method will reduce and simplify the properties processing for the experts.

The similarity measure between the properties of the nominal type on the basis of the signs complex was analyzed by the following methods: taxonomic analysis of E. S. Smirnov [2], coefficient of Pearson’s mutual contingency [3], Pearson’s chi-squared test [4].

The most revealing on the test data defined Pearson’s chi-squared test.

References

1. Maria Glava, Eugene 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 (EIT16), May 2327, 2016, Odesa, Ukraine, 2016, pp. 97100, ISBN 978-1-5090-2224-3.
2. Smirnov E. S. Taxonomic analysis. M.: Publisher Moscow University, 1969 (in Russian).
3. Gromyko G. L.: Textbook. M.: INFRA-M. 2005. P. 476 (in Russian).
4. Lapach S. N., Chubenko A. B., Babich P. N. Statistics in science and business. Kiev: Morion, 2002 (in Russian).