UDK 004.8

Olena A. Arsirii[1]**,** Doctor of the Technical Sciences, Professor, Head of the Department of Information Systems, E-mail: e.arsiriy@gmail.com, ORCID: https://orcid.org/0000-0001-8130-9613, Odessa, Ukraine

Oksana Yu. Babilunha[1], Candidate of the Technical Sciences, Department of Information Systems, E-mail: babilunga.onpu@gmail.com , ORCID: https://orcid.org/0000-0001-6431-3557, Odessa, Ukraine

Olga S. Manikaeva[1]**,** postgraduate student Department of the Information Systems, E-mail: manikaeva@gmail.com, ORCID: http://orcid.org/0000-0002-0631-8883, Odessa, Ukraine

Oleksii I. Rudenko[1], Department of the Information Systems, E-mail: icetear.gm@gmail.com, ORCID: https://orcid.org/0000-0002-0753-2443, Odessa, Ukraine[1]Odessa National Polytechnic University, Shevchenko Ave., 1, Odessa, Ukraine, 65044

AUTOMATION OF THE PREPARATION PROCESS WEAKLY-STRUCTURED MULTI-DIMENSIONAL DATA OF SOCIOLOGICAL SURVEYS IN THE DATA MINING SYSTEM

Abstract. *In order to obtain knowledge about the target audience, the preparation process of weakly-structured multi-dimensional data of sociological surveys were automated. The following techniques have been developed for automating data preparation: machine representation, preprocessing of the data from the sociological surveys in order to clean and filter it, transformation of data into feature space based on a formalized research objective, nonlinear dimensionality reduction and visualization of the multi-dimensional data. As research has shown, the procedures associated with obtaining of primary and secondary feature spaces are the most significant. The Orange3 framework, which includes component-based data mining software and is used as a module for Python, were used to create IT of preparing weakly structured multidimensional data of sociological surveys in the Data Mining system. Approbation of the automated preparation process of weakly-structured multi-dimensional data within the sociological surveys Data Mining system allowed to increase the reliability of decision-making on the lifestyle of the respondents compared to a sociologist of the master's qualification level and the respondents own responses***.

Keywords: *information technology; data mining; pre-processing; preparation process of data; sociological surveys*

Introduction and formulation of the research problem. Conducting various sociological surveys, questioning or interviewing with subsequent analysis of the resulting empirical weakly-structured multi-dimensional data is the most common way to obtain information about the target audience in the form of *patterns* – non-trivial knowledge, having considerably smaller dimensionality as compared with the original data [1].

Patterns consist of structured quantitative metrics or qualitative assessments of behavioral, social demographic, geographical, psychophysical, medical or any other characteristics of the data. Such patterns can be presented in a convenient visual form and with the interpretation of an expert-sociologist; they turn into *knowledge* about the target audience. Such knowledge helps experts in the subject area to make informed decisions when planning business strategies. For example, in order to make decisions about the organization of a specialized library and information services, an expert analyst needs to assess the interests, requirements and preferences of the target audience, using personal data (gender, age, educational background), activity (frequency and purpose of library visits, the basis for literature choice), literary preferences (entertaining, professional, etc.) [2].

A complete examination of a graduate's knowledge with the subsequent formation of recommendations on the specifics of employment is impossible without assessing data from sociological surveys on the adequacy of curricula and technology classes, etc. [3].

As a result, a class of data mining tasks has been formed. *Data mining* makes it possible to detect previously unknown, practically useful and accessible interpretations of knowledge necessary for making decisions regarding the target audience being studied in raw empirical sociological data [4]. The current level of development of information technologies allows the automation of the *preparation process* of multi-dimensional empirical weakly structured data of sociological surveys to obtain information about the target audience in the form of *patterns* of smaller dimensionality for visual representation in the decision space [5-6].

Analysis of existing scientific achievements and publications. From the point of view of the expert-sociologist, the process of obtaining information about the target audience based on the analysis of empirical data which consists of a cyclical sequence of the following mandatory steps [7-8]:

DOI://10.15276/hait.01.2018.1
ISSN 2663-0176 (Print)

1) Awareness of theoretical or practical insufficiency of available knowledge about the target audience.

2) Formulation of the problem and hypotheses (in qualitative research hypotheses are usually formulated at later stages of the study).

3) Collection (selection) of empirical material on the basis of which hypothetical assumptions can be confirmed or refuted.

4) Analysis of empirical data using various methods, strategies, research programs and models.

5) Interpretation of the processed data and decision-making, explanation of the social phenomenon.

6) Redefining and refining a problem or hypothesis leading to a new research cycle.

In the modern world, conducting sociological research is associated with the need to analyze large and rapidly growing volumes of empirical data that exceed a person's ability to process them [9-10]. Therefore, Data Mining methods are gradually becoming the most popular tools for an expert-sociologist. The advantages of the following technologies were used in the development of data mining technology:

– *online transaction processing (OLTP)*, which is limited to relational databases technology (*SQL*) as the main tool for efficient storage, retrieval and management of large amounts of data [4; 9; 10];

– *online analytical processing (OLAP)*, which is focused on the use of non-relational databases and data warehouses (*NoSQL*), includes data cleaning and integration and uses for data analysis functions such as consolidation, aggregation, summarizing, viewing information “from different angles” [4; 9; 10].

Traditionally, in the process of *Data Mining* the following stages are distinguished (Fig. 1) [4]:

1. Collection or *Selection* of data for analysis.

2. *Preprocessing*. At this stage missing values, inconsistencies and random “noise” are imputed from the source data (cleaning/filtering), data is merged from several possible sources into one storage (integration).

3. *Transformation*. At this stage, the data is converted to a form suitable for analysis. Data aggregation, attribute discretization, data compression, and dimensionality reduction are often used.

4. *Data mining*. Within this stage, Data Mining algorithms are used to extract patterns.

5. *Interpretation/Evaluation* of found patterns. This stage may include visualization of the extracted patterns, identification of useful patterns based on the utility function.

6. Use of interpreted/visualized patterns and new *Knowledge* by subject matter experts.

The tasks of data processing solved at the stages of selection, preprocessing and transformation of data that precede the actual data mining are proposed to be combined into a comprehensive stage of *Data Preparation* for further modeling.

Let's consider some problems in automating the process of preparing sociological surveys data in terms of Data Mining.

At the stage of collecting (selecting) data for analysis (Fig. 1), an expert-sociologist, based on the existing knowledge of the target audience or its insufficiency, solves the problem of *hypothesizing*, which must be confirmed or refuted as a result of a sociological research. At the same time, the hypothesis at the qualitative level establishes a connection between empirical sociological facts or groups of facts and some explanation of a sociological phenomenon. For example, as a formalized goal of a sociological survey, it is hypothesized that the dependent variable (lifestyle) changes depending on some reasons (quality of food, alcohol consumption, playing sports, etc.), which are independent variables. Next, an expert sociologist solves problems related to the actual collection (selection) of *targeted* sociological empirical data that can confirm/refute the hypothesis put forward. Such data can be defined as primary information of any kind, obtained as a result of one of the many types of sociological information gathering [7-8].

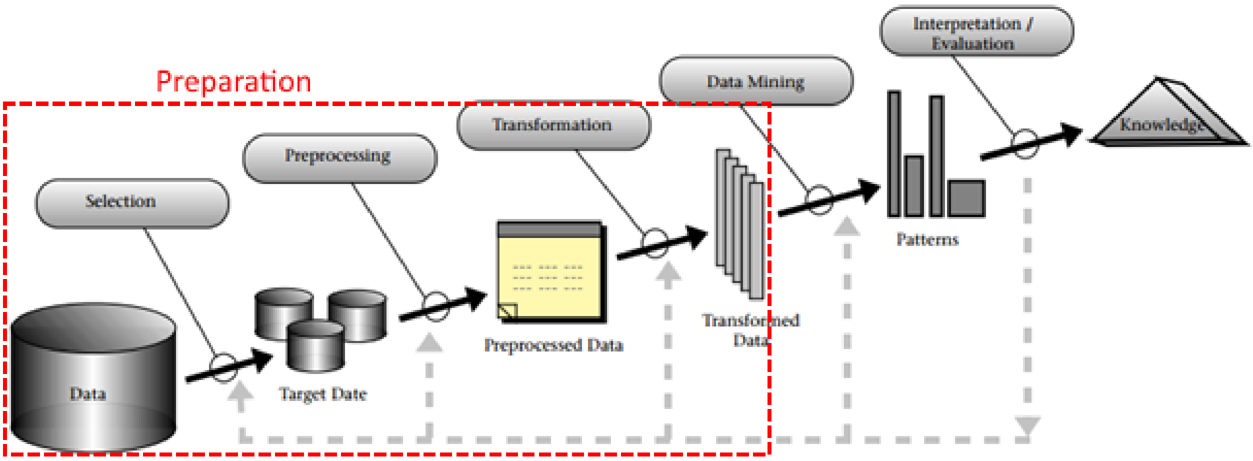


Fig. 1. The sequence of knowledge extraction stages based on the Data Mining approach

As a rule, to conduct in-depth analytical studies, data is collected through questionnaires and interviews with a “complex” structure. Any empirical data is always somehow structured. Depending on the degree of structuration of the data, it is divided into the following types [7-8]:

– *unstructured data* is text-type data obtained in the process of conducting various types of interviews, texts of answers to open questions, and any other texts or documents that the sociologist refers to;

– *strongly-structured data* is data that exists in the form of matrices of any type (i.e. tables), obtained through a part of the questionnaire (interview) according to the closed-question scheme;

– *weakly-structured* data is data of an intermediate type, not only quantitative, but also categorical (nominal and ordinal) types, as well as data which exists in textual form. However, this type of data is specially organized. Examples of such data are data with a limited number of unique values or categories. As a rule, they contain answers to open questions of the questionnaire (interview) with a limited field of searching for answers, either obtained by the method of unfinished sentences (test for sentence completion) or by the method of repertory grids (G. Kelly's theory of personality constructs) [11]

At the stages of *Data Preprocessing* and *Data Transformation* (Fig. 1), an expert-sociologist is faced with the task of formalizing and structuring data related to the organization of *multi-dimensional feature space*. Hence, special methods of text recognition and analysis (*OCR* and *Text Mining*) are used for the transformation of unstructured data into the feature space for carrying out further modeling [12]. Strongly-structured data, namely a quantitative data type is easily formalized and automatically transformed into a multi-dimensional feature space. For formalization and structuring of weakly-structured data, various methods of data preprocessing are used: cleaning, filtering, normalization, encoding, etc. As a result of such preprocessing, weakly-structured data of sociological surveys is transformed into a multi-dimensional feature space. Before applying data mining techniques, the task of *dimensionality reduction* of the feature space should be solved.

Thus, the complexity of developing Data Mining technologies and systems for solving problems of extracting knowledge about the target audience is associated with the need at the *Preparation stage* of using data which is weakly-structured and uncertain because it is collected from various sources, interpreted using different scales and often contradicts itself. An expert decision to classify the respondent to a particular class based on the analysis of the characteristics of such data is ambiguous and depends on the qualification of an expert-sociologist. This leads to the need to automate the process of preparation of weakly-structured multi-dimensional data of sociological surveys in Data Mining Systems.

The goal and tasks of the research. The goal of the research is to develop an information technology to automate the preparation process of weakly-structured multi-dimensional data of sociological surveys in order to extract knowledge about the target audience, which will increase the accuracy of decision-making.

To achieve this goal, the following tasks must be solved:

1. Development of a preprocessing method for sociological surveys data for the purpose of cleaning and filtering it.

2. Development of a transforming method for preprocessed data, hence we can turn it into a feature space, taking into account the formalized goal of the study.

3. Using the technique of nonlinear dimensionality reduction and visualization of multi-dimensional variables of the transformed data.

Research methods. As a scientific basis of the research, a systematic approach is used when conducting social surveys, questioning and interviewing and methods and technology of Data Mining sociological data. Object-oriented design was used to develop an information technology. Methods of computer simulation were used in the approbation of information technology.

Presentation of the main research material To create the information technology to automate the preparation process of weakly-structured multi-dimensional data of sociological surveys at the *first stage*, a preprocessing method of sociological surveys data was developed with the goal of cleaning and filtering the data. The preprocessing method of sociological surveys data requires the following steps:

1) Presentation of text data and their names in English, using the *Google translate* and storing data in open text format for structured hierarchical data, such as comma-separated values (*csv*);

2) Checking the data of sociological surveys for compliance with the range of permissible values of each variable, removal of invalid values;

3) Identifying variables which contain a large number of missing values, deleting such variables if they are not informative, or reducing the number of missing values by replacing them with a mean/most frequent value.

As a result of such preprocessing, the amount of empirical data which is subject to further processing decreases on average by 25 %. Hence, (Fig. 2) shows statistics on the number of variables that contain missing values for the sociological survey “Ukraine – lifestyle”. The original dataset contained the answers of 1143 respondents to 114 questions [13].

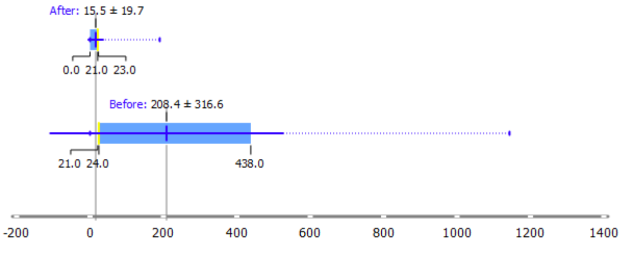


Fig. 2. Boxplot of an average number of missing values in the dataset “Ukraine – lifestyle” per variable before and after variables imputation

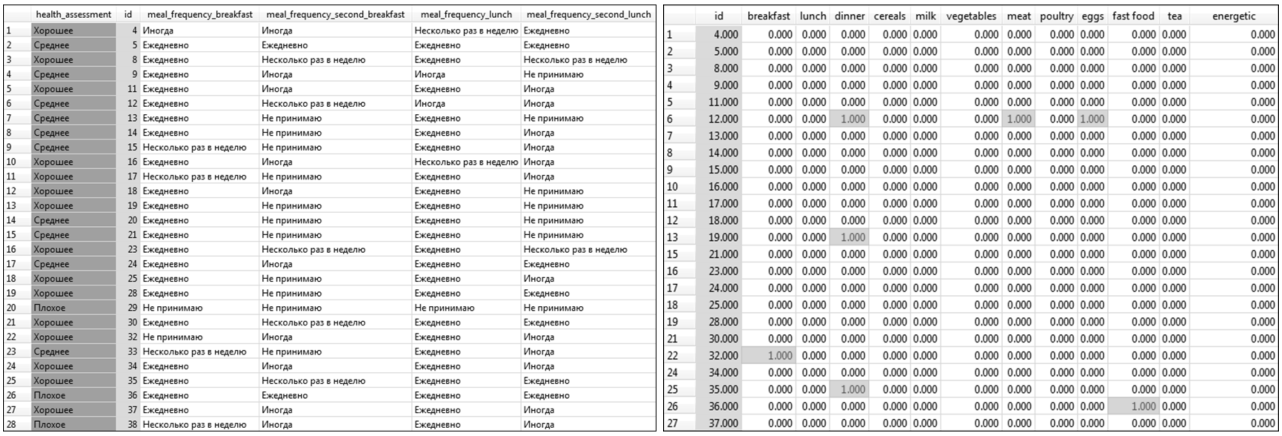
When determining variables with a large number of missing values, the number of missing values was calculated for each variable in the dataset. The results in the form of minimum, maximum, lower and upper quartiles, mean and standard deviation are displayed in the form of a box plot *before* and *after* variables imputation (Fig. 2).

At the *second stage*, a method was developed for transforming the preprocessed data of sociological survey into the feature space taking into account the goal of the research. The transformation technique consists of three steps:

1) Formalization of the purpose of sociological research, i.e. the *hypothesis* that the *dependent variable* changes depending on the *independent variables*. For example, for the sociological survey “Ukraine – lifestyle” as a formalized goal, the hypothesis “lifestyle” is set. According to it, 1143 respondents participating in the survey should be divided into three so-called *meta-classes* “good”, “average” and “bad”. Classification of respondents is made depending on their answers to 114 questions of the questionnaire, in which they were asked to determine, for example – the quality of their food, alcohol consumption, sports, etc.

2) Definition of independent variables that are *relevant* to the research's goal and removal of irrelevant variables from the *primary feature space*; from the point of view of an expert-sociologist, only variables, which affect directly or indirectly the respondent's lifestyle are relevant. For example, for the sociological survey “Ukraine – lifestyle”, variables which describe information sources and taste preferences were removed from the primary feature space, as well as motives for engaging in sports, etc.

3) Development of procedures for the normalization and binarization of variables relevant to the goal of the sociological research. For the subsequent use of Data Mining methods, the values of variables should be on the same scale. Continuous variables are reduced to equal ranges, for example, from – 1 to 1, or from 0 to 1. The values of “yes” and “no” or “male” and “female” are binarized, the values of categorical variables, if necessary, are converted to a Likert like scale, and then converted using *One-Hot Encoding* (*OHE*) [14]. In particular, a feature that characterizes sport activity can take 5 different values: “I don't exercise at all”, “I exercise rarely and irregularly”, “I exercise several times a month”, “I exercise several times a week”, “I exercise everyday”. In this case, after encoding a feature using *OHE*, 5 binary features are created, the values of all of which are zero except for one. Fig. 3 shows fragments of values in the dataset “Ukraine – lifestyle” after the first and second stages of data preparation process.



	health_assessment	id	meal_frequency_breakfast	meal_frequency_second_breakfast	meal_frequency_lunch	meal_frequency_second_lunch
1	Хорошее	4	Иногда	Иногда	Несколько раз в неделю	Ежедневно
2	Среднее	5	Ежедневно	Ежедневно	Ежедневно	Ежедневно
3	Хорошее	8	Ежедневно	Несколько раз в неделю	Ежедневно	Несколько раз в неделю
4	Среднее	9	Ежедневно	Иногда	Иногда	Не принимаю
5	Хорошее	11	Ежедневно	Иногда	Ежедневно	Иногда
6	Среднее	12	Ежедневно	Несколько раз в неделю	Иногда	Иногда
7	Среднее	13	Ежедневно	Не принимаю	Ежедневно	Не принимаю
8	Среднее	14	Ежедневно	Не принимаю	Ежедневно	Иногда
9	Среднее	15	Несколько раз в неделю	Не принимаю	Ежедневно	Иногда
10	Хорошее	16	Ежедневно	Иногда	Несколько раз в неделю	Иногда
11	Хорошее	17	Несколько раз в неделю	Не принимаю	Ежедневно	Иногда
12	Хорошее	18	Ежедневно	Иногда	Ежедневно	Не принимаю
13	Хорошее	19	Ежедневно	Не принимаю	Ежедневно	Не принимаю
14	Среднее	20	Ежедневно	Не принимаю	Ежедневно	Не принимаю
15	Среднее	21	Ежедневно	Не принимаю	Ежедневно	Не принимаю
16	Хорошее	23	Ежедневно	Несколько раз в неделю	Ежедневно	Несколько раз в неделю
17	Среднее	24	Ежедневно	Иногда	Ежедневно	Ежедневно
18	Хорошее	25	Ежедневно	Не принимаю	Ежедневно	Иногда
19	Хорошее	28	Ежедневно	Не принимаю	Ежедневно	Ежедневно
20	Плохое	29	Не принимаю	Не принимаю	Не принимаю	Не принимаю
21	Хорошее	30	Ежедневно	Несколько раз в неделю	Ежедневно	Ежедневно
22	Хорошее	32	Не принимаю	Иногда	Ежедневно	Иногда
23	Среднее	33	Несколько раз в неделю	Не принимаю	Ежедневно	Иногда
24	Хорошее	34	Ежедневно	Иногда	Ежедневно	Иногда
25	Хорошее	35	Ежедневно	Несколько раз в неделю	Ежедневно	Ежедневно
26	Плохое	36	Ежедневно	Ежедневно	Ежедневно	Ежедневно
27	Хорошее	37	Ежедневно	Иногда	Ежедневно	Иногда
28	Плохое	38	Несколько раз в неделю	Иногда	Ежедневно	Иногда

a

	id	breakfast	lunch	dinner	cereals	milk	vegetables	meat	poultry	eggs	fast food	tea	energetic
1	4.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	5.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	8.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	9.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	11.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	12.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
7	13.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	14.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	15.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	16.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	17.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	18.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
13	19.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	21.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	23.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	24.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	25.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	28.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
21	30.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
22	32.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
24	34.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25	35.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
26	36.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
27	37.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

b

Fig. 3. Fragments of data in the dataset “Ukraine – lifestyle”:
a – after preprocessing;
b – after transformation into a primary feature space

At the *third stage*, before using Data Mining methods such as cluster analysis for grouping respondents, and classification for subsequent assignment of new respondents to one of the existing groups, was used techniques of dimensionality reduction and visualization of multi-dimensional data of sociological surveys in the space of secondary features. When choosing a dimensionality reduction method for the feature space, the capabilities of t-distributed *Stochastic Neighbor Embedding* (*SNE*) and *Multi-Dimensional Scaling* (*MDS*) were compared to make it possible to visualize data of sociological surveys.

In general, this task can be formulated as follows: there is data of sociological surveys and each record of it is described by a multi-dimensional variable. It is necessary to obtain a new variable that exists in the two or three-dimensional *secondary* feature space, which would preserve the structure of the initial data as much as possible.

Originally, the classical *SNE* algorithm [15] transforms the multi-dimensional Euclidean distance between points into conditional probabilities, reflecting the similarity of points:

$$p_{ji} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)} . \quad (1)$$

Expression (1) shows how close the point x_j is to the point x_i with the Gaussian distribution of the values of Euclidean distances around x_i with a given deviation σ. The value of σ will be different for each point. It was chosen so that the points in areas with higher density have less variance. For this, the evaluation of perplexes is used:

$$Perp(P_i) = 2^{H(P_i)} , \quad (2)$$

where $H(P_i)$ – Shannon entropy in bits:

$$H(P_i) = -\sum_j p_{ji} \log_2 p_{ji}. \quad (3)$$

In this case, the perplexion can be interpreted as a smoothed estimate of the effective number of neighbors for point x_i. It is set as a parameter of the method. The authors recommend using a value in ranging from 5 to 50. The value of σ is determined for each pair of x_i and x_j using a binary search algorithm. For two-dimensional or three-dimensional mappings of the pair x_i and x_j, we denote them as y_i and y_j. Conditional probability y_i and y_j can be estimated using the formula (1).

$$q_{ji} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} . \quad (4)$$

If the mapping point’s y_i and y_j correctly model the similarity between the original points of high dimension x_i and x_j, then the corresponding conditional probabilities p_{ji} and q_{ji} will be equivalent. As an assessment of the quality with which q_{ji} reflects p_{ji}, the Kullback-Leibler divergence is used.

Studies show that the use of the classical *SNE* algorithm can be associated with difficulties in optimizing the loss function and the crowding problem [15]. The use of the *t-Distributed Stochastic Neighbor embedding* (t-SNE) algorithm significantly eases these difficulties [16].

The loss function *t-SNE* has two fundamental differences. First, *t-SNE* has a symmetric form of similarity in a multidimensional space and a simpler version of the gradient. Secondly, instead of the Gaussian distribution for points into the visualization space, the t-distribution (Student's distribution) is used, the use of which facilitates optimization and solves the problem of crowding.

By reducing the dimensionality of the feature space and obtaining "*secondary"* features using the *Multi-Dimensional Scaling* (*MDS*) algorithm [17], which is fed with the input matrix of pairwise distances *D* between objects of the set *S*. A matrix *X* is formed at the output, containing the coordinates (*x*, *y*) of the objects *S* on the plane. In this case, the matrix *D* may contain gaps.

The *MDS* algorithm is based on minimization of the mean square error of approximation of the initial distances of matrix *D*. The task of minimizing the function, called "stress" [18], should be solved:

$$S = \sum_{i,j} w_{ij}\left(D_{ij} - r_{ij}\right)^2, \tag{5}$$

summation in expression (5) is carried out over all pairs of points (*i*, *j*) from the set *S*, for which distances D_{ij} are given in matrix *D*, $w_{ij} = D_{ij}$, *Weight Power* are the weights of objects, $r_{ij}^2 = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$ – Euclidean distances between the *i*-th and *j*-th objects on the plane.

The exponent of *Weight Power* allows the expert to orient the process of placing points on a more accurate reflection of far (*Weight Power* > -2) or close (*Weight Power* < -2) distances. It is considered that the most adequate result is achieved when *Weight Power* = -2. In this case, the "stress" functional acquires the meaning of potential energy in a system of points connected by elastic bonds, and the minimization problem acquires a clear physical meaning for finding a stable equilibrium [18].

Multi-dimensional scaling let us to calculate the coordinates of objects in Euclidean space, based on known pairwise distances between them. In the current implementation, using the *MDS* algorithm (5), a space of "secondary" features is formed, which are specified using coordinates on the plane of 1143 objects (respondents), which are initially represented in the 23-dimensional feature space. The calculated coordinates of each respondent on a plane (similarity map) are set using the values of radius *R* centered at the origin of the coordinates (Fig. 4). On the similarity map, it's not the coordinates of points that are important, but their relative positions. Studies have shown that use of the *MDS* algorithm in comparison with the *t-SNE* is more preferable if the objects in the initial feature space (see Fig. 3, b) are represented using binary vectors.

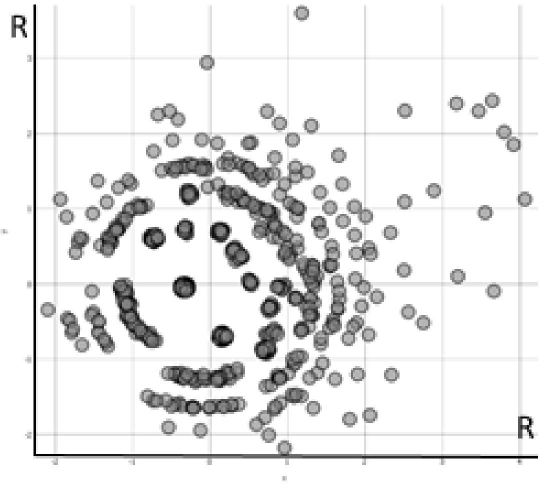


Fig. 4. Map of respondents similarity, built using MDS projections by the dataset "Ukraine – lifestyle"

Approbation of sociological surveys data preparation IT using the study "Ukraine – lifestyle"

The following basic tools were used to create *IT of preparing weakly structured multidimensional data of sociological surveys in the Data Mining system*: an interpretable, interactive, object-oriented programming language *Python*; *Orange3* framework, which includes component-based data mining software and is used as a module for Python; the *Scikit-Learn* library, which implements machine learning for Python, as well as interacting with the numeric and scientific libraries of Python *NumPy* and *SciPy*.

For approbation of the developed IT, the sociological surveys data obtained directly from 1143 respondents in the form of their answers to 114 questions on the topic "Ukraine – Lifestyle" [13] were used. The purpose of the analysis is to determine how healthy the respondents are. Initially, all respondents, depending on their responses to the questionnaire, should be divided into three meta-classes – "good", "average" and "bad".

The analysis of the effectiveness of the developed IT data preparing for automation of the first three data processing stages in the Data Mining system (see Fig. 1) was carried out using *classification accuracy* (*CA*) metric:

$$CA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where *TP* – true positive; *TN* – true negative; *FP* – false positive; *FN* – false negative solutions.

As an *input vector* (*feature vector*), data from two-dimensional space of "secondary" features were used, which was obtained using the *MDS*-projection of values of radius *R* with the center at the origin.

To obtain two-dimensional space of "secondary" features were used (Fig. 5):

– *case* 1 – 77-dimensional data, represented by variables which were obtained from the initial dataset after the first stage of IT data preparation (Fig. 5, a);

– *case* 2 – 23-dimensional data, represented by feature space after the second stage of IT data preparation (Fig. 5, b);

– *case* 3 – 36-dimensional data obtained by an ex-pert- sociologist as answers of respondents to 12 questions relevant to the research being execute (Fig. 5, c).

When visualizing the results of the classification of respondents in Fig. 5, the following classification solutions were used as an *output vector* (*target vector*):

– *case* 1 – a respondents' own decision about their belonging to one of the three meta-classes (Fig5,a);

– *case* 2 – an automated decision on the affiliation of respondents to one of 6 clusters, which was made by Data Mining system (Fig. 5, b);

– *case* 3 – a decision on the affiliation to one of the three meta-classes, which was taken by an expert-sociologist on the basis of using their own method (Fig. 5, c).

The method used by an expert-sociologist in the distribution of the respondents to the met classes based on the analysis of their responses to the “Ukraine – lifestyle” survey consisted of the following steps:

a) 3 meta-classes were defined, taking into account the formalized goal of the research;

b) 12 relevant features were determined for the subsequent respondent’s classification;

c) the response space for each relevant feature was divided into categories, in accordance with a given number of meta-classes;

d) if a respondent chose a definite answer from a category to a specific question, he'd get 1 point in the corresponding category;

e) the respondent belongs to the meta-class corresponding to the category in which he or she scored the maximum amount of points;

f) if the maximum cannot be determined, the respondent is ignored.

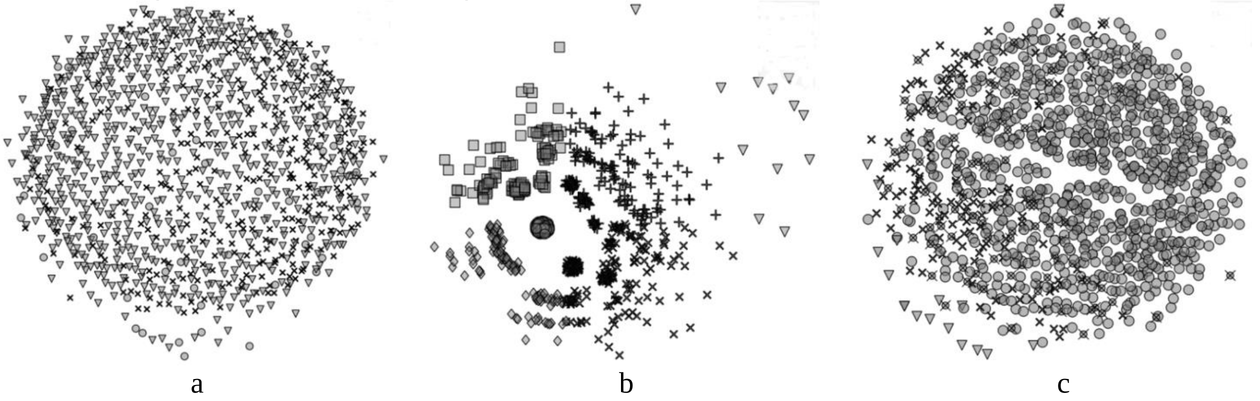


Fig. 5. Visual presentation of the results of the respondents classification by the dataset “Ukraine – lifestyle”:
a – in accordance with their own decision (3 meta-classes: ○ – “bad”, × – “average”, ▽ – “good”);
b – in accordance with an automated solution of the Data Mining system
(6 meta-classes ○ –“good”, × – “bad”, ▽ – “very bad”, + – “rather bad”, ◇ – “very good”, □ – “average”);
c – in accordance with the decision of an expert-sociologist
(3 meta-classes: ○ – “good”, × – “average”, ▽ – “bad”)

Analysis of the visual representation of classification results of respondents to the “Ukraine – lifestyle” sociological survey allows us to conclude that it *is impossible to obtain reliable decisions about the belonging of the respondents to certain meta-class without using data preparation IT*. Labels of the 3 meta-classes (Fig. 5, a & Fig. 5, c) are located in the intersecting areas. While for data obtained using the developed IT, the respondents fit into six practically non-intersecting classes (Fig. 5, b). Also, it was fairly easy to build classifying surfaces.

To obtain *quantitative characteristics of the reliability of classification decisions*, the following experiment was conducted. From the initial dataset, a new dataset was formed based on the respondents' answers to questions which are “guaranteed” by the expert to be significant for the analysis. They refer to two polar meta-classes – “bad” and “good”. Thus, 144 respondents were assigned to the “bad” class and 375 of them to the “good” class.

Figure 6 shows the classification accuracy of 519 respondents for the two polar meta-classes. The classification was carried out using the Data Mining system. Feature and target vectors were obtained from a new artificially formed dataset in accordance with the conditions of the previous experiment. According to it, feature vector consisted of data from two-dimensional secondary feature space, which were obtained using the *MDS*-projections of values of radius *R* centered at the origin for *cases* 1, 2 and 3.

The following classification decisions were used as the target vector: respondent's answer (*case* 1), Data Mining system's (*case* 2), and a expert-sociologist's (*case* 3).

The classification accuracy was determined according to (6).

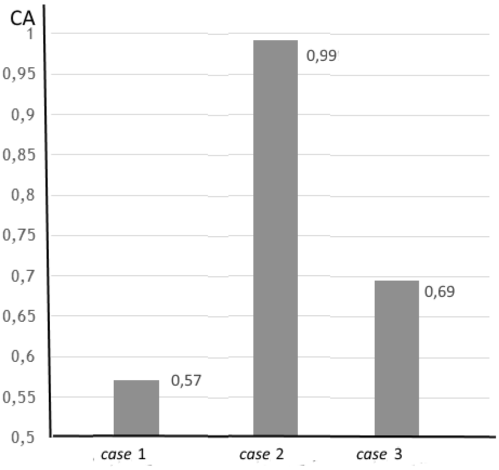


Fig. 6. The results of the accuracy of classification of respondents by lifestyle for two meta-classes with the proposed processing case

Thus, the use of the developed information technology made it possible to increase the accuracy of the classification of decisions for two artificially created polar classes by 30 % compared with the expert-sociologist and by 43 % compared with the respondents own decision.

Conclusions and prospects for further research. Information technologies have been developed to automate the preparation process of weakly-structured multi-dimensional data from sociological surveys to extract knowledge about the target audience. The technological solutions are based on the following developed techniques:

– preprocessing of the data from the sociological surveys in order to clean and filter it;

– transformation of data into feature space based on a formalized research objective;

– nonlinear dimensionality reduction and visualization of multi-dimensional variables.

As research has shown, the procedures associated with obtaining of “primary” and “secondary” feature spaces are the most significant. The use of the developed information technology by the dataset “Ukraine – lifestyle” made it possible to increase the accuracy of classification decisions for two artificially created polar classes by 30 % compared with the expert-sociologist and by 43 % compared with the respondents own decision.

The developed information technology for the preparation of sociological surveys data in conjunction with the Data Mining System is recommended for the analysis of multi-dimensional weakly-structured data obtained from various sociological surveys, questionnaires or interviews. Presented in a visual form convenient for experts, the knowledge gained about the target audience makes it easy to take it into account when making informed decisions.

References

1. Aleskerov, F. T., Belousova, V. Yu., Yegorov, L. G., & Mirkin, B. G. (2013). Analiz dannyh i intellektual'nye sistemy, Analiz patternov v statike i dinamike [Data analysis and intelligent systems, Analysis of Patterns in Statics and Dynamics], Part 1: Literature Review and Concept Refinement. Interdisciplinary Scientific and Practical Journal of Nru Hse “Business Informatics”, No. 3 (25), pp. 3-18 (in Russian).

2. Arsiriy, E. A. (2011). Nejronnye seti raspoznavanija obrazov chitatelej publichnoj biblioteki dlja organizacii specializirovannyh bibliotechnyh uslug [Neural network pattern recognition of readers of the public library for organizing specialized library services]. Proceedings of the Odessa Polytechnic University, Ukraine, No. 1 (35), pp. 118-124 (in Russian).

3. Arsiriy, E. A. (2009). Nejrosetevoe formirovanie integral'nyh professional'nyh harakteristik v sisteme distancionnogo obuchenija, [Neural network formation of integral professional characteristics in the system of distance learning], MOODLE. Proceedings of the Odessa Polytechnic University, Ukraine, No. 2 (32), pp. 161-166 (In Russian).

4. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). “From Data Mining to Knowledge Discovery in Data bases” [Текст], AI Magazine, Vol. 17, No. 3, pp. 37-54. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

5. Rudenko, A. I., & Arsiry, E. A. (2018). Metod intellektual'nogo analiza slabo strukturirovannyh mnogomernyh dannyh sociologicheskih oprosov [Method of intellectual analysis of weakly structured multidimensional data of sociological surveys]. Materials of the Eighth International Scientific Conference of Students and Young Scientists “Modern Information Technologies 2018”, Ministry of Education and Science of Ukraine, Odessa National Polytechnic University; Institute of Computer

Systems, Odessa, Ukraine, *Ecologic Publ.*, pp. 168-169 (in Russian).

6. Arsiry, E. A. (2015). Razrabotka podsistemy podderzhki prinjatija reshenij v sistemah raspoznavanija obrazov nejrosetej s ispol'zovaniem statisticheskoj informacii [Development of a subsystem for decision-making support in systems of neural network pattern recognition using statistical information]. East European Journal of Advanced Technologies. Mathematics and Cybernetics – Applied Aspects. Vol. 6, No. 4 (78), pp. 4-12 (in Russian).

7. Semenov, V. You. (2009). Analiz i interpretacija dannyh v sociologii: uchebnoe posobie. Vladimirskij gosudarstvennyj universitet [Analysis and interpretation of data in sociology: a tutorial. Vladimir State University], Vladimir, Russian Federation, *VlSU Publ.*, 131 p. (in Russian).

8. Kislova, O. N. (2005). Intellektual'nyj analiz dannyh: vozmozhnosti i perspektivy ispol'zovanija v sociologicheskih issledovanijah [Intellectual Data Analysis: Opportunities and Prospects for Use in Sociological Studies]. Methodology, theory and practice of sociological analysis of modern society: Collection of scientific works. Kharkiv, Ukraine, pp. 237-243 (in Russian).

9. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0: Step- by-Step Data Mining Guide. SPSS, Copenhagen

10. Wirth, R., Hipp, J. (2000). “CRISP-DM: Towards a Standard Process Model for Data Mining”, Proceedings of the 4-th international conference on the practical applications of knowledge discovery and data mining, pp. 29-30.

11. Tararuhina, M. I. (2018). Technique of repertory grids by J. Kelly [Digital Resource]. – Available at : https://studfiles.net/preview/594463/ (In Russian).

12. Lande, D. (2003). Deep text analysis – technology for efficient analysis of text data. Data mining [digital resource]/ – Available at : http://dwl.kiev.ua/art/dz/ index.html (In Russian).

13. The data of the sociological research “Ukraine-style of life” [Digital Resource]/ – Available at : http://edukacjainauka.pl/ limesurvey/index.php/lang-pl. – 01.11.18.

14. Guido, S., & Müller, A. C. (2016). “Introduction to Machine Learning with Python: A Guide for Data Scientists”, O’Reilly Media.

15. Hinton, G. E., & Roweis, S. T. (2003). “Stochastic Neighbor Embedding. In Advances in Neural Information Processing Systems*”,* Vol. 15, pp. 833-840.

16. L. J. P. van der Maaten, & G. E. Hinton (2008). “Visualizing data using t-SNE”. Journal of Machine Learning Research, 9 (Nov): pp. 2431-2456.

17. Buja, A., Swayne, D. F, Littman, M. L, Dean N., Hofmann H., & ChenData L. (2008), Visualization With Multidimensional Scaling, Journal of Computational and graphical Statistics, Vol. 17, Issue 2, pp. 444-472.

18. Wojciech, B. (2001). “Visualisation of Abstract Data” [Текст], Basalaj Wojciech, Technical reports published by the University of Cambridge Computer Laboratory is freely available via the Internet, ISSN 1476-2986. – Available at : https://www.cl.cam.ac.uk/techreports/ UCAM-CL-TR-509.pdf.

Received 18.11.2018

[1]**Арсірій, Олена Олександрівна,** д-р техніч. наук, проф., зав. каф. інформаційних систем,
E-mail: e.arsiriy@gmail.com,, ORCID: https://orcid.org/0000-0001-8130-9613, Одеса, Україна
[1]**Бабілунга, Оксана Юріївна,** канд. техніч. наук, доцент каф. інформаційних систем,
E-mail: babilunga.onpu@gmail.com, ORCID: https://orcid.org/0000-0001-6431-3557, Одеса, Україна
[1]**Манікаєва, Ольга Сергіївна,** аспірант каф. інформаційних систем, E-mail: manikaeva@gmail.com,
ORCID: http://orcid.org/0000-0002-0631-8883, Одеса, Україна
[1]**Руденко, Олексій Ігорович,** каф. інформаційних систем, E-mail: icetear.gm@gmail.com,
ORCID: https://orcid.org/0000-0002-0753-2443, Одеса, Україна
[1]Одеський національний політехнічний університет, пр. Шевченка, 1, Одеса, Україна, 65044

АВТОМАТИЗАЦІЯ ПРОЦЕСУ ПІДГОТОВКИ СЛАБО СТРУКТУРОВАНИХ БАГАТОВИМІРНИХ ДАНИХ СОЦІОЛОГІЧНИХ ОПИТУВАНЬ В СИСТЕМІ DATA MINING

Анотація. *Для отримання знань про респондентів соціальних досліджень при розробці інформаційної технології інтелектуального аналізу автоматизовано етап підготовки слабо структурованих багатовимірних даних соціологічних опитувань. Для автоматизації підготовки даних розроблено інформаційну технологію яка базується на наступних методиках: машинного представлення , очищення та фільтрації даних; трансформації очищених даних в простір первинних ознак з урахуванням формалізованої мети дослідження; нелінійного зниження розмірності багатовимірного простору первинних ознак для побудови двовимірного простору вторинних ознак та їх подальшої візуалізації. Апробація інформаційної технології підготовки багатовимірних слабо структурованих даних спільно з системою Data Mining на даних соціологічних опитувань дозволила підвищити достовірність прийняття рішень по стилю життя респондентів у порівнянні з соціологом кваліфікаційного рівня магістр та із власним визначенням респондентів. Як показали дослідження розробленої інформаційної технології підготовки даних соціологічних опитувань, найбільш впливовими на результат аналізу є процедури, що пов'язані з побудовою просторів первинних і вторинних ознак для подальшого проведення кластерізації та класифікації Представлені в зручному для експертів візуальному вигляді, отримані знання про досліджувану цільової аудиторію дозволяють легко враховувати їх при прийнятті обґрунтованих рішень фахівцями в предметної області.*

Ключові слова: *інформаційна технологія; інтелектуальний аналіз даних; системи Data Mining; попередня обробка; соціологічні опитування*

[1]**Арсирий, Елена Александровна,** д-р технич. наук, проф., зав. каф. информационных систем,
E-mail: e.arsiriy@gmail.com, ORCID: https://orcid.org/0000-0001-8130-9613, г. Одесса, Украина
[1]**Бабилунга, Оксана Юрьевна,** канд. технич. наук, доцент каф. информационных систем,
E-mail: babilunga.onpu@gmail.com, ORCID: https://orcid.org/0000-0001-6431-3557, г. Одесса, Украина
[1]**Маникаева, Ольга Сергеевна,** аспирант каф. информационных систем, E-mail:
manikaeva@gmail.com, ORCID: http://orcid.org/0000-0002-0631-8883, г. Одесса, Украина
[1]**Руденко, Алексей Игоревич,** каф. информационных систем, E-mail: icetear.gm@gmail.com,
ORCID: https://orcid.org/0000-0002-0753-2443, г. Одесса, Украина
[1]Одесский национальный политехнический университет, пр-т Шевченко, 1, г. Одесса, Украина

АВТОМАТИЗАЦИЯ ПРОЦЕССА ПОДГОТОВКИ СЛАБО СТРУКТУРИРОВАНИХ МНОГОМЕРНЫХ ДАННЫХ СОЦИОЛОГИЧЕСКИХ ОПРОСОВ В СИСТЕМЕ DATA MINING

Аннотация. *Для получения знаний о респондентах социальных исследований при разработке информационной технологии интеллектуального анализа автоматизирован этап подготовки слабо структурированных многомерных данных социологических опросов. Для автоматизации подготовки данных разработана информационная технология, которая базируется на следующих методиках: машинного представления, очистки и фильтрации данных; трансформации очищенных данных в пространство первичных признаков с учетом формализованной цели исследования; нелинейного снижения размерности многомерного пространства первичных признаков для построения двумерного пространства вторичных признаков и их последующей визуализации. Апробация информационной технологии подготовки многомерных слабоструктурированных данных в совместно с системой интеллектуального анализа на данных социологических опросов позволила повысить достоверность принятия решений по стилю жизни респондентов по сравнению с социологом квалификационного уровня магистр и собственному определению респондентов***.*** *Как показали исследования разработанной информационной технологии подготовки данных социологических опросов, наиболее влияющими на результат анализа являются процедуры, связанные с получением пространств первичных и вторичных признаков. Представленные в удобном для экспертов визуальном виде, полученные знания об исследуемой целевой аудитории позволяют легко учитывать их при принятии обоснованных решений специалистами в предметной области.*

Ключевые слова: *информационная технология; интеллектуальный анализ данных; системы Data Mining; предварительная обработка; социологические опросы*