ISSN 0147-6882, Scientific and Technical Information Processing, 2017, Vol. 44, No. 4, pp. 273–279. © Allerton Press, Inc., 2017.
Original Russian Text © A.E. Anisimova, A.A. Ryazanova, A.Yu. Shcherbakov, 2017, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 1: Organizatsiya i Metodika Informatsionnoi Raboty, 2017, No. 11, pp. 31–37.

A Semantic Kernel as a Universal Instrument for Classification and Systematization of Unstructured Information in the Field of Human Capital

A. E. Anisimova[a], *, A. A. Ryazanova[a], and A. Yu. Shcherbakov[b], **

[a] *All-Russian Institute of Scientific and Technical Information, Russian Academy of Sciences, Moscow, 125190 Russia*
[b] *Federal Research Center Computer Science and Control, Russian Academy of Sciences, Moscow, 119333 Russia*
**e-mail: ushsekr@viniti.ru*
***e-mail: x509@ras.ru*

Received August 29, 2017

Abstract—This paper considers a technique and software complex for isolating the semantic kernel of a text array to ensure a high degree of accuracy in the selection of CVs for relevant vacancies, the requirements for which are specified in an arbitrary text form. Such selection makes it possible to compile a stable database of vacancies and a corresponding resume database. The availability of these two databases opens the path for further highly effective automated analysis of key skills, implemented on the basis of the semantic kernel.

Keywords: human capital, large data, employment sites, information search, semantic core, cluster analysis, joint use of words

DOI: 10.3103/S0147688217040050

INTRODUCTION

From the point of view of human-resource management, the main efforts in the process of regulating the labor market at the current level are aimed at removing obstacles in the search for a workplace and an appropriate employee. There are two main reasons that qualified specialists do not find a suitable workplace. The first of these is related to the fact that the employee and employer describe the same range of skills and competencies using different terms and use different names for vacancies. The reverse situation is also possible, and even more common, where one term (for example, a technologist) describes fundamentally different specialties (for example, a food-production technologist, a sewing-equipment technologist, or a engineering communications technologist). In this case, it is necessary to use more sophisticated search systems than the search for the headers of ads, which is established in almost all modern employment sites.

One of the purposes of this study is to determine the range of specialties that are of greatest interest to prospective students in terms of future employment and wages. To compile a list of key specialties, automatic tools that determine the ratio of vacancies and resumes are important. From the list of resumes that have a suitable name, based on the semantic analysis of the text, it is suggested to delete CVs that in fact do not meet the requirements of employers. When the number of vacancies exceeds the number of resumes, one can question the prospects of the specialty in the labor market. In the reverse situation, a conclusion can be made about the diminishing importance of the profession or about the lack of correspondence between the qualifications of the available specialists and the requirements set by employers.

Technological aspects of research in the field of human capital development are related to the means of interaction among the labor market, the education system, and the technological infrastructure, which makes it possible to find specialists who are suitable for a particular job, as well as to reveal systemic gaps in the training of workers. The algorithm of interaction in the first approximation can be presented in the form of a "black box," where the input contains the requirements of the employer for candidates for vacancies and characteristics of real employees and their resumes and the output is the conclusion about the compliance of qualifications of employees with the labor market requirements and, as a consequence, proposals for state or municipal bodies of professional education in the field of professional development or retraining of workers [1].

Automated data processing and the search for relevant information are greatly simplified through a single vocabulary of terms. However, a single updated dictionary of specialties, narrowly characterizing a certain type of activity, has yet to be created. To pre-

pare a block of vacancies from the resume database most fully complying the advertisements, the methodology of multiple semantic text analysis is applicable. At the same time, it is of fundamental importance to automate the analysis procedure, preceded by studies of tests that are necessary for determining the level of preliminary manual processing of files.

LITERATURE REVIEW

In the scientific literature, a significant place is given to publications devoted to topic modeling, by which it is possible to determine characteristic topic blocks in large texts [2]. Topic modeling is widely used in information searches for the purpose of detecting latent non-classified information, in some cases fundamentally new information. The technology of topic modeling makes it possible to reduce a text consisting of millions of terms to several hundred topics based on the probability with which different words are generated in some topic.

The most common methods of topic modeling are the latent Dirichlet allocation (LDA) and probabilistic latent semantic indexing (pLSI). The foundations of the LDA were laid by D. Blei, A. Ng, and M. Jordan in 2003 [3]. From the point of view of the latent placement of Dirichlet, service words have the same probability in different topics, and any topic has the probability of generating specific words. With an appropriate rubricator that attributes specific words to a certain topic it is a fairly simple task to determine a set of topics in any text. This method of textual analysis is used in various scientometric studies [4].

The above techniques are extremely useful for determining the qualitative similarity between two documents and highlighting the key terms that characteristically describe a specific text. Using key terms, by analyzing their pairwise distribution in each of the text segments, it is possible to undertake a cluster analysis, whose result will be the construction of a hierarchical picture of terms. Within the framework of library and information sciences, metric studies have become extremely popular in recent years and include analysis of the joint use of words [5–7].

Semantic systems based on the use of ready-made dictionaries are being increasingly applied in the field of higher education management. Thus, the staff of the information systems school of the Singapore University of Management and experts in the field of teaching computer science carried out a joint study on the automatic processing of curricula in computer science based on the analysis of the developed vocabulary of professional competences [8]. From a certain point on, the problem of the completeness and integrity of university education began to be considered from the point of view of mastering the competencies that are necessary to start a career. The competence approach and use of the term "learning outcomes" in the design of a training course bring the necessary clarity and transparency to training courses that are being developed. In order to identify the necessary competencies, the role of systems for the semantic processing of large texts increases. The essence of curriculum analysis is to evaluate the components of the curriculum and develop proposals for its improvement.

A single study in selected universities in the United States and Singapore in 2008 specifically developed a list of skills in the field of information systems for business [9], which was formed by providing expert assessments and taking the views of employers and the personal experience of IT professionals into account. Since 2014, for European universities there is a single qualification matrix for the computer-science specialty, while in other countries the general classification is only still being formed manually by removing repetitions and combining related skills. There are other studies devoted to the analysis of curricula based on the joint use of terms-competencies [10].

From the point of view of the development of national human capital, analytical processing of rapidly changing data on national employment sites is of great importance. One particular example is [11], which was devoted to the analysis of positive job experiences on Taiwanese job sites via task-technology fit (TTF) using specialized software (UTAUT2). The TTF model spread after the publication by D. Goodhue and R. Thompson in 1995, which justified the advantage of technologies that were created specifically for a specific task [12]. Popular science videos exist that explain the power and significance of the new type of technology (https://youtu.be/R9UGr5SpzlQ).

A similar technology focused on the solution of a specific task, that is, the search for actual skills of growing professions, is necessary for any regional employment sphere.

A METHODOLOGY FOR THE FORMATION OF DATABASES BASED ON SEMANTIC ANALYSIS

Since the requirements of an employer for job candidates and the characteristics of real employees and their resumes are unstructured text data, the content of the black box can be processed using various types of semantic algorithms. These include text indexing, their comparison, identification of meaning, statistical analysis of texts, and searching in texts. In the first approximation, the task of finding the resume that is most suitable from the employer's point of view is to compare two unstructured texts.

The task of comparing the two texts in computer science belongs to classical ones. A sufficient number of semantic textual search methods have been developed based on the analysis of the frequency of occurrence of

words, their correspondence to a certain topic, and their joint use and probabilistic distribution [13].

If the first word of the first text is selected and compared with all the words of the second text when it is found it can be stated that this word occurs in both the first and second text, while if it is not found, then this word is only in the first text. After the completion of the procedure, the second text contains the words found only in it. On average, the complexity of this procedure is the product of the lengths of texts in words by the average length of the text.

For the optimal solution of this problem, the apparatus of non-biased (ambiguous) mappings with the following properties is used.

Let W be a word of an arbitrary length L in some alphabet A. Consider the transformation $H(W) = h$, which maps words of arbitrary length to a word of fixed length.

This transformation must have the following property: for a random equiprobable choice of two words W_1 and W_2 in the alphabet A from the set of possible words, their corresponding words h_1 and h_2 must be different with high probability.

If the transformation H is a Shannon stirring transformation, then, as a rule, the length of the word h is used to estimate the probability.

Suppose that the length of the hash word is 3 bytes. Then, the conditional probability $P(h_1 = h_2/W_1$ is not equal to $W_2)$ is estimated to be of the order of 2^{-24}, that is, 10^{-7} (taking the fact into account that $2^{10} = 1024$ is approximately $1000 = 10^3$).

Therefore, the following transformation can be applied to any text in any language: every single word of the text W that is longer than the length h can be replaced by the value (hash value) of its function H (hash function). For generality, the hash value can replace all words regardless of their length.

As a result, the text is converted into a sequence of binary numbers, let us call them hash words, each of which is of the length $|h|$, that is, the length of the hash value (in the example given, 3 bytes). The principal result of this transformation is that any constructions for comparison and searching become equal in length and there is no need to compare words of different lengths.

Further, for each text T_i, simultaneously with its transformation to hash words, a dictionary D_i is constructed consisting of non-repeated hash values and the corresponding words.

Dictionary D_i is a metric of the content of the text and makes it possible not only to optimize the search in the text (we are primarily looking for the words of the search query in the dictionary; if available, we search for them in the text, as modern search engines do), but also compare texts containing unstructured data between each other.

The complex of indexing and text analysis used by Prof. A.Yu. Shcherbakova operates on the basis of the described algorithm [14]. The text-indexing program m_ind, when launched in the format m_ind[.exe] filename.ext, creates three files:

filename.csv: a list of words (in the Windows encoding) that occur in the indexed text (dictionary). The file consists of 35-byte records, of which 32 bytes are space-separated words, the ";" character, and two newline characters;

filename.lmd: index file;

filename.num: a file of double-byte values; the *i*th field is equal to the number of words with the *i*th entry number in the dictionary, found in the indexed text.

The text-comparison program tcmp when launched in the format Tcmp [.exe] filename1.ext1 filename2.ext2 produces a set-theoretic comparison of the two texts specified in the arguments (the files filename1.ext1 and filename2.ext2 must be previously indexed by the m_ind program) and creates three files:

- onlyone.csv: words that occur only in the first text (filename1),
- onlytwo.csv: words that occur only in the second text (filename2),
- common.csv: words that occur in both filename1 and filename2.

The program estimates the similarity metric of texts in files filename1.ext1 and filename2.ext2. For this, three metrics are defined [12].

Let T1 be the number of words in the first text (filename1.ext1), T2 be the number of words in the second text (filename2.ext2), O1 be the number of words in the file onlyone.csv, O2 be the number of words in the onlytwo.csv file, and C be the number of words in the common.csv file.

Then, C is the cardinality (the number of elements) of the intersection of the two sets filename1.ext1 and filename2.ext2. The similarity metric of texts should be equal to:

(1) 0, if C = 0, that is, if the intersection of the sets of two texts is empty;

(2) 1, if O1 = O2 = 0 and the texts are the same.

In this case, the equality T1 + T2 = O1 + O2 + 2C holds, which follows from the equalities T1 = O1 + C and T2 = O2 + C.

Then, we can define metric R1 = 1/2(C/T1 + C/T2). As experiments show, this metric will take maximum values when assessing the similarity of texts.

In addition, we can define a "natural" metric R2 = 2C/(T1 + T2), which, as is easily seen, takes the value 0 if the texts do not match (C = 0), and 1 if C = T1 = T2 in case of matching texts.

Finally, we can also define a third and minimal metric of the similarity of texts R3 = C/(O1 + O2 + C).

Metric R3 is necessary for cases when some words are deleted during comparison of texts, for example,

words of short lengths (2 symbols: prepositions, conjunctions, and interjections) or other words at the choice of the analyst.

Let us illustrate this with concrete examples. For this we take one typical requirement for a vacancy and four resumes of applicants from the site hh.ru. For the analysis using the tcmp program, we have five files: base_t, which in an unstructured arbitrary form contains the requirements of the employer to the vacancy engineer–technologist in the field of oil production and refining; the resume engineer–technologist in the field of oil production and refining of the first applicant, file rt1; the resume engineer–technologist in the field of oil production and refining of the second applicant, rt2 file; as well as, for illustration, the technology files of two resumes: sewing manufacturing engineer–technologist, files rt3 and rt4. Here are the results of a pairwise comparison of all the resumes with the file base_t.

For the first applicant:

Min word length in COMMON => 0

Read pages.....

Successful comparison! See onlyone,onlytwo and COMMON files

Files:

[rt1.txt]=398 words [base_t.txt]=171 words All=569

[onlyone]=347 [onlytwo]=120 [common]=51 All=569

Files metrics is correct

first Equal metric = 0.213193 [21%] –>Hihg

Null-Equal metric = 0.179262 [17%] –>Medium

second Equal metric = 0.098456 [9%] –>Down

Medium = 0.163473 [16%]

For the second applicant:

Min word length in COMMON => 0

Read pages.....

Successful comparison! See onlyone,onlytwo and COMMON filesFiles:[rt2.txt]=181 words [base_t.txt]=171 words All=352[onlyone]=158 [onlytwo]=148 [common]=23 All=352

Files metrics is correct

first Equal metric = 0.130787 [13%] –>Hihg

Null-Equal metric = 0.130682 [13%] –>Medium

second Equal metric = 0.069909 [6%] –>Down

Medium = 0.110349 [11%]

For the third applicant:

Min word length in COMMON => 0

Read pages.....

Successful comparison! See onlyone,onlytwo and COMMON files

Files:

[rt3.txt]=277 words [base_t.txt]=171 words All=448

[onlyone]=252 [onlytwo]=146 [common]=25 All=448

Files metrics is correct

first Equal metric = 0.118226 [11%] –>Hihg

Null-Equal metric = 0.111607 [11%] –>Medium

second Equal metric = 0.059102 [5%] –>Down

Medium = 0.096215 [9%]

For the fourth applicant:

Min word length in COMMON => 0

Read pages.....

Successful comparison! See onlyone,onlytwo and COMMON filesFiles:[rt4.txt]=315 words [base_t.txt]=171 words All=486[onlyone]=275 [onlytwo]=131 [common]=40 All=486

Files metrics is correct

first Equal metric = 0.180451 [18%] –>Hihg

Null-Equal metric = 0.164609 [16%] –>Medium

second Equal metric = 0.089686 [8%] –>Down

Medium = 0.144771 [14%]

In these results of the program, the metric R1 occurs for the text first Equal metric, metric R2 occurs for Null-Equal metric, and metric R2 occurs for second Equal metric.

As is easy to see, the resume of the first applicant as much as possible corresponds to the requirements of the employer; the first metric takes the value of 21% (first Equal metric = 0.213193 (21%)). For the other applicants, the first metric takes the values 13, 11, and 18%, respectively.

It is not difficult to explain the fact that the summary of the second applicant is much less consistent with the requirements of the vacancy: his experience in the field of oil production and refining is less than 3 years, while the first applicant's experience is more than 18 years; it is much smaller in volume and contains fewer skills in the specialty.

An unexpectedly high metric of similarity was revealed in the comparison of the fourth resume sewing manufacturing engineer–technologist of with the basic vacancy (the length of the text of the resume is comparable to the length of the text of the first resume), which indicates the presence of a large number of words common to most engineer–technologist specialties in different fields of activities, such as production, technological, products, development, preparation, compilation, etc.

To improve the proposed methods, we perform the following. Consider, for example, the specialty engineer–designer of radio electronic equipment. We combine all the requirements for the vacancy engineer–designer of radio electronic equipment in one text file and apply the m_ind transformation. As a result, we obtain a vocabulary of the vacancy that contains over 500 positions: basic words that describe the requirements for the applicant that occur at least once in the combined text. Next, we carry out manual

Table 1. Comparative analysis of vacancies and a group of resumes using the complex of text indexing and analysis (KIAT), truncated at up to 231 words

Vacancy	Resume	Average value from 3 metrics
Engineer–designer of radio electronic equipment	Engineer–designer of radio electronic equipment	0.101325 [10%]
Engineer–designer of radio electronic equipment	Sewing-manufacturing Engineer	0.074569 [7%]
Engineer–designer of radio electronic equipment	Oil industry Engineer–technologist	0.032651 [3%]

expert processing of the text: we exclude prepositions and conjunctions, auxiliary words, and words that are of little importance for the specialty. If the first file is truncated to 231 words, we compare it with the resume of the radio electronic equipment engineer, sewing manufacturing engineer, and oil industry engineer–technologist (Table 1).

For the first applicant:

Min word length in COMMON => 0

Read pages.....

Successfull comparison! See onlyone,onlytwo and COMMON files

Files:

[treb2.txt]=231 words [rezrea.txt]=235 words All=466

[onlyone]=203 [onlytwo]=207 [common]=28 All=466

Files metrics is correct

first Equal metric = 0.120181 [12%] –>Hihg

Null-Equal metric = 0.120172 [12%] –>Medium

second Equal metric = 0.063927 [6%] –>Down

Medium = 0.101325 [10%]

For the second applicant:

Min word length in COMMON => 0

Read pages.....

Successfull comparison! See onlyone,onlytwo and COMMON files

Files:

[treb2.txt]=231 words [rt4.txt]=315 words All=546

[onlyone]=207 [onlytwo]=291 [common]=24 All=546

Files metrics is correct

first Equal metric = 0.090043 [9%] –>Hihg

Null-Equal metric = 0.087912 [8%] –>Medium

second Equal metric = 0.045977 [4%] –>Down

Medium = 0.074569 [7%]

For the third applicant:

Min word length in COMMON => 0

Read pages.....

Successfull comparison! See onlyone,onlytwo and COMMON files

Files:

[treb2.txt]=231 words [rt2.txt]=181 words All=412

[onlyone]=223 [onlytwo]=173 [common]=8 All=412

Files metrics is correct

first Equal metric = 0.039415 [3%] –>Hihg

Null-Equal metric = 0.038835 [3%] –>Medium

second Equal metric = 0.019802 [1%] –>Down

Medium = 0.032651 [3%]

Thus, when the requirements file is truncated to the keywords in the specialty, the result of the comparison becomes very convincing; when the dictionary of the first file (vacancy) is truncated to the keywords–terms that characterize the employee's competencies, the methodology that we propose may be applicable for selecting job announcements with resumes that are suitable for the respective vacancies.

THE SEMANTIC KERNEL AS A BASIS FOR FORMULATING A LIST OF KEY SKILLS

Any labor market can be considered as a dynamically developing structure; this has a significant influence on the choice of a relevant database. Among researchers, there has been a noticeable tendency to search for updated data in Internet resources and social networks. Sources of information that are replenished in real time have themselves been the subject of a special study for many years. To work with large Internet data, rules and special techniques are created. D. Lewandowski in 2012 introduced the basic rule of information retrieval when working with large arrays of open data, which, in particular, consisted in creating stable data files [15]. For Lewandowski, this rule made sense in the context of analyzing user queries when searching for information in networks or library directories connected to open sources of information. In fact, it extends to any types of information retrieval related to large data in the WWW.

For a review of the possibilities of semantic analysis of vacancies in the allocation of key skills for a particular specialty, we will consider job announcements in the field of engineer–designer. If it is not a question of soft skills (for example, the ability to work with documents) but rather of narrow professional competencies, then an expert can quite easily identify them in an announcement (as a rule, these competences are listed under the heading requirements). When working with

Table 2. Ranking of terms by the frequency of references to the specialty (engineer–designer)

Term	Frequency of reference
AutoCad	14
SolidWorks	9
Kompas-3D	5
Scad	3
Lira	3
Revit	2
Inventor	2

Table 3. Joint use of terms taking their "proximity" into account

Pair no.	Term 1	Term 2	Joint use (proximity)
1	AutoCad	SolidWorks	8 (0.3)
2	AutoCad	Revit	1 (0.8)
3	AutoCad	Lira	3 (0.6)
4	AutoCad	Kompas-3D	4 (0.5)
5	Kompas-3D	Inventor	2 **(0.4)**
6	SolidWorks	Lira	1 (0.8)
7	Revit	Lira	1 (0.6)
8	AutoCad	Inventor	2 (0.7)
9	AutoCad	Scad	2 (0.7)
10	SolidWorks	Kompas-3D	4 (0.4)
11	Scad	Revit	1 (0.6)
12	Scad	Lira	1 (0.6)

job announcements, seven terms that were used in them more than once were identified. For this specialty, the level of professionalism is determined by the set of specialized software products the applicant has proficiency in. The terms with spelling variants were reduced to uniformity for the convenience of further work (Table 2).

When analyzing key terms in any text, not only is the frequency of the use of terms of great importance but also the search for satellite terms that themselves can rarely be used but always follow key terms with a large number of uses. To distinguish these satellites, we will make up a general scheme for joint use of terms (one line corresponds to the list of terms mentioned in one announcement):

AutoCad, SolidWorks, Revit, Lira
AutoCad, SolidWorks
SolidWorks
AutoCad, Kompas-3D, Inventor, SolidWorks
AutoCad, Kompas-3D, SolidWorks
AutoCad, Scad
AutoCad
SolidWorks, Kompas-3D, AutoCad.
SolidWorks, AutoCad
AutoCad
Scad, Revit
SolidWorks, Kompas-3D
AutoCad, Lira
AutoCad, Inventor, Kompas-3D
AutoCad, SolidWorks
AutoCad, Scad, Lira
SolidWorks, AutoCad

For each pair of terms, we calculate the proximity of their mutual arrangement according to a formula that takes the number of joint use of two terms into account, as well as reference to each of the terms minus the number of joint uses. The proximity between terms 1 and 2 is determined by the formula[1]:

$$(A + B)/(2C + A + B),$$

where: C is the joint use of term 1 and 2; A is reference to term 1 minus C; and B is reference to term 2 minus C.

The result of this step is a digital indicator of the proximity between the two terms (Table 3).

The thresholds for the significant proximity of terms can be determined depending on the difference in proximity and study objectives. In our case, we define the proximity threshold from 0 to 0.5 and analyze the terms close to the three most commonly used terms: AutoCad, SolidWorks, and Compass-3D (see Table 2). Analysis of the closeness of terms draws attention to the term Inventor. If the threshold significance was determined in the range from 0 to 0.6, the term Lira would also be significant.

Thus, by identifying the semantic kernel of job advertisements and further statistical analysis of terms it is possible to create a structure of key skills for a particular specialty, which is of great utility for further labor market analytics.

CONCLUSIONS

The described technique for selecting the semantic kernel of a textual array makes it possible to compile a stable database of vacancies and a corresponding resume database. Automatic matching of texts will be effective if it is performed on the basis of an expertly determined basic semantic structure of the test being analyzed.

ACKNOWLEDGMENTS

This publication was prepared as part of a grant supported by the Russian Foundation for Basic Research, grant no. 16-33-01023 (introduction, literature review, use of a complex of text indexing and analysis in the research of employment sites, study of the role of the semantic kernel for identification of key

[1] In this case, the metric of Lance and Williams is used, which takes a value from 0 and 1 and considers only those observations for which at least one trait is present.

skills), and within the framework of works supported by the Russian Foundation for Basic Research, grant no. 15007-08522 (creation of a complex of text indexing and analysis, description of the program operation principle).

REFERENCES

1. Anisimova, A.E. and Gagel'strom, A.O., Russian higher education and labor market, *Ross.: Tendentsii Perspekt. Razvit.,* 2016, no. 3, pp. 717–726.
2. Kun Lu, Xin Cai, Ajiferuke, I., and Wolfram, D., Vocabulary size and its effect on topic representation, *Inf. Process. Manage.,* 2017, vol. 53, pp. 653–665. http://dx.doi.org/ doi 10.1016/j.ipm.2017.01.003
3. Blei, D., Ng, A.Y., and Jordan, M.J., Latent Dirichlet allocation, *J. Mach. Learn. Res.,* press, 2003, vol. 3, pp. 993–1022.
4. Hassan, S.U. and Haddawy, P., Analyzing knowledge flows of scientific literature through semantic links: A case study in the field of energy, *Scientometrics,* 2015, vol. 103, no. 1, pp. 33–46.
5. Hu, J. and Zhang, Y., Research patterns and trends of recommendation system in China using co-word analysis, *Inf. Process. Manage.,* 2015, vol. 51, no. 4, pp. 329–339. http://dx.doi.org/ doi 10.1016/j.ipm.2015.02.002
6. Khasseh, A.A., Soheili, F., and Moghaddam, H.S., Intellectual structure of knowledge in iMetrics: A co-word analysis, *Inf. Process. Manage.,* 2017, vol. 53, pp. 705–720. http://dx.doi.org/10.1016/j.ipm.2017.02.001
7. Ronda-Pupo, G.A. and Guerras-Martin, L.A., Dynamics of the evolution of the strategy concept 1962–2008: A co-word analysis, *Strategic Manage. J.,* 2012, vol. 33, no. 2, pp. 162–188. doi 10.1002/smj.948
8. Gottipati, S. and Shankararaman, V., Competency analytics tool: Analyzing curriculum using course competencies, in *Education and Information Technologies,* Springer, 2017. doi 10.1007/s10639-017-9584-3
9. Ducrot, J., Miller, S., and Goodman, P.S., Learning outcomes for a business information systems undergraduate program, *Commun. Assoc. Inf. Syst.,* 2008, vol. 23, art. 6. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1920&context=tepper.
10. GnanaSingh, A.A. and Leaveline, E.J., Competency-based calisthenics of learning outcomes for engineering education, *Int. J. Educ. Learn.,* 2013, vol. 2, no. 1, pp. 25–34.
11. Kuo-Yu, Huang. and Yea-Ru, Chuang., Aggregated model of ttf with utaut2 in an employment website context, *J. Data Sci.,* 2017, vol. 15, pp. 187–204.
12. Googhue, D.L. and Thompson, R.L., Task-technology fit and individual performances, *MIS Q.,* 1995, vol. 19, no. 2, pp. 213–236.
13. Drosou, M., Jagadish, H.V., Pitoura, E., and Stoyanovich, J., Diversity in big data: A review, *Big Data,* 2017, vol. 5, no. 2, pp. 73–84. doi 10.1089/big.2016.0054
14. Ryazanova, A.A. and Shcherbakov, A.Yu., To the problem of metrics of similarity of tests for methods of their automated comparison, *Tekhnicheskie nauki: Nauchnye prioritety uchenykh. Sb. nauchn. tr. po itogam mezhdunarodnoi nauchno-prakticheskoi konferentsii* (Technical Sciences: Scientific Priorities of Scientists. Proc. Int. Sci.-Pract. Conf.), Tolyatti, 2017.
15. Lewandowski, D., A framework for evaluating the retrieval effectiveness of search engines, in *Next Generation Search Engines,* Louis, C., Biskri, I., Ganascia, J.-G., and Roux, M., Eds., Hershey, PA: IGI global, 2012, pp. 456–479. doi 10.4018/978-1-4666-0330-1.ch020

Translated by K. Lazarev