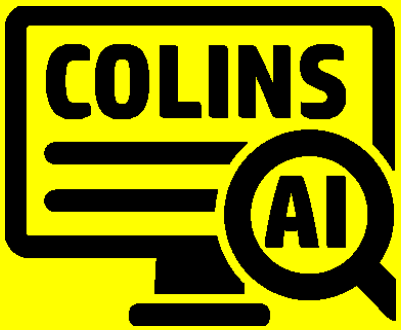
Vasyl Lytvyn
Victoria Vysotska
Thierry Hamon
Natalia Grabar
Natalia Sharonova
Olga Cherednichenko
Olga Kanishcheva
(Eds.)



COMPUTATIONAL LINGUISTICS AND INTELLIGENT SYSTEMS

Proceedings of the 4th International Conference, COLINS 2020. Volume I: Main Conference

Lviv, Ukraine
April, 2020

Lytvyn, V., Vysotska, V., Hamon, T., Grabar, N., Sharonova, N., Cherednichenko, O., Kanishcheva, O. (Eds.): Computational Linguistics and Intelligent Systems. Proc. 4th Int. Conf. COLINS 2020. Volume I: Workshop. Lviv, Ukraine, April 23-24, 2020, CEUR-WS.org, online

This volume represents the proceedings of the Workshop Conference, with Posters and Demonstrations track, of the 4th International Conference on Computational Linguistics and Intelligent Systems, held in Lviv, Ukraine, in April 2020. It comprises 83 contributed papers that were carefully peer-reviewed and selected from 165 submissions. The volume opens with the abstracts of the keynote talks. The rest of the collection is organized in two parts. Parts II contain the contributions to the Workshop COLINS Conference tracks, structured in two topical sections: (I) Computational Linguistics; (II) Intelligent Systems.

Preface

It is our pleasure to present you the proceedings of the Workshop Conference of COLINS 2020, the fourth edition of the International Conference on Computational Linguistics and Intelligent Systems, held in Lviv (Ukraine) on April 23-24, 2020.

The main purpose of the CoLInS conference is a discussion of the recent research results in all areas of Natural Language Processing and Intelligent Systems Development.

The conference is soliciting literature review, survey and research papers comments including, whilst not limited to, the following areas of interest:

- mathematical models of language;
- machine learning;
- discourse analysis;
- segmentation, tagging, and parsing;
- speech recognition;
- sentiment analysis and opinion mining;
- text categorization and topic modeling;
- text mining;
- information retrieval;
- artificial intelligence;
- information extraction;
- statistical language analysis;
- text summarization;
- data mining and data analysis;
- computer lexicography;
- social network analysis;
- question answering systems;
- web and social media;
- NLP applications;
- machine translation;
- intelligent text processing systems;
- memory systems and computer-aided translation tools;
- computer-aided language learning;
- corpus linguistics.

The language of COLINS Conference is English.

The conference took the form of oral presentation by invited keynote speakers plus presentations of peer-reviewed individual papers. There was also an exhibition area for poster and demo sessions. A Student section of the conference for students and PhD students run in parallel to the main conference.

This year Organizing Committee received 165 submissions, out of which 83 were accepted for presentation as a regular papers. The papers are submitted to the following tracks: discourse analysis (3 paper), data mining and data analysis (12 paper), web and social media (9 papers), segmentation, tagging, and parsing (2 papers), text categorization and topic modeling (7 paper), statistical language analysis (12 papers), text mining (7 paper), information retrieval (9 papers), artificial intelligence (29 papers), Natural Language Processing (32 papers), NLP applications (18 paper), corpus linguistics (9 papers), sentiment analysis and opinion mining (3 paper), computational lexicography (2 papers), automatic ontology building (5 paper), morphological analysis (3 paper), content analysis (5 papers), intelligent text processing systems (9 papers), intelligent computer systems building (21 papers) and problem of classification (4 papers). The papers directly deal with such languages: Ukrainian, Polish, Russian, Spanich, Island, English and Germany.

These papers and extended abstracts were published in this Volume I of COLINS 2020 proceedings.

The conference would not have been possible without the support of many people. First of all, we would like to thank all the authors who submitted papers to COLINS 2020 and thus demonstrated their interest in the research problems within our scope. We are very grateful to the members of our Program Committee for providing timely and thorough reviews and, also, for being cooperative in doing additional review work. We would like to thank the Organizing Committee of the conference whose devotion and efficiency made this instance of COLINS a very interesting and effective scientific forum.

April, 2020

Vasyl Lytvyn
Victoria Vysotska
Thierry Hamon
Natalia Grabar
Natalia Sharonova
Olga Cherednichenko
Olga Kanishcheva

Associative Methods as a Tool to Improve the Quality of Knowledge Control

Nataliya Maslova[1][0000-0002-9078-0973], Olha Polovynka[1,2][0000-0002-0575-4587]

[1]Donetsk National Technical University, Pokrovsk, Shibankova square, 2, Ukraine
[2]National Technical University "Kharkov Polytechnic Institute", Kharkov, Ukraine
masgpp2@gmail.com, olga.polovinka1@gmail.com

Abstract. One of the tools to control the level of knowledge is testing. When compiling a control test, it is necessary to ensure full coverage of educational material, to avoid the identity of educational and control tests. A common approach is random selection of test questions from a given list by their conditional numbers. The questions database should have a large size: the total number of questions in the database should significantly exceed the number of questions in a single test. But database size does not guarantee that the control test will not include questions that have already been used at the stages of training testing or the thematic control test. In addition, this approach does not take into account the logical relationship of questions, which may affect the reliability of knowledge assessment.

To solve this problem, use of associative rule search algorithms at the stage of selecting test tasks and including them in the control test was proposed. Associative methods allow identifying the frequency of occurrence of particular questions, to discard the most frequently used ones, to offer options for choosing subsequent questions, taking into account their thematic relationship.

The relative newness of research is the use of a modification of associative rule search algorithms in the formation of control test tasks. Practical use of the proposal allows increasing the completeness of the coverage of educational material, objectively evaluating student knowledge level.

Keywords: testing, quality of knowledge, educational material, associative methods.

1 Introduction

The constant development of the education system leads to the need to improve the methods of quality control of learning outcomes. Controlling, evaluating of the knowledge and skills of students is an integral part of the process of didactic diagnosis. Diagnosis involves the control, verification and evaluation of knowledge, and, in addition, the collection of statistical data, their analysis, prediction of further developments. One of the methods for monitoring and measuring student knowledge is testing. For its implementation, it is necessary to have test tasks, clear rules for conducting, analysis and processing of results. Testing allows you to determine the initial

level of knowledge, identify poorly understood topics, adapt the necessary materials in the learning environment, and make the cognition process more active and productive [1].

The advantages of testing are the accuracy of measuring the quality of knowledge, objectivity of evaluation and the economic efficiency of the procedure. But, despite the advantages, the testing method also has disadvantages. The main disadvantage is the complexity of compiling a test that would ensure reliably assesses the level of knowledge of the subjects, reduce the probabilistic component of the results repetition. The problem is especially acute when compiling a control test. In this case, compilers strive to ensure the fullest possible coverage of the educational material, to avoid the identity of educational and control tests.

The methods for selecting test tasks with the required level of knowledge of the subjects are not perfect. The database of questions has a sufficiently large size. The total number of questions should significantly exceed the number of questions in a single test. A common approach is to randomly select test questions from a given list by their conditional numbers. But this does not guarantee that the control test will not include questions that have already been asked at the stages of training testing or the thematic control. In addition, this approach does not take into account the logical relationship of issues and coverage of topics, which can affect the reliability of knowledge assessment.

A fairly new direction is the use of Data Mining in the process of knowledge control. Data Mining is one of the relevant and sought-after areas of processing structured and unstructured data of large volumes.

The direction includes many methods for processing data and detecting implicit and unknown dependencies. The revealed interactions and mutual influences of data are used in the future to solve non-trivial problems in various fields. Data Mining is particularly effective when analyzing large-scale unstructured data. The main tasks for the application of which intellectual analysis is used are the tasks of classification, clustering and the search for associative rules.

The methods of associative rules search are used to find typical sets of purchases (logistics), in cryptography (intrusion detection), to search for information on web resources (Web Mining), to solve production problems (continuous production), and to analyze medical and pharmacological information. Widely used methods of Data Mining in Microsoft software products, in the Oracle Data Mining system, in the development of Kaspersky Lab.

One of a new direction is the application of associative rules to the analysis of the educational process of the university, namely, to identify the relationship between academic performance and data on the educational process [2].

A prerequisite for applying the methods of searching for associations in the field of education is the rather large amounts of data of various structures. The task of monitoring the educational process, the search for factors affecting its effectiveness, the relationship of these factors is an important task of pedagogy. But the use of Data Mining and, in particular, associative methods in constructing tests and analyzing the results of knowledge testing, according to the authors, in scientific publications is missing.

The aim of the work is to demonstrate the results of the use of associative methods in the preparation of control tests, the analysis of qualitative indicators according to the experiment.

2 Associative Methods

Associative methods are a mechanism for searching for logical patterns between related elements, events or objects.

The problems of constructing associative rules include the task of determining the list of objects that are often found in a data set. The indicated possibility of associative methods allows sifting, or vice versa, to select frequently occurring sequences, to compile (create, form) filtered data sets that meet the requirements of the task.

Let A be a finite set of unique elements (list of items). In the general case, a set consists of n binary attributes called objects. From these components, many sets of items can be made up, such that $X \subseteq A$ [3].

The work of the associative method begins with the formation of associative rules of the form $X \rightarrow Y$ and sifting out non-informative values in a descending order.

First, sets of one element are considered. At the second iteration, two element sets are analyzed. On the third - three-elemental and so on, while the algorithm provides informative results, forms the rules. The volume of the analyzed data in this case is 2^T-1 and therefore finding all the frequent sets in the database is difficult.

Clipping uninformative values is performed using special indicators. The basic indicators that are used in almost all algorithms are support and confidence. The support metric shows how often a particular set of items is found in the database. Confidence is an indicator of how often the rule is true. In addition to these, there are a number of other indicators, for example, the indicator lift (the degree of connectedness of events), but for this study, other indicators were not used.

So, if X is a certain set of elements, then the support for the set of elements is calculated by the formula:

$$support(X)= T_s(X)/T,$$

where T_s (X) is the number of rows in the table satisfying the condition of joining the set (the number of transactions in the database that contain the set X), T is the total number of rows in the transaction table.

The confidence determination formula is:

$$con(X \rightarrow Y)= sup(X \rightarrow Y)/ sup(X).$$

At the end of each iteration, the Apriori method divides the sets of elements into two groups: the support value, which is greater than some predefined control value and those where this value is less than the support level.

Of the sets of elements, the Apriori algorithm leaves for further use only those of them whose support value is greater than some predetermined control value, that is, those for which the condition is satisfied:

$$\sup(X) \geqslant \text{podr}, \quad (1)$$

where podr is the minimum (control, threshold) value for selecting an element and inclusion in the set.

The procedure continues until all sets are considered.

From the sets of elements remaining after dropping out by condition (1), rules of the form $X \rightarrow Y$ are formed. Of these rules, those for which the condition is not fulfilled are eliminated:

$$\text{conf}\,(X \rightarrow Y) \geqslant \text{dost}, \quad (2)$$

where dost is the minimum value of the reliability of the rules.

Rules that have passed verification under condition (2) are the desired associative rules. For these rules, a conclusion is drawn about the relationship in the source data and the value of the next data element is predicted.

As noted, the size of the analyzed variants for the formation of the sample increases significantly from step to step, therefore, the analysis requires the use of special algorithms and programs.

Both commercial and freely distributed software products that implement the work of associative algorithms are presented on the market. The variety of algorithms is explained by the lack of unification in solving the problems of searching for associative rules and the use of various programming environments.

The most famous and frequently used associative rules search algorithms are Aprori, Eclat, and FP-Growth [4].

3 The Use of Associative Algorithms in Testing

The Apriori algorithm is designed to search for repeating sets of elements and to identify, on their basis, the relationships in large data sets.

In this case, an element is understood as a separate test task, the code of which is written in the database cell, as shown in Table 1.

A set is several elements of test tasks that occur simultaneously within the same test.

There is a methodology for determining the number of tasks recommended for inclusion in the test. This amount depends on many factors. For example, testing goals, audience, age of test participants, etc. But in general, a set, and in this case a test can include a fairly large number of test questions. So, in [5] information security tests are proposed, including from 150 to 450 questions. Apriori algorithms are designed to work with large amounts of data. If the length of the test is 150 questions, the number of possible combinations of 150 questions of 30 (without repetition) is 3.219878534049457e + 31, which makes it expedient to use association search methods.

The control test is composed of many test items. When compiling it is necessary to observe the following conditions:

- minimize the repetition of already asked questions;

- identification and inclusion of rarely used questions in the test;
- the exclusion of related and complementary issues.

The last requirement, according to the authors, should reduce the likelihood of guessing or intuitively receiving an answer.

A fragment of the database for processing and identifying the number of repetitions is shown in table 1. The sequence of the experiment is described in detail in [6].

Each test receives a unique code consisting of module numbers, topics, lectures, sections, test question number and the question type attribute (1 open-form question, 2 - compliance questions, 3 — sequence questions, 4 — closed-form questions).

For example, m1t02L04r07q13h2 - module 1, topic 2, lecture 4, section 7, question number 13, type of question - compliance questions (type 2).

Table 1. Database fragment

Number	Test Question Codes
T1	m1t02L02r02q03h4
T2	m1t02L02r01q06h4, m1t02L02r01q07h2
T3	m1t03L03r02q03h1
T4	m1t02L02r01q08h4, m1t03L04r02q03h4, m1t04L06r01q05h3, m1t05L07r02q02h2
T5	m1t02L02r02q03h4, m1t04L06r01q05h3, m1t02L07r01q08h2

This encoding is not redundant. On the contrary, it allows you to find questions related to one module, to the same topic, with the same number of lectures and sections. Then you can check if this question is unique.

It is this coding that avoids the lexical analysis of the content of the question in order to determine its uniqueness. The use of special codes made it possible to use the mechanism for searching for associative rules, the Apriori algorithm, to determine the frequency of repetition of a question and its uniqueness.

In our case, the Apriori algorithm has been modified. Scheme of the modified Apriori algorithm is shown in Figure 1.

The necessity to modify the original algorithm is due to the fact that the classic Apriori allows you to find the most commonly used sets, and to solve the problem you need to make a list of questions that are least used in tests conducted in the learning process.

If we use the above notation, then in our case, from the test sets, the modified algorithm leaves for later use only those whose support value is less than some predetermined control value, that is, those for which the condition is satisfied:

$$\sup(X) \leq \text{podr}$$

At the first iteration, elements whose support level is higher but not lower than the specified one were cut off. This allowed us to get a list of questions that appeared less frequently in work tests. At the second stage, pairs of non-repeating questions were

compiled so that they related to different topics and had a different level of complexity. At the same time, rules are formed that allow proposing whether or not to include next question in the test.

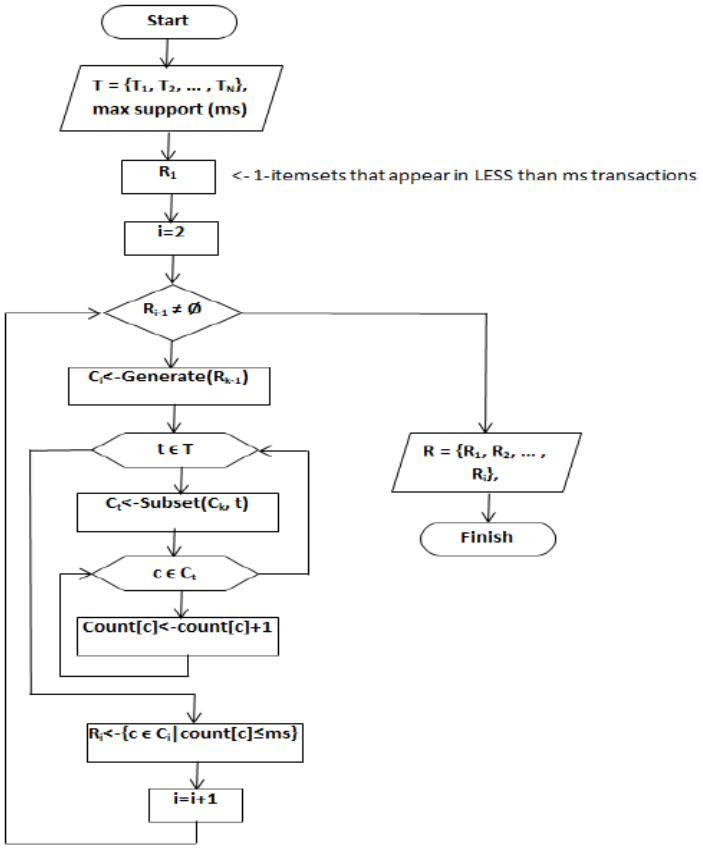


Fig.1. Scheme of the modified Apriori algorithm

4. Description of the Experiment

The research was conducted with the involvement of first, second and third year students who study information technology. As a base, information security tests were selected. The reason for this is the relevance and importance of ensuring information protection, extensive development in the field of testing on this topic, a significant amount of materials in the Internet space, the need for high-level knowledge in this direction.

In general, the research consisted of several stages.

At the first stage, the initial level of knowledge of students who did not have previous training in the field of information security (trial testing) was determined.

The initial level of knowledge of students who do not have skills in the field of information security obtained using ICT tests [7] is presented in Figure 2.

Fig.2. Entry level of students' knowledge

A gradual improvement in the quality of answers is apparently associated with the skills acquired over three years of training as a result of the use of computer equipment and information technologies that are necessary for every modern person.

Analysis of test answers showed that the largest number of correct answers were given to well-known questions in the field of information security. Answers to special questions contained the largest number of errors. This confirms the lack of preliminary training of the participants in the experiment.

The questions of this stage of the study are divided into three sections. The first section allows you to evaluate the knowledge of the general principles of data protection. The second analyzes the availability of Internet and email skills. The third contains questions on the use of social networks.

In general, the number of correct and incorrect answers, taking into account the topics of the questions, is presented in table 2.

Table 2. Made mistakes structure

Test questions topics	Mistakes (%)
The general principles of data protection	57,14%
The availability of Internet and email skills	22,86%
The use of social networks	20,00%
Total	100%

At the second stage, the degree of mastering the educational material was assessed. At this stage, tests were used, composed of the training materials of the original program of the course on information security. To master the learning materials, students could re-take trial testing. Tests were randomly selected using a random number generator. There was no ban on reusing the test question.

This stage for this study became a source of collecting material on the frequency of use of test questions. From an educational point of view, this stage has confirmed the need to control the level of assimilation of educational material. Students for whom the level of mastering the educational material was controlled mastered the material in

a shorter time, in a larger volume and with higher indicators (average value of the coefficient KSR1 = 0.87) than students in groups without control of this indicator.

At the third and last stage, control testing was organized, during which non-repeating questions were used, selected from the general list using the modified Apriori algorithm.

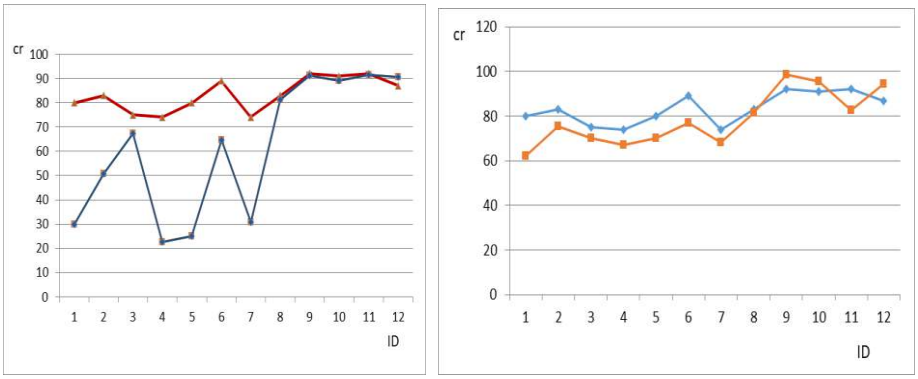
The third stage is the most interesting for the presented study.

5. The Obtained Results

The research was conducted with the involvement of first, second and third year students who take a course in information technology. As the base selected tests on information security. The reason for this is the relevance and importance of ensuring information protection, voluminous developments in the field of testing on this topic, a significant amount of materials in the Internet space, the need for high-level knowledge in this direction.

A preliminary test using various variants showed an initial knowledge level of 62%. In the learning process, all students took a basic course on the basics of information security and testing was repeated. The average mark of students rose to 80%, which indicates the achievement of a sufficient level of knowledge in the selected subject (Fig. 3, a)). The top line is the results after training. The line below is pre-test data.

To confirm the objectivity of the assessment, the test results were compared with students' exam scores. The pair correlation coefficient was 0.8 when using tests, the formation of which used the methods of intellectual analysis. When comparing examination scores with assessment results using tests of random selection of questions, the correlation coefficient was 0.83. Thus, the use of tests in the process of assessing students' knowledge, in the preparation of which the associative rules are used confirms a fairly high objectivity of assessment and can be recommended for use in the educational process (Fig. 3, b)).



a) results before and after training b) correlation dependence “test – exam”

Fig.3. Test results

To assess the completeness of the display of educational material, the value Pz = Yz / Q * 100 was calculated, where Yz is a parameter that shows how many different test questions were used in the preparation of the tests, Yz = Sum (Ai) and Q is the total number of test questions. The results of building tests using the associative rules search algorithm (Y2) and tests without processing by Data Mining methods (Y1) for test material from 150 questions are collected and shown on Figure 4.

№	quantity, Y_1	%, Y_1	quantity, Y_2	%, Y_2
1	127	84,7	138	92
2	113	75,3	142	94,7
3	136	90,7	143	95,3
4	128	85,3	149	99,3
5	116	77,3	137	91,3
6	141	94	139	92,7
7	139	92,7	148	98,7
8	125	83,3	146	97,3
9	119	79,3	143	95,3
10	132	88	138	92
average number	127	85,1	142	94,9

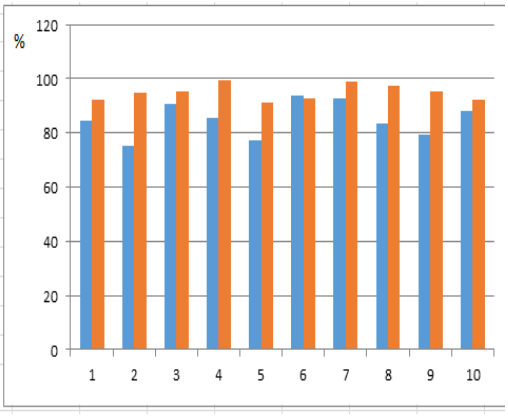


Fig.4. Completeness of study material

When conducting an experiment on a test of 150 questions, without using the methods of intellectual analysis in the formation of tests, 113 to 140 questions were recorded (an average of 127), which had not previously been encountered in preliminary and educational tests. In an experiment using the Apriori method, from 137 to 149 unique questions (on average 142) were used to build the test. Thus, when compiling tests, from 9 to 15 questions were selected from topics with a low percentage of use. The test, in the compilation of which associative methods were used, showed a higher (by 9.8%) completeness of the coverage of educational material.

5. Discussion and Conclusion

The paper presents the results of a study on the application of Data Mining methods in the preparation of control tests.

The authors developed an Apriori algorithm with cutting off the most common elements whose support level is higher, but not lower than indicated. This made it possible to obtain a list of questions that were less frequently encountered in working tests and formulate rules that allow inclusion in the test or rejection of the next question from the number of test items entered into the database, account its uniqueness and frequency of use.

It has been experimentally confirmed that the use of tests in the process of assessing students' knowledge in the preparation of which associative rules are used has a higher objectivity in assessing the level of knowledge.

In addition, associative selection revealed rarely used questions in the database of test items. The inclusion of such questions in the control test increased the completeness of the display of educational material, expanded the variability of the tests, and made it possible to more fully evaluate the knowledge of the subjects.

According to the results of the study, training programs were adjusted. Duplicate material and sections that do not carry a new information load have been removed. As a result, students mastered a large volume of educational material, achieved higher indicators with a ball assessment of knowledge.

The authors believe that the study has elements of scientific novelty, which consists in using a modification of the association search method in the formation of control test tasks. This allows you to ensure the completeness of the educational material, objectively assess the level of knowledge of students.

In the process of forming the knowledge base, along with the test tasks of the original development, the questions of the ICT security skills barometer test were used. Researchers were guided by the framework and requirements of the Erasmus + project "Digital competence framework for Ukrainian teachers and other citizens" (598236-EPP-1-2018-1-LT-EPPKA2-CBHE-SP), one of the participants of which is Donetsk National Technical University.

References

1. Bilyakovs′ka O. O., Myshchyshyn YA., Tsyura S. B.: Didactics of food schools [Dydaktyka vyshchoyi shkoly: navch. posib.], LNU imeni Ivana Franka, 360p. (2013)
2. Kazmina I.I., Nuzhnov E.V.: Modification of the Apriori algorithm for the analysis of the educational process of the university [Modyfikatsiya apriornoho alhorytmu dlya analizu danykh navchal′noho protsesu vuzu]. Yzvestyya PFU. Tekhnichni nauky, pp..28-39 . (2016)
3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Chapter 6. Association Analysis: Basic Concepts and Algorithms // Introduction to Data Mining. — Addison-Wesley (2006).
4. Shitikov V.K., Mastitsky S.E.: Classification, regression, Data Mining algorithms using R **[**Klassifikatsiya, regressiya, algoritmy Data Mining s ispol'zovaniyem R**].** - E-book, https://github.com/ranalytics/data-mining**,** (2017)
5. Hlobal′na systema testuvan, https://testserver.pro/index/pro/it/os, last accessed 2020/02/3
6. Maslova N.O., Polovynka O.L.: Application of association search methods for creating information security tests [Zastosuvannya methods for making an assessment at the time of testing with informational safety", Zbírnik naukovikh prats' Donets'kogo natsíonal'nogo tekhníchnogo uníversitetu] Seriya: "Computer science is a cybernetics and a calculating technique.", no.1 (29), pp 47-53. (2019).
7. ICT security skills barometer. Test your skills, http://dev.ecdl.lt/project/eguardtest/index.php?lang=en, last accessed 2020/01/15