Digital technologies for searching and processing unstructured information in modern higher education

Mikhnev I.P.
Volgograd Institute of Management, branch of the Russian Academy of National Economy and Public Administration,
Department of Information Systems and Mathematical Modeling
Volgograd, Russia
mkmco@list.ru

Salnikova N.A.
Volgograd Institute of Management, branch of the Russian Academy of National Economy and Public Administration,
Department of Information Systems and Mathematical Modeling
Volgograd, Russia
ns3112@mail.ru

Mikhneva S.V.
Volgograd Institute of Management, branch of the Russian Academy of National Economy and Public Administration,
Department of Theory and History of Law and the State,
Volgograd, Russia
svet-mihneva@list.ru

***Abstract* — The article presents the results of research on the application of Big Data digital technologies for unstructured information in modern higher education. Considered the importance of technology Big Data at the present stage of development, analyzed the prospects for application in higher education. An overview of the most common mathematical methods for extracting content from arrays of unstructured information is given: prediction method, TF-IDF method, hidden Markov model, PageRank reference ranking model, document indexing model, formalization method of processing unstructured Web-documents data. The present state of the art, trends in the development of Big Data and Web Mining technologies, and examples of companies that have implemented Big Data technologies in their activities are outlined.**

Keywords* — *analysis of unstructured data; Big Data technologies; digital computer learning; digital technology; higher education; unstructured information.

I. INTRODUCTION

The Big Data technologies are able to transform modern higher education into a product with an excellent predictable result, and a university teacher – into a super-professor [1]. Big Data analysis allows you to speed up the solution of various research, scientific and pedagogical shortcomings. When studying statistics, it is possible to work with individual vectors and with educational systems of the global level [2]. The Big Data technologies help to significantly improve the so-called “teacher design” in those places where a student “falls asleep from boredom”; it is better to replace such an education system. This technology automates the behavior of the educational system and gives actual recommendations and tips if students are stuck in one place. And if these recommendations and tips do not help, technology notifies teachers and parents that such students need additional help [3].The Big Data technology captures various data from a student and analyzes how a particular student learns. It fixes where the student is mistaken, what he decides slowly, what he decides quickly, when he is distracted and makes a full detailed portrait of the student: how much time and for what actions he spent, correctly or not, how much he moved with the mouse on the screen, how many times he returned to the decision of one and the same tasks. The Big Data helps to process the experience of hundreds of thousands of teachers and students, and on the basis of analysis to obtain an effective educational methodology [4]. Such an educational technique becomes a product of mass experience. With the help of the Big Data technology, you can personalize content to the needs of each student [5]. For example, the Big Data analyzes hundreds of thousands of pages of textual information in the global network and selects the text that contains the necessary number of new words and phrases needed to learn [6]. One of the features of the functioning of educational systems is the need to process large amounts of unstructured data, filter them and adequately interpret, which is a priority task for the modern system of higher education [7]. To fulfill the functions of processing unstructured data within the framework of the implementation of the concept of industrial Internet of things in a distributed automated system, the Web Mining technology is implemented, using the Data Mining methods to research and extract information from Web documents and services [8].

Volumes of heterogeneous and rapidly arriving digital information cannot be processed with traditional instruments. The analysis of the data itself allows you to see certain and imperceptible patterns that a person cannot see [9]. This allows you to optimize all areas of our life – from government to production and education. Usually big data comes from three sources:

- Internet (social networks, forums, blogs, media and other sites);
- corporate archives of documents;

- sensors, devices and other devices.

II. Big data in education

The education system has always generated a significant amount of data. The only question was how to start working with this data at the system level: analyze them, make decisions based on them. In the field of education, there are five main types of data:

- personal information;
- data on the interaction of students with electronic learning systems (electronic textbooks, online courses);
- data on the effectiveness of educational materials;
- administrative (system-wide) data;
- forecast data.

Let us show by examples how different types of data are used in educational institutions, and how this helps education.

A. *Big Data and Economics*

According to statistics, 400,000 students are deducted annually in the USA. Many students take loans to study, and deduction for them is not only a big risk of non-payment of debt, but also a deterioration of the entire credit history. The outflow of students also negatively affects the educational institutions themselves: the greater the outflow, the lower the profits and government financial support. In addition to the economic factor, the percentage transition freshmen to the next course affects the position of the university in the national rankings. To solve the problem, the University of Virginia University together with the research company Education Advisory Board conducted a study that helped identify students at risk and help them. Students who began to skip classes or receive bad grades left the school more often. A platform was created for the university that aggregates all student grades and finds problems. Employees can work with them individually – for example, to offer the student a tutor or other assistance. During one semester, the number of students who completed the course increased by 16%, and the number of students who took the next course of study increased by 8 percent.

B. *Learning Personalization*

One of the popular strategies for personalizing learning is to offer an additional online course to a lagging student. As the student answers questions, the platform will be able to predict his readiness for new topics. For example, the Arizona University of Technology needed to develop a new course in mathematics, as students had to prepare for the exam for a whole year. After using additional courses based on the Knewton platform, about half of the students were able to pass the exam at least a month earlier. Another area of application for big data is predictive modeling. American colleges and universities themselves send letters of appeal to prospective students, inviting them to enter this or that educational institution. Each university seeks to invite the most promising students who will probably enter. To facilitate the work of the admissions office, analysts from ForecastPlus collected and analyzed several types of student data: ethnicity, performance on a number of subjects, final works, and grades. ForecastPlus predictive modeling has proven its effectiveness in more than one hundred US campuses. Thus, Creighton University in Nebraska was able to exclude 35,000 students who were not the most promising and did not send letters to them, which allowed them to save more than 30 thousand dollars.

C. *Improving the Quality of Teaching*

More and more educational institutions are beginning to use technology, producing a huge flow of data. In Roosevelt Elementary School near San Francisco, teachers use the DIBELS program with reading assignments to help identify students who are lagging behind and offer them help. This allows the teacher to quickly prepare and adapt their lessons to the needs of schoolchildren. Assessing the quality of teaching with the help of tests cannot be truly effective: as a result, teachers simply train students on certain types of tasks. By analyzing the data on the educational process, the school administration can better assess the teachers and, if necessary, make changes.

D. *Choosing a Future Profession*

In six technology colleges in South Carolina, there is a program for obtaining a new profession, SC ACCELERATE, aimed at people over 25 years old. Data analysis allows participants to choose an education and career that best fits their experience and personal qualities. The CareerChoice GPS program carries out a prognostic analysis and helps to determine the choice of a career: the service examines the student’s personality traits, his learning success, and previous work experience. Applicants submit applications to the most suitable universities - and the latter benefit from it. It is also beneficial for employers: they get specialists who are already prepared for work.

E. *Virtual Campus*

In Russia, the use of big data in education is just beginning. Unfortunately, there are quite a few implemented initiatives. The authors propose to develop a student identification card that combines a number of functions: a travel and student ID card, a passbook and a pass to the territory. Using the map, it will be possible to collect data on the time and location of the student on campus. By creating a personal account, students will monitor progress, communicate with teachers, learn class schedules, and study the university catalog. All of these services will allow collecting and processing data in order to generate further recommendations for improving the learning experience [10].

III. Overview of the most common mathematical methods for extracting content from arrays of unstructured information

The methods used in Big Data analysis come from pattern recognition, computer learning, statistics and psychometrics. The DataShops special repositories that exist today allow you to collect data and analyze it right there. Today, the most popular repository, PSLC DataShop, collects and stores information processed in more than 260,000 hours spent by students in educational programs, this is approximately 55 million actions, responses and various results [11].

Big Data combines technology and state-of-the-art technologies that find meaning from a variety of data at an ultra-extreme practicality limit. Big Data requires special approaches, tools and methods that are significantly different from traditional classical ones. New technologies of the future based on Big Data – this is a whole series of approaches, tools and methods for processing a significant variety of unstructured data of huge volumes to obtain user-perceived new results. Also effective in conditions of continuous growth, dislocation across various nodes of the global network, alternative to traditional DBMS and Business Intelligence class solutions. In this series of approaches include the latest means of mass simultaneous processing of unstructured big data, first of all, the capabilities of such categories as NoSQL, MapReduce algorithms, special program circuits and Hadoop libraries.

A. Forecasting Method

One of the most interesting models of working with Big Data is forecasting, where a combination of known data allows you to provide a forecast of the unknown unknown. Multiple data is collected from records of Internet services, student systems, surveys, social networks and various observations during experiments. The collection and processing of such data is a huge matter, since you need to know which moments to look at and be able to identify the necessary useful information. The model can work to predict the present, using statistics for the past hour, finding out if a student is interested in watching an online course, or predicting the future (using previous estimates), whether the student can solve the following problem and with what result. Modern algorithms of Big Data technology take into account the cost of the error and the efficiency of the correct use of the educational system. For example, if in one minute a student learns 0.05% of the course, then the wrong prediction "costs" him one extra minute of training, and the correct one adds 0.03% [12].

B. TF-IDF method

TF-IDF (TF – term frequency, IDF – inverse document frequency) A statistical measure used to assess the importance of a word in the context of a document that is part of a document collection or corpus. The weight of a word is proportional to the frequency of use of the word in the document and inversely proportional to the frequency of the word in all documents in the collection.

The TF-IDF measure is often used in problems of text analysis and information retrieval, for example, as one of the criteria for the relevance of a document to a search query, when calculating the measure of document proximity for clustering.

The TF-IDF method involves the use of two concepts (metrics):

1. The frequency of a term TF (term frequency), defined as the ratio of the number of occurrences of a certain term to the total number of words in a document. This metric allows us to estimate the weight of the term t_i within a single document:

$$tf(t,d) = \frac{n_i}{\sum_k n_k}, \tag{1}$$

where n_i is the number of occurrences of the i-th term in the document, $\sum_k n_k$ is the total number of words in this document, including common and connecting words [13].

2. "Inverted document frequency" IDF (inverse document frequency) Inversion of the frequency of occurrence of a given term is found in array documents, allowing to reduce the weights of connecting and commonly used words. For each unique word within a specific collection of documents, there is only one IDF value:

$$idf(t,D) = log\,\frac{|D|}{|\{t\epsilon d_i\,|d_i\,\epsilon D\}|}, \tag{2}$$

where |D| – the number of documents in the array, $|\{t\epsilon d_i|d_i\epsilon D\}|$ – the number of documents in the array D, in which t occurs.

Thus, the measure TF-IDF is determined by the following formula:

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,d). \tag{3}$$

When using this measure, terms with a high frequency within one document and with a low frequency of use in other documents receive maximum weight.

Various modifications of the TF-IDF model are also known. One example is the Okapi BM25 measure [14]. The most significant limitation of the use of TF-IDF is the need for the data set to remain unchanged during the entire calculation time, which greatly complicates the calculation if it is necessary to conduct it in real time.

The TF-IDF method is often used to represent a collection of documents in the form of numerical vectors that reflect the importance of using each word from a certain set of words (the number of set words determines the dimension of the vector) in each document.

C. Hidden Markov Model

Of greatest interest are hidden Markov models (HMM). The hidden Markov model is a statistical model that simulates the operation of a process similar to a Markov process with unknown parameters, and the task is to guess unknown parameters based on the observed ones. The obtained parameters can be used in further analysis, for example, for pattern recognition. The HMM can be viewed as the simplest Bayesian trust network.

Consider the formal description of the hidden Markov model. Each model is determined by the following parameters:

1. The set $S = \{s_1, s_2, \ldots, s_N\}$ of N states.

2. The initial probability distribution $P = \{p_i\}$.

3. The probability matrix of transitions between states $A = \{a_i\}$.

4. Matrix of the probability of generating observations $B = \{b_j(O_t)\}$, where $b_j(O_t)$ is the probability of generating observations O_t at time t in the state $q_t = s_j$, $b_j(O_t)=P(O_t|q_t = s_j)$.

The model presented in this way is one-dimensional, but for searching and processing documented information, pseudo-multidimensional HMMs are of interest, consisting of a finite number of elements, called superpositions, each of which in turn is a separate HMM [15].

In the framework of the functioning of an automated search engine using the HMM, the document may be presented in the form of a random n-dimensional discrete signal with n signs forming an index. In turn, the index (vector document) can be extracted from the document in various ways, such as the Karhunen – Loeve transform [16].

The resulting vector documents are distributed according to the model states, which determine some key search features. For example, a HMM implementing a search for journal articles may consist of six superstates corresponding to the parameters of the article (Name, author (s), journal name, year, number, ISBN), each of which is divided into separate states.

In this section, the search for documents is based on the identification of similarities in multidimensional space. A measure of similarity is a function that determines the value of similarity between two or more objects based on some predefined criteria or metrics. A metric is a function of distance r, defined on a metric set, for any points a, b, c of which the following system of conditions is true:

$$\begin{cases} r(a,b) = r(b,a) \\ r(a,b) = 0 \leftrightarrow a = b \\ r(a,b) + r(b,c) \geq r(a,c) \end{cases} . \tag{4}$$

Consider the process of finding information as follows. Let O be the array object, L the query object, K the key documents, r the metrics. On the basis of (4), one can assert the truth of the following inequalities:

$$\begin{cases} r(O,K) \leq r(O,L) + r(L,K) \\ r(L,K) \leq r(O,L) + r(L,K) \end{cases} . \tag{5}$$

Accordingly, the distance from the query object to the array object is r(O,L) ≥ r(O,K) − r(L,K).

Thus, by successive comparison of the objects of the array and the request with the key object, the lower limit of similarity metrics between the document and the request can be selected, allowing to cut off that part of the array of documents that does not satisfy the request.

D. *Reference Ranking PageRank Model*

The first representation of a mathematical model for the formalized representation of the task of searching and processing information within the framework of the concept of the industrial Internet of things is a mathematical model of reference ranking [17].

PageRank is one of the reference ranking algorithms. The algorithm applies to a collection of documents linked by hyperlinks (such as web pages from the world wide web), and assigns each of them a certain numerical value, which measures its “importance” or “credibility” among other documents. Generally speaking, the algorithm can be applied not only to web pages, but also to any set of objects linked by reciprocal links, that is, to any graph.

Consider Web-documents and links to them in the form of a graph in which documents are vertices, and links – arcs of the graph. Let G(V, E) be an oriented graph, where V is the set of vertices, and E is the set of arcs. Then the model can be represented as a sparse matrix S – the adjacency matrix of the graph G, consisting of the elements:

$$S_{ij} = \begin{cases} 0, (i,j) \epsilon E \\ 1, otherwise \end{cases}, i, j \epsilon V. \tag{6}$$

The matrix can simultaneously contain links of the form (i, j), and (j, i). Let Θ – some set of topics, T(i, j, t) is the function of reference weight (j, i) in the theme t ϵ Θ (the so-called subject weight), deg(i, t) is defined as:

$$\deg(i,t) = \sum_{j=0}^{N} T(i,j,t) \times S_{ij}. \tag{7}$$

Expression (7) determines the sum of the weights of all outgoing links of the i-th vertex in the subject t, then the probability of a user clicking on the link (i, j) is determined by the following expression:

$$P = \frac{T(i,j,t)}{\deg(i,t)}. \tag{8}$$

The matrix of thematic scales for links (i, j) can be written in the following form:

$$\forall e \epsilon \Theta : S(i,j) = S_{ij} \times P(i,j,t) \tag{9}$$

Thus, thematic ranking of the i-th vertex can be represented as a system of iterative linear algebraic equations:

$$X_i^{k=1} = (1-d) + d \times \sum_{j=1}^{N} X_j^k \times S(i,j), \tag{10}$$

where d is the attenuation coefficient, k is the iteration number, X_j^k is the value of the rank of the j-th vertex during the k-th iteration k.

E. *Document Indexing Model*

The second mathematical representation of the problem of searching and processing unstructured documented information is the document indexing model.

Search index – a data structure that contains information about documents and is used in search engines. Indexing performed by a search engine is the process of collecting, sorting and storing data in order to provide fast and accurate information retrieval. Index creation includes interdisciplinary concepts from linguistics, cognitive psychology, mathematics, computer science, and physics. Web indexing is the process of indexing in the context of search engines designed to search for Web pages on the Internet [18].

The mathematical model of indexation can be formed in the form:

$$S = \{D, T, Q, R\} \tag{11}$$

where D = $\{d_i\}_n$ is the set of documents in the array, T = $\{t_j\}_m$ is the set of terms indexing the semantic load of documents, Q = $\{q_i\}_m$ is the set of user or system operator requests, R = $\{r_i\}_n$ is the set the references to documents issued as a result of the search, n, m, respectively, the number of documents in the array and the terms used.

where D = $\{d_i\}_n$ is the set of documents in the array, T = $\{t_j\}_m$ is the set of terms indexing the semantic load of documents, Q = $\{q_i\}_m$ is the set of user or system operator requests, R = $\{r_i\}_n$ is the set the references to documents issued as a result of the search, n, m, respectively, the number of documents in the array and the terms used.

If W = $\{w_{ij}\}_{n \times m}$ is a matrix that defines the relationship of terms to documents, where w_{ij} the weight of the j-th term in the i-th document, taking into account all the documents for which w_{ij} ϵ [0;1], then w_{ij} = 0, if the j-th term is not found in the i-th document; on the contrary, w_{ij} = 1, when the j -th term of 100% corresponds to the i-th document.

The work of the information retrieval system can be defined as the calculation of the response vector R = W × Q by converting the query vector Q in accordance with the matrix W.

To implement the transformation, it is necessary to determine the weighting factors of the terms. This definition can be represented as the following algorithm:

1. The determination of the weight of the j-th term for the i-th document. Denote it as:

$$f_{ij} = \frac{u_{ij}^{t}}{u_i^{w}}, \tag{12}$$

where u_{ij}^{t} is the total number of detected occurrences of the j-th term in the i-th document, u_i^{w} is the total number of words in the i-th document. Then f_{ij} is the weight coefficient of the jth term in the i-th document without taking into account the rest of the array (in the case when this coefficient is maximum for j ϵ [1, N], the j-th term is defined as reflecting the content of the i-th document.

2. The determination of the total weight of the j-th term within the entire array of documents.

By comparing the ratio of the total number of documents in the array n and n_j^d – the number of documents in which the j-th term occurs (the so-called document frequency), it is possible to determine whether a word is a significant term for this document (the smaller the n_j^d, value, the greater the weight j-th term in the document). For normalization, it is possible to perform the operation of the natural logarithm of n and n_j^d. Thus, the inverse document frequency is defined as:

$$f_i^{d} = \ln(n) - \ln\left(n_j^{d}\right) = \ln\left(\frac{n}{n_j^{d}}\right). \tag{13}$$

The total weight of the j-th term in the i-th document is calculated as:

$$w_{ij} = f_{ij} \times f_j^{d} = \left(\frac{u_{ij}^{t}}{u_i^{w}}\right) \times ln\left(\frac{n}{n_j^{d}}\right). \tag{14}$$

Thus, we can talk about the concentration of the j-th term in the i-th document while increasing its frequency in this document and reducing the number of documents containing the j-th term [19].

Based on the above method, it is possible to build a matrix of relations W, which is the basis of the database of information retrieval system indexes.

F. Formalizing the Processing of Unstructured Data from Web documents

The third view is the formalization of processing unstructured data from Web documents.

This task can be formulated as follows: there are many Web-oriented documents P = {p_1, p_2, ..., p_n}, each of which is defined by a set of attributes (variables) p_j = {x_1, x_2, ..., x_m, y}, where x_m is a block of information contained in a document defining the value of the variable y – the value of information of interest to the user, cleared of noise [20].

In turn, each variable x_m can take values from some set Z = {z_1, z_2, ...}. Thus, a finite set of relevant data, cleared of noise, separated from an unstructured array, over which further structuring actions are possible, is defined as D = {Z → y}.

IV. Conclusion

The presented formalized representations are the basis for the further development of methods and algorithms for processing unstructured information in distributed automated systems focused on work in modern higher education, implementing the concept of the industrial Internet of things.

On the basis of the developed ideas, the creation and software implementation of an integrated corporate information system module for managing digital education with distributed information arrays is proposed. In addition, there are practical prospects for applying development results in the field of the Web Mining to create software tools for extracting unstructured information from Web pages.

According to the results of preliminary studies and forecasts, it is possible to assume an increase in the efficiency of the work of program modules using the implementation of the developed formal representations by an average of 10–15% compared to the existing systems for searching and processing unstructured information.

References

[1] A. G. Kravets, A. G. Belov, and N. P. Sadovnikova, “Models and Methods of Professional Competence Level Research. Recent Patents on Computer Science,” vol. 9, No 2, pp. 150-159, 2016.

[2] A. Isaev, A. Kravets, L. Isaeva, and S. Fomenkov, “Distance Education: Educational Trajectory Control,” Proceedings of the International Conference e-Learning 2013, pp. 151-158, 2013.

[3] Mikhnev I.P., Novikova A.A., Petrosyan M.K. "Big Data and new technologies of the future for global information processing ", Scientific research and modern education, Collection of materials of the 2nd International Scientific and Practical Conference, pp. 235-239, 2018.

[4] A.G. Kravets, M.A. Kanavina, and N.A. Salnikova, “Development of an Integrated Method of Placement of Solar and Wind Energy Objects in the Lower Volga,” International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), pp.1-5, 2017.

[5] Valery A. Kamaev, Ilya P. Mikhnev, and Natalia A. Salnikova, "Natural Radionuclides as a Source of Background Irradiation Affecting People Inside Buildings", Procedia Engineering, vol. 150, pp. 1663-1672, 2016.

[6] Alla Kravets, Nikita Shumeiko, Boris Lempert, Natalia Salnikova, and Natalia Shcherbakova, “Smart Queue” Approach for New Technical Solutions Discovery in Patent Applications,” Communications in Computer and Information Science, vol. 754, CIT&DS 2017, Volgograd, Russia, pp. 37-47, 2017.

[7] I.P. Mikhnev, N.A. Salnikova, and M.B. Lempert, "Modern Condition of Dose Loads from Construction Materials and Main Sources of Ionizing Impact on the Population of the Volgograd Region", Materials and Technologies in Construction and Architecture, Materials Science Forum, vol. 931, pp. 1007-1012, 2018.

[8] Hieu Tran, and M. V. Shcherbakov, “Detection and Prediction of Users Attitude Based on Real-Time and Batch Sentiment Analysis of Facebook Comments,” Computational Social Networks, 5th International Conference, (CSoNet 2016), Ser. Lecture Notes in Computer Science, vol. 9795, Proceedings, Springer, Ho Chi Minh City, Vietnam, pp. 273-284, 2016.

[9] I.P. Mikhnev, N.A. Salnikova, and M.B. Lempert, "Research of Activity of Natural Radionuclides in Construction Raw Materials of the Volgograd Region", Solid State Phenomena, vol. 265, pp. 27-32, 2017.

[10] A. P. Tyukov, O. A. Khrzhanovskaya, A. A. Sokolov, M. V. Shcherbakov, and V. A. Kamaev, “Fast Access to Large Timeseries Datasets in SCADA Systems,” Research Journal of Applied Sciences, vol. 10, No. 1, pp. 12-16, 2015.

[11] Ilya P. Mikhnev, Natalia A. Salnikova, Mikhail B. Lempert, and Kirill Yu. Dmitrenko, "The Biological Effects of Natural Radionuclides from the Construction Materials on the Population of the Volgograd Region", 8th International Conference on Information, Intelligence, Systems and Applications (IISA 2017), pp. 1-6, 2017.

[12] A. P. Shiryaev, A. V. Dorofeev, A. R. Fedorov, L. G. Gagarina, and V. V. Zaycev, “LDA models for finding trends in technical knowledge domain,” Proceedings of the 2017 IEEE Russia section Young researchers in electrical and electronic engineering conference (ElConRus 2017), Saint-Petersburg, pp. 551–554, 2017.

[13] Ilya P. Mikhnev, Svetlana V. Mikhneva, and Natalia A. Salnikova, "Studies of radon activity in civil engineering and environmental objects", International Journal of Engineering and Technology, S. l, vol. 7, No. 2.23, pp. 162-166, 2018.

[14] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” Journal of Documentation, MCB University Press, vol. 60, No 5, pp. 493–502, 2004.

[15] I.P. Mikhnev, N.A. Salnikova, and S.V. Mikhneva, "Effect of Thermal Treatment of Building Materials on Natural Radionuclides Effective Specific Activity", Materials Science Forum, Vol. 945, pp. 30-35, 2019.

[16] Okapi BM25. URL: https://ru.wikipedia.org/wiki/Okapi_BM25 (Date of the application: 15.03.2019).

[17] Introduction to Hidden Markov Models. URL: https://scholar.harvard.edu/files/adegirmenci/files/hmm_adegirmenci_2014.pdf (Date of the application: 15.03.2019).

[18] Hidden Markov Model Toolkit Book, Cambridge University Engineering Department, 2001–2009. URL: http://htk.eng.cam.ac.uk/prot-docs/htk_book.shtml (Date of the application: 15.03.2019).

[19] Ilya Mikhnev, Natalia Salnikova, and Svetlana Mikhneva, "New Industrial Technologies and Innovations for the Production of Nanostructured Materials", Advances in Social Science, Education and Humanities Research, Vol. 240, pp. 83-89, 2019.

[20] A. G. Kravets, and A. Gurtjakov, “Corporate Intellectual Capital Management: Learning Environment Method,” Proceedings of the IADIS International Conference ICT, Society and Human Beings 2013, Proceedings of the IADIS International Conference e-Commerce 2013 pp. 3-10, 2013.