INFORMATION TECHNOLOGY & INTERACTIONS (SATELLITE)

04 December, 2020
Kyiv, Ukraine

ISBN 978-966-2399-61-5

TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV
(FACULTY OF INFORMATION TECHNOLOGY,
FACULTY OF COMPUTER SCIENCE AND CYBERNETICS)
NATIONAL TECHNICAL UNIVERSITY OF UKRAINE "IGOR SIKORSKY KYIV
POLYTECHNIC INSTITUTE"
VIKTOR GLUSHKOV INSTITUTE OF CYBERNETICS OF THE NAS OF UKRAINE
INSTITUTE OF INFORMATION TECHNOLOGY AND LEARNING TOOLS OF THE NAES OF
UKRAINE
INSTITUTE OF INFORMATION REGISTRATION PROBLEMS OF THE NAS OF UKRAINE
INSTITUTE OF SOFTWARE SYSTEMS OF THE NAS OF UKRAINE
THE COUNCIL OF YOUNG SCIENTISTS OF THE FACULTY OF COMPUTER SCIENCE AND
CYBERNETICS AND THE FACULTY OF INFORMATION TECHNOLOGY OF
TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV

VII INTERNATIONAL CONFERENCE

Information Technology and Interactions (Satellite)

04 December, 2020

Conference Proceedings

Kyiv
Stylos
2020

UDC 004(082)
I-60

I-60 Information Technology and Interactions (Satellite): Conference Proceedings, December 04, 2020, Kyiv, Ukraine / Taras Shevchenko National University of Kyiv and [etc]; Vitaliy Snytyuk (Editor). – Kyiv: Stylos, 2020. – 388 p.

ISBN 978-966-2399-61-5

This book includes abstracts of the 7th International Conference "Information Technology and Interactions (Satellite) – 2020". Philosophical, theoretical and applied aspects which describe the results, problems and prospects of the creation and use of intelligent computing methods and creating of information systems and technology on their basis are reviewing.

Main tracks of the conference are: Artificial Intelligence Technologies, Cyberspace Protection Technologies, Data Analytics, Digital Project Management Technologies, E-commerce, E-government and E-learning Technologies, Mathematical Foundations of Information Technology, Network and Internet Technologies.

UDC 004(082)

ISBN 978-966-2399-61-5

CONTENTS

ARTIFICIAL INTELLIGENCE TECHNOLOGIES

CYBERSPACE PROTECTION TECHNOLOGIES

DATA ANALYTICS

DIGITAL PROJECT MANAGEMENT TECHNOLOGIES

MATHEMATICAL FOUNDATIONS OF INFORMATION TECHNOLOGY

NETWORK AND INTERNET TECHNOLOGIES

ARTIFICIAL INTELLIGENCE TECHNOLOGIES

[1] **Zinnur Abduramanov**
Senior Lecturer
[2] **Zarema Seidametova**
Professor
[3] **Niiare Valiieva**
Lecturer
[1-3] *Crimean Engineering-Pedagogical University*

COLOR RECOGNITION DEEP LEARNING MODEL

Hardware or software systems of certain human activities (for example, the recreation of intelligent reasoning and actions using software systems and devices, or the imitation of a device of intellectual behavior inherent in humans). Machine Learning (ML) and Deep Learning (DL) are AI subsets. In ML we create models (algorithms) which are first trained on the basis of available data (datasets), and then allow making predictions based on the data. DL uses an artificial neural network structure imitating the neurons located in the human brain. The term "deep" is used to refer to multiple layers in an artificial neural network. The basic idea is that the network of artificial neurons, constructed from interconnected switches can learn to recognize patterns in the same way as is done by the brain and the nervous system of animals. Deep learning can be defined as neural networks with many parameters and layers. One of the largest neural networks "The Sparsely-Gated Mixture-of-Experts Layer" is described in the paper [1]. Authors of the paper [1] proposed a method to increase the capacity of the model without linearly increasing the number of calculations.

Classic machine learning (ML\DL) algorithms, which are not subset of DL algorithms, train and make predictions much faster than deep learning (DL) algorithms [2] – one or more processors are enough to train a classical model. Deep learning models require additional hardware for training as well as for large-scale deployment of software infrastructure; without this, it takes a lot of time to train the model.

There are problems for which classical machine learning algorithms build a "good enough" model. But there are problems for which classical machine learning algorithms don't work very well. For example, deep learning algorithms are used for natural language processing tasks (text translation, discourse analysis, morphological segmentation, object recognition, natural language generation, natural language understanding, mood analysis, and speech recognition). Deep learning algorithms also are used in the fields where image classification is required (image classification with localization, object detection, object segmentation, transfer of image styles, coloring, reconstruction, super-resolution and image synthesis) [3]. In addition, deep learning is used in pharmaceutical industry in the development of new drugs (to predict how molecules will interact; to search for subatomic particles, etc.).

An interdisciplinary group of experts within the framework of an independent initiative of the Human-Centered Artificial Intelligence Institute at Stanford University prepared and published the study [4] on the impact of artificial intelligence on various aspects of social development. Authors of the papers [5, 6, 7, 8] presents examples of

using the frameworks, programming language and libraries for solving artificial intelligence problems and building deep learning models. The papers [9], [10], [11] describe libraries and tools for machine and deep learning (OpenAI Gym, TensorFlow, Keras, Scikit-Learn).

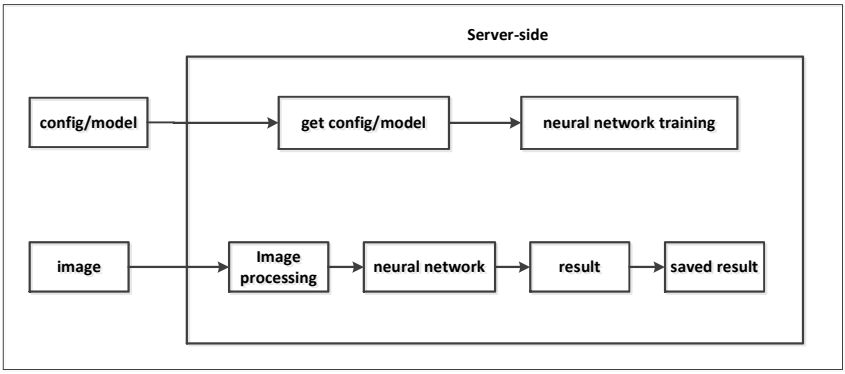

Figure 1 – Schematic DLMCR client-server application architecture

After the user has sent the image to the server, it is necessary to obtain its mathematical representation. We need to get the number of the image pixels and their representation in the three channels of R, G, B. Thus, we get two models of the input image (Fig. 2).

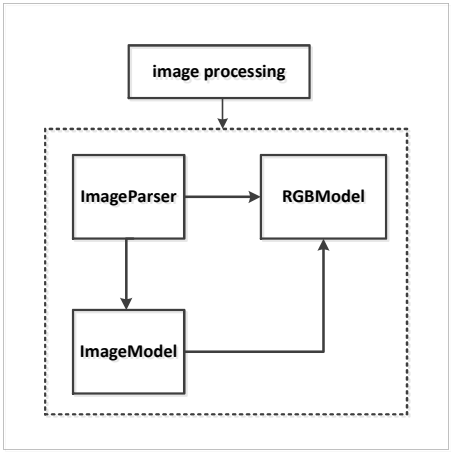

Figure 2 – Input image representation in the format required for neural network

In order to create a deep learning model, it is important to prepare a dataset that stores each color value of the RGB model and the value of the color opacity. The obtained data should be normalized, this is due to the nature of all machine and deep

learning algorithms, including neural networks. Data that is different in physical meaning and arrives at the input of a neural network can lead to its incorrect operation: incorrect predictions, non-existent data, it can also slow down the learning process and the modeling process.

After normalization, all the numerical values coming to the input of the neural network are in one narrow intermediate interval (in our case, an interval from 0 to 1).

After receiving data from the input signals, they are transmitted to the subsequent (hidden) layers of the neural network. The first such layer in the entire hierarchy determines the number of all colors in the image and assigns its own class to each. Then delving deeper into the layers of the network, each color is responsible for its own layer of the network. It means that some layer of the network can only be responsible for the defined color and it will give a result about whether the given image has this color or not. The result from each deep (hidden) layer of the network is analyzed and summarized at the output of the external, global neural network. It can give results in the form of what percentage a given color is in the image.

The number of colors neural network knows depends on the dataset on which network was trained and trained. The more colors it was given for training, the higher the accuracy of the result produced by the network.

After training, the neural network will know unique colors. The array of output values is an array of ones and zeros: [0,0,1,0,0,0,0,0,0,0,0,0], where 0 is no this color. Each color in this array has its own index. Then this dataset transforms into a json file that transferred it to train the neural network.

To implement a deep learning color recognition model, we choose the multi-paradigm JavaScript programming language that allows to cover a large number of areas for solving diverse problems: development and support of direct server services for receiving and sending data; access to the content of files, and transformation of the content into the necessary data structures; deployment and implementation of artificial intelligence systems for image analysis; transforming the output data into a human-readable format; use of the received output data for visual presentation in the client part.

Neural networks can solve almost any problem posed to a person in any area of the activity, they do it faster, more accurately and more efficiently. The goal of the project was to design and develop a deep learning model for building a color recognition system, which can be applied in the work of the designers, photographers, medical workers and in the agricultural industry.

For the analysis of existing models and architectures of neural networks, libraries that allow to implement neural networks we use Synaptic.js as the main library for implementing a neural network.

References:

1. N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, arXiv preprint arXiv:1701.06538. 2017. URL: https://arxiv.org/abs/1701.06538.

2. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning (Vol. 1, p. 2). Cambridge: MIT press. 2016.

3. I. Arel, D.C. Rose, T.P. Karnowski, Deep Machine Learning – a New Frontier in Artificial Intelligence Research. Computational Intelligence Magazine, IEEE. 2010. Vol. 5, No. 4. P. 13-18. DOI: 10.1109/mci.2010.938364

4. Artificial Intelligence Index Report 2019 / Human-Centered Artificial Intelligence Institute, Stanford University. URL: https://hai.stanford.edu/ai-index/2019

5. P. Bezak, Building recognition system based on deep learning. In 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR) (pp. 1-5). IEEE. 2016.

6. F. Marra, G. Poggi, C. Sansone, L. Verdoliva, A deep learning approach for iris sensor model identification. Pattern Recognition Letters, 113, pp.46-53. 2018.

7. M. Zhang, P. Wang, X. Zhang, Vehicle Color Recognition Using Deep Convolutional Neural Networks. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science. pp. 236-238. 2019.

8. A. Tang, K. Lu, Y. Wang, J. Huang, H. Li, A real-time hand posture recognition system using deep neural networks. ACM Transactions on Intelligent Systems and Technology (TIST), 6(2), pp.1-23. 2015.

9. T. Hope, Y.S. Resheff, I. Lieder, Learning tensorflow: A guide to building deep learning systems. "O'Reilly Media, Inc.". 2017.

10. Building a Color Recognizer in Python. URL: https://towardsdatascience.com/building-a-color-recognizer-in-python-4783dfc72456

11. K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mane, D. Fritz, D. Krishnan, F.B. Viégas, M. Wattenberg, Visualizing dataflow graphs of deep learning models in tensorflow. IEEE transactions on visualization and computer graphics, 24(1), pp.1-12. 2017.

[1] **Maryna Antonevych**
Student of the Faculty of Information Technology
[2] **Anna Didyk**
Student of the Faculty of Information Technology
[3] **Vitaliy Snytyuk**
Dr. of. Sci., Professor, Dean of the Faculty of Information Technology
[1,2,3] *Taras Shevchenko National University of Kyiv*

CHOICE OF BETTER PARAMETERS FOR METHOD OF DEFORMED STARS IN N-DIMENSIONAL CASE

Abstract. This paper is devoted to the problem of optimization of a function in n -dimensional space, which in the general case is polyextreme and undifferentiated. In contrast to the classical method of deformed stars [4], we obtained a method that solves problems in n-dimensional space, where the population of solutions consists of 3-, 4-, and 5-point groups. The advantages of the developed method over genetic algorithm [1], differential evolution [2] and evolutionary strategy [3] as the most typical evolutionary algorithms are shown. Testing was performed to investigate the best configuration of method of deformed stars parameters.

Keywords: Function, optimization, method of deformed stars (MODS), n-dimensional space.

Introduction. A large number of modern practical problems belong to the class of constraint satisfaction problems (CSPs). Stochastic search, combinatorial optimization methods, and evolutionary algorithms are used to solve such tasks. Exactly, the use of evolutionary algorithms does not require strict target functions constraints and also does not guarantee the finding of a global optimum, although according to certain conditions there is a probability convergence.

General algorithm of the method.

Step 1. We initialize the parameters of the algorithm, let $t = 0$.

Step 2. Generate m potential solutions in domain D (population P_t).

Step 3. Form w figures (triangles, rectangles or pentagons) $F_i, i = \overline{1, w}$.

Step 4. For each F_i find the vertex in which the function f takes the best value and consider it the best vertex.

Step 5. Find the center of each figure as the average of all its vertices.

Step 6. For each figure F_i find a compressed figure T_i in which the best point is transferred along the line connecting the center of the figure and the best point, and all the others are transferred to it.

Step 7. For each figure F_i find a compressed figure U_i, in which all points are compressed to the best vertex.

Step 8. For each figure F_i find the figure Q_i, obtained by rotating F_i around the best vertex.

Step 9. For each figure F_i find the figure B_i, obtained by rotating F_i around the center of the figure.

Step 10. For each figure F_i find a modified figure R_i.

Step 11. Form a general population P_t, which will contain all the new points created in the previous steps. Thus, $P_t = P_t + T_i + U_i + Q_i + B_i + R_i, i = \overline{1,w}$.

Step 12. For all potential solutions from P_t find the value of the function f and sort the potential solutions from the best to the worst.

Step 13. Leave in P_t only m best solutions and check the fulfillment of the stop criterion.

Step 14. If the stop condition is not met, go to step 3. Otherwise, complete the algorithm and the best element in the population will be considered the best solution.

Note that the stop criterion may be:

- a given number of iterations;
- the worst value of fitness function in neighboring populations is less than specified;
- the average value of fitness function in neighboring populations is less than specified, etc.

The experiment results. Table 1 shows the results for the Schwefel 2.20 function in 10-dimensional space. The comparative graph of results is presented in fig. 1. It shows how well-known methods and MODS found a solution during the execution with the stop condition by iterations. As we can see, MODS found a solution on the initial iterations, in contrast to other known methods.

Table 1.

Comparison of the results of methods under three conditions of completion

Test function	Global minimum	1000 iterations, 10 launches, accuracy = 10^{-5}					
		GA	ES	DE	MODS-n3	MODS-n4	MODS-n5
	f	f	f	f	f	f	f
Schwefel 2.20	0	0.143	194.1793	79.7148	0.0	0.0	0.0
Worst fitness-function between populations, epsilon = $1*10^{-10}$							
Schwefel 2.20	0	8.4094	191.1157	283.9126	0.0	0.0	0.0
Average fitness-function between populations, epsilon = $1*10^{-10}$							
Schwefel 2.20	0	3.6056	244.8562	170.2587	0.0	0.0	0.0

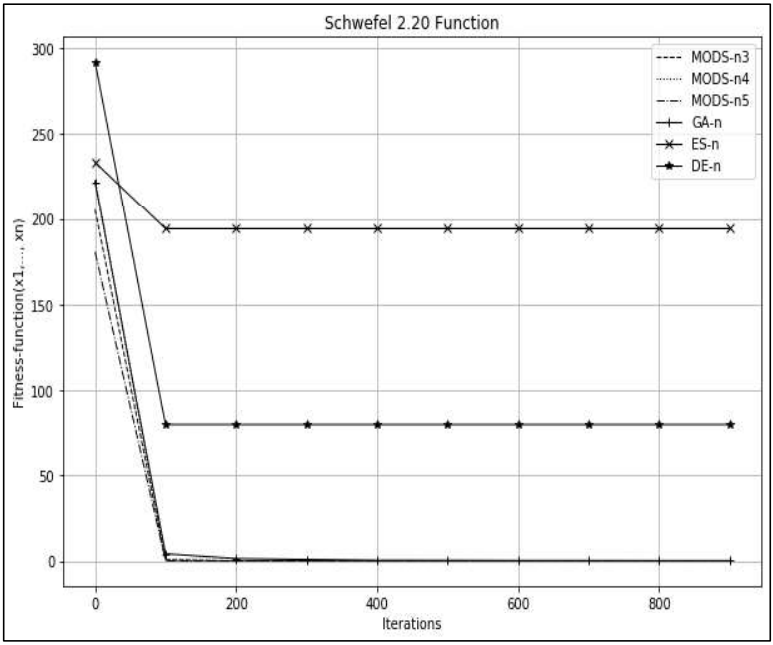

Figure 1 - Convergence plots for the Schwefel 2.20 function, the condition by iterations, coefficient = 2

Choice of better parameters for MODS. As we can see, first experiments were made with the parameters, which are used to perform transformations in MODS populations, equal to 2, for all stop conditions.

And MODS with these parameters was used for comparison with well-known methods in Table 1.

To investigate the best configuration of the MODS, it was decided to conduct testing, considering the different values of the input parameters of the method. For the study, it was decided to set the values of all parameters, which are used to perform transformations in MODS populations, equal to 1.5.

After conducting experiments, it was found that even with the change of parameters, MODS was able to find the correct extremes of the tested functions.

Thus, we can say that when changing the parameters of the method, the solution will still be found.

However, it is important to note, that for coefficients = 1.5 the method worked faster. Therefore, the change in the value of the parameter affected the speed of finding the result.

The comparative graph of results is presented in fig. 2. It shows how well-known methods and MODS (but, in this case, with parameters, that are equal to 1.5) found a solution with the execution of the stop condition by iterations.

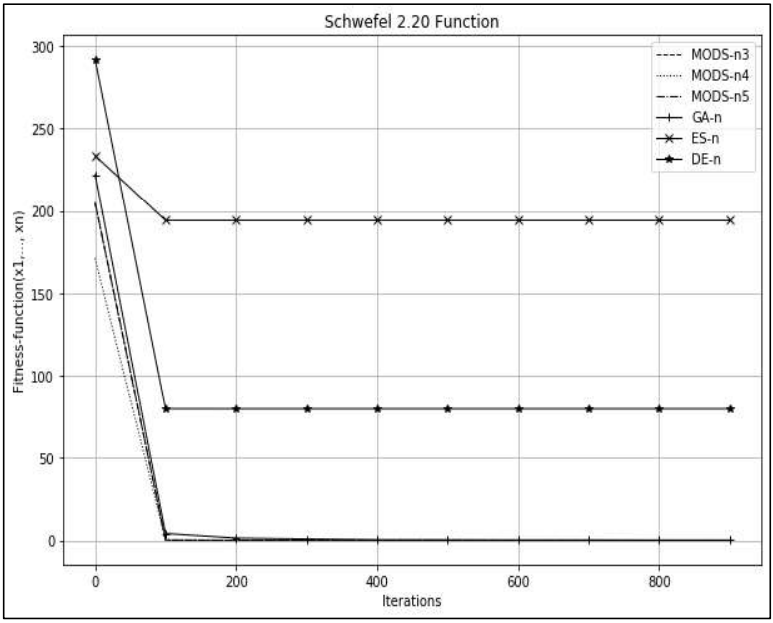

Figure 2 - Convergence plots for the Schwefel 2.20 function, the condition by iterations, coefficients = 1.5

Conclusion. In implementing the method of deformed stars, significantly fewer steps performed in the wrong direction, in contrast to genetic algorithm, the method of differential evolution and evolutionary strategy as representatives of classic evolutionary paradigm. The accuracy of the obtained solutions is, on average, higher than that of competing algorithms due to a deeper study of the solution search area. It is also important to note that in the course of research it was found that when changing the parameters of the MODS solution will still be found.

References:

1. J.H. Holland (1975) Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, Michigan; re-issued by MIT Press (1992).

2. Storn, R., Price, K. Differential Evolution – a Simple and Efficient Heuristic for Global Optimization over Continuous. Journal of Global Optimization 11, 341-359 (1997).

3. I. Rechenberg (1973) Evolutions strategic: Optimierung technischer Systeme nach Prinzipen der biologischen Evolution, Frommmann-Holzboog Verlag, Stuttgart (2nd edition 1993).

4. V. Snytyuk, "Method of Deformed Stars for Multi-extremal Optimization. One- and Two-Dimensional Cases', in International Conf. Mathematical Modeling and Simulation of Systems. MODS 2019, Advances in Intelligent Systems and Computing, vol 1019. Springer, Cham.

5. Antonevych M., Didyk A., Snytyuk V. Optimization of functions of two variables by deformed stars method // In Proc. 2019 IEEE International Conference on Advanced Trends in Information Theory ATIT, Kyiv, Ukraine.

[1] **Anton Astakhov**
Master's student
[2] **Oleh Ilarionov**
PhD in Engineering Science, Associate Professor
[1,2] *Taras Shevchenko National University of Kyiv*

ANALYSIS OF SPEECH EMOTION RECOGNITION METHODS

Abstract. Human beings can naturally recognize emotions from speech, but building automated speech emotion recognition systems is a challenging and relevant problem. Such systems could be used for interactive voice-based assistants and customer satisfaction analysis.

Keywords: Speech emotion recognition; Speech Datasets; Classification; Emotion Recognition

Speech is the most common way for humans to communicate with each other. People can not only understand the meaning of words, but can also extract a lot of useful information from the speaker's emotions.

Speech emotion recognition (SER) — is the process of identifying human emotions embedded in their speech [1]. However, human beings can naturally solve this problem, it is still a challenging task to automatically recognize human emotions. Therefore, SER is a relevant topic for scientific research. We have examined the structure of SER systems and identified the crucial parts of them, such as datasets and classifiers.

Speech emotion recognition system usually consists of three main components: signal acquisition, feature extraction, emotion recognition. General framework is shown in Figure 1 [2].

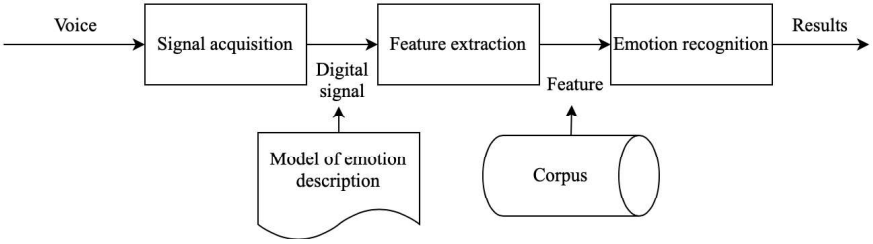

Figure 1 — SER system block diagram

One of the key stages to building SER systems is to find correct and complete dataset. Audio recordings should be performed by diverse groups of people (men and

women of different ages with different accents), who can express the whole range of emotions (anger, happiness, sadness, disgust, surprise, fear, etc.). Usually, this kind of audio recordings are made by professional actors [3]. Moreover, there are some other criteria to match when choosing datasets, such as English language and free to use license (at least for scientific purposes). Datasets, that match all of the criteria, are presented in Table 1.

Table 1 — Review of SER datasets

Dataset	**Emotions**	**Size**
Surrey Audio-Visual Expressed Emotion (SAVEE) (Surrey audio-visual expressed emotion database, 2019)	Anger, disgust, fear, happiness, sadness, surprise, neutral, common	14 speakers (male), 120 utterances
Toronto Emotional Speech Database (TESS) (Toronto emotional speech database, 2019)	Anger, disgust, neutral, fear, happiness, sadness, pleasant, surprise	2 speakers (female), 2800 utterances
eNTERFACE'05 Audio-Visual Emotion Database (Martin et al., 2006)	Anger, disgust, fear, happiness, sadness, surprise	42 speakers (34 male, 8 female) from 14 nationalities, 1116 video sequences
SAMAINE Database (McKeown et al., 2011)	Valence, activation, power, expectation, overall emotional intensity	150 speakers, 959 conversation
TUM AVIC Database (Schuller et al., 2009)	Five level of interest; 5 non-linguistic vocalizations (breathing, consent, garbage, hesitation, laughter)	21 speakers (11 male, 10 female), 3901 utterances
AFEW Database (Kossaifi et al., 2017)	Anger, disgust, surprise, fear, happiness, neutral, sadness	330 speakers, 1426 utterances from movies, TV-shows

Another key component of building a SER system is to choose the right classifier, which is a method to classify underlying emotions for a given utterance [4]. There are several different options to choose from, such as machine learning, traditional classifiers, deep learning algorithms. However, there is no consensus on what should be the ultimate classifier to solve the problem. Classifiers used in literature are presented in Table 2.

Table 2 — Review of classifiers, studies, datasets, and results

	Number of studies	**Datasets**	**Results**
HMM	7 [5, 6]	Berlin Emo DB	79%
GMM	4 [5, 7]	Berlin Emo DB, EPSAT, EMA, GES, SES, WSJ	75%
SVM	10 [5, 8]	LDC, Berlin Emo DB, AIBO DB, ABC DB, SUSAS DB, EMA DB, VAM I-II DBs, VAM DB	74%
MLP	5 [5, 9]	Berlin Emo DB	72%
kNN	2 [9, 10]	One natural and one acted speech corpora in Mandarin	66%
Decision Tree	3 [5, 8]	AIBO DBUSC IEMOCAP DB	58%
Rule Based Fuzzy Estimator	1 [11]	EMA DB, VAM I-II DBs	70%
DNN	2 [12]	IEMOCAP DB	54%
CNN	5 [5, 13]	SAVEE DB, Berlin EMO DB, DES DB, MES DB, RECOLA DB	70%
RNN	6 [5, 14]	IEMOCAP DB	63%

Among the most common classifiers are Hidden Markov models, Support vector machines, and Recurrent Neural Networks. The most precise are Hidden Markov models and Gaussian Mixture Models. The most underrated are Convolutional Neural Networks and Rule Based Fuzzy Estimators.

References:

1. Mehmet Berkehan Akçay, «Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers», Department of Software Engineering, Izmir University of Economics, Izmir, Turkey, 13.12.2019.

2. Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng, «A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM», 12.08.2016.

3. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA. A review on speech emotion recognition: Case of pedagogical interaction in classroom. In: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE; 2017. pp. 1-7.

4. Helen Chan, Travis Ebesu, Caleb Fujimori, «Speech Emotion Detection and Analysis», COEN296: Natural Language Processing, Department of Computer Engineering Santa Clara University, 04.12.2018.

5. Babak Basharirad and Mohammadreza Moradhaseli, «Speech emotion recognition methods: A literature review» (2017)

6. B. Schuller, G. Rigoll, M. Lang, «Hidden markov model-based speech emotion recognition», Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on, 2, IEEE (2003), pp. II-1.

7. O.-W. Kwon, K. Chan, J. Hao, T.-W. Lee «Emotion recognition by speech signals», Eighth European Conference on Speech Communication and Technology (2003).

8. M. Borchert, A. Dusterhoft, «Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments», Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, IEEE (2005).

9. B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll, «Speaker independent speech emotion recognition by ensemble classification», Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, IEEE (2005).

10. J. Rong, G. Li, Y.-P.P. Chen, «Acoustic feature selection for automatic emotion recognition from speech», Inform. Process. Manag., 45 (3) (2009).

11. M. Grimm, K. Kroschel, E. Mower Provost, S. Narayanan, «Primitives-based evaluation and estimation of emotions in speech», Speech Commun., 49 (2007).

12. K. Han, D. Yu, I. Tashev, «Speech emotion recognition using deep neural network and extreme learning machine», Fifteenth annual conference of the international speech communication association (2014).

13. Q. Mao, M. Dong, Z. Huang, Y. Zhan, «Learning salient features for speech emotion recognition using convolutional neural networks», IEEE Trans. Multimed., 16 (8) (2014).

14. S. Mirsamadi, E. Barsoum, C. Zhang, «Automatic speech emotion recognition using recurrent neural networks with local attention», 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2017).

[1] **Tamara Bondar**
Bachelor of Computer Science
[2] **Hryhorii Hnatiienko**
PhD
[1,2] *Taras Shevchenko National University of Kyiv*

VIDEO REGISTRATION AND FACE RECOGNITION TECHNOLOGY ON STREAM VIDEO

Abstract. In this paper, the analysis of modern methods of face recognition on streaming video in real-time is carried out and the system analysis for the development of the intelligent system and the user interface for work with it is carried out.

Keywords: video registration, recognition, database, streaming video.

One of the systems that can increase security in public places and educational institutions is the face recognition system. Currently, many systems of this type are used in the world, the most famous of which are SkyNet [1], FacePro [2]. Based on the analysis of existing systems [3], it was decided to develop and adapt a similar system for education.

During the analytical review of existing solutions, a general description of the process of visitor identification was formed in the form of an IDEF0 diagram (Fig. 1).

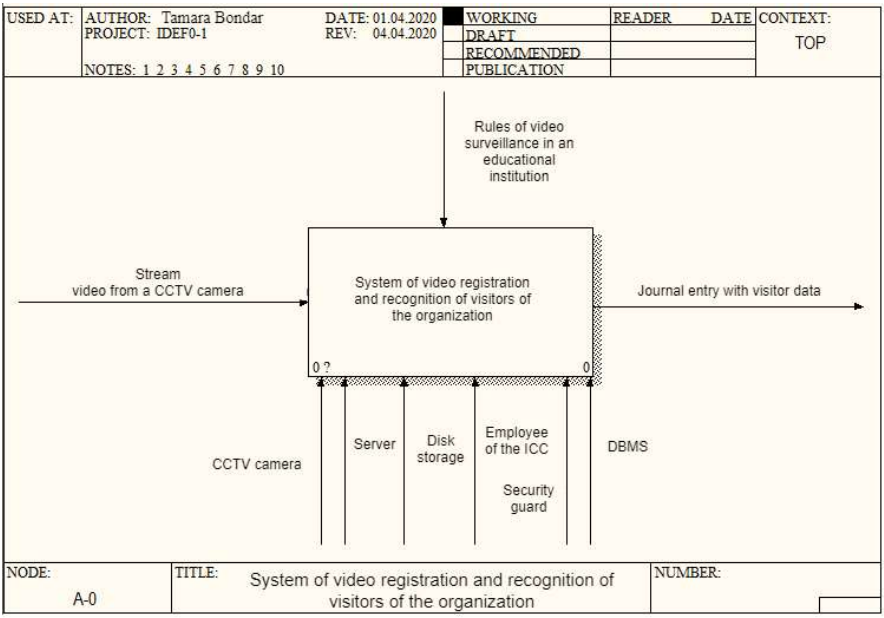

Figure 1 - IDEF0 diagram of the process of identifying visitors to the organization

At the entrance, the system must receive a streaming video, which will be broadcast from a video surveillance camera located in the direction of the flow of visitors to the organization. The result of the system should be an entry with the visitor's data in the electronic journal.

The operation of the video recording and visitor recognition system of the organization includes three processes:

- search and highlight faces on streaming video;
- search for a visitor in the database;
- making a record of the visitor in the electronic journal.

The intelligent system of video registration and identification of visitors of the organization must receive streaming video at the entrance and interpret the information received in it based on the results of personal identification. Comparative analyzes described in the sources [4-5] prove that convolutional networks will fully satisfy the requirements.

The main feature of the developed network: it was based on the MobileFaceNet architecture [6], which uses the ReLU activation function, alternation of deep cores, and bottleneck structures. This type of architecture is ideal for the limited capacity of the technical resources of the organization.

In order to determine the level of similarity of facial characteristics, it is necessary to find the distance from the characteristic removed by the convolutional neural network to the nearest similar to it among the characteristics of identified visitors stored in the database. The Euclidean distance square is used to give more weight to those class implementations that are significantly separated from each other. This feature is ideal for face recognition, as it allows to more accurately separate the facial features of visitors from each other.

References:

1. Weida Li, China updates Skynet system with facial recognition, 2018. URL: https://bit.ly/3onLqCQ

2. FacePRO: Panasonic Facial Recognition System, Panasonic Security System, 2019. URL: https://bit.ly/35rIfBs.

3. Tamara Kopchyk, Analysis of intelligent search and face recognition systems on streaming video. Information technologies and interactions: materials VI International. scientific-practical conf., Kyiv, December 20, 2019 / Ministry of Education and Science of Ukraine, "National Taras Shevchenko University of Kyiv", and [other] - p. 232-233.

4. Nicolas Delbiaggio, A comparison of facial recognition's algorithms, Degree Programme in Business Information Technology, Bachelor's Thesis, 2017. URL: https://bit.ly/31CSH8g.

5. Thai Hoang Le, Applying Artificial Neural Networks for Face Recognition / Thai Hoang Le. – Ho Chi Minh City, 2011. – p. 16. – (Department of Computer Science, Ho Chi Minh University of Science).

6. X.Gao, Y. Liu, S. Chen, Z. Han, MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices URL: https://arxiv.org/abs/1804.07573.

[1] **Oleksandr Derevianchenko**
PhD., Associate Professor Faculty of Computer Science and Cybernetics
[2] **Andrii Nikolaiev**
M.S. CS
[1,2] *Taras Shevchenko National University of Kyiv*

IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE MODULE FOR LEARNING PURPOSES

Abstract. Games are a great environment to explore different methods and algorithms. Developers often rely in their research on the ability to train AI with the help of games. It is also known that children acquire complex skills by learning and applying different patterns of behavior with a low level of risk while playing games. The purpose of the work is to implement AI module using modern machine learning methods and show some benefits for learning purposes.

Keywords: gaming artificial intelligence, reinforcement learning, gamification.

For the base algorithm was taken the idea of combining a tree-based search with a learned model which is presented in the [1] by Google DeepMind (2019). It is a reinforcement learning algorithm which could master games without having any knowledge about the game rules and environment dynamics making it better then state of the art model-based and free-based domain algorithms respectively.

According to the algorithm idea, there are provided four components that run simultaneously in the one loop. The shared storage holds the latest neural network weights, the self-play uses the weights to generate self-play games and place them in the replay buffer. The played games are used to train a network and store the weights in the shared storage (Figure 1).

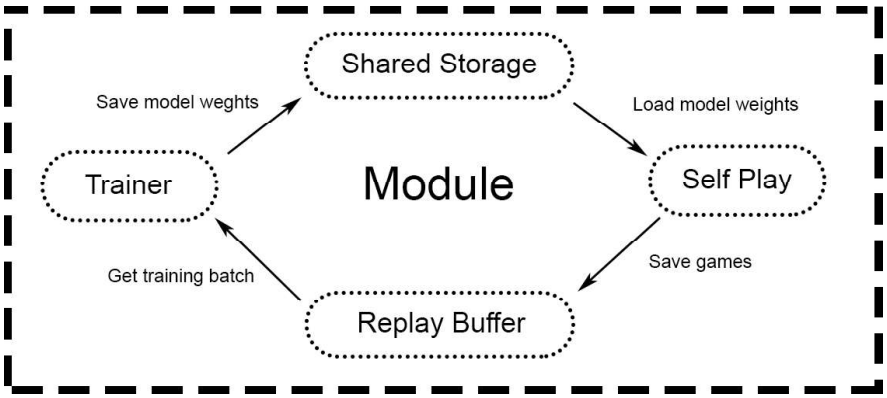

Figure 1 – How the module works.

Google Colaboratory was used for the model training as it has some powerful tools for machine learning in cloud computing [2]. The model was making about $T_{steps} = 100$ training steps per minute during the training what makes the module competitive in such game as Connect 4 less than a few hours (Figure 2).

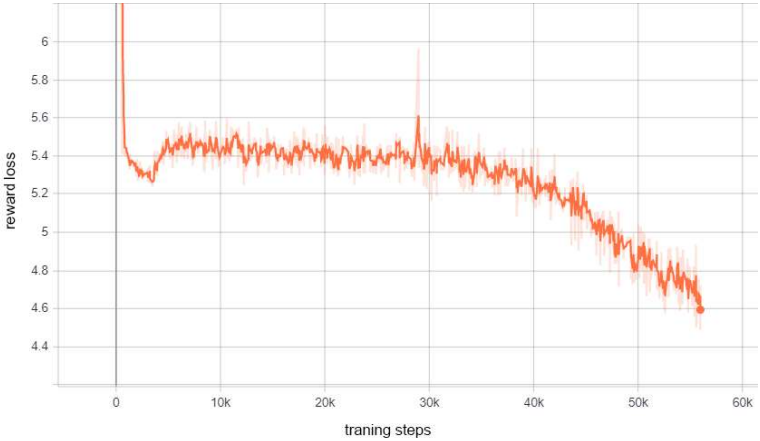

Figure 2 – Reward loss for the game Connect 4.

After reaching the certain level of playing, the developed module can be applied for the learning purposes: e.g. in the games position analysis; for searching the best move at the given position; to provide an expert evaluation for professional or beginner players (in chess and other board games).

The idea could be also applied for some educational projects as a gamification tool which could help students to master some of the board games and develop their own critical and logical thinking. Such method is highly approved as players tend to have more motivation to learn [3].

The strong point about the work is that the module could be used for different game domains (including more visually complex ones) which could bring higher variety of environments, where such learning skills could be developed.

References:

1. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2019). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. ArXiv, abs/1911.08265.

2. Google Colaboratory. URL: https://research.google.com/colaboratory/faq.html.

3. Burgers, C, Eden, A, van Engelenburg, M. (2015) How feedback boosts motivation and play in a brain-training game. Computers in Human Behavior 48: 94–103.

[1] **Yuliia Hlavcheva**
Deputy Director of the Library
[2] **Maksym Glavchev**
PhD in Economics, Associate Professor
[3] **Victoria Bobicev**
PhD in in Computer Science, Associate Professor
[3] **Olga Kanishcheva**
PhD in in Computer Science, Associate Professor
[1,2,4] National Technical University "Kharkiv Polytechnic Institute", Ukraine
[3] Technical University of Moldova, Republic of Moldova

LANGUAGE-INDEPENDENT FEATURES FOR AUTHORSHIP ATTRIBUTION ON UKRAINIAN TEXTS

Abstract. Authorship attribution is the natural language processing task of identifying the author of an input text. The main goal of this task is to define salient characteristics of documents that capture the writing author's style. In this paper, we analyze language-independent features for authorship attribution. For the experiments we used ML methods. The experimental results on scientific text classification showed that DT method outperforms most other ML methods, and these language-independent features are appropriate for the Ukrainian scientific documents authorship attribution.

Keywords: Writing Style, Language-Independent Features, Authorship Attribution, Text Classification, Machine Learning Methods.

The task of authorship identification is not new. The results of authorship detection studies are actively used in various spheres of human life. Authorship research can be divided into three main areas: authorship identification, authorship characterization, similarity detection [1]. Similarity detection is most often used to identify potential academic plagiarism [2]. This topic is relevant and important.

This paper focuses on the third direction of similarity identification. The decision is based on the group of properties that reflect the author's style measurement and comparison. Stylometric properties by which the author's style can be identified make up a list of style markers. All style markers are equally effective when used for different languages [3]. We distribute the stylometric characters into two groups: language-dependent features and language-independent features. This paper focuses on the study of the statistical characteristics of scientific texts in the Ukrainian language, which can be attributed to language-independent stylometric properties.

For the experiments we used our own preprocessed text corpus. The text corpus consists of individual scientific publications in Ukrainian. For stylometric properties we used only the paper main text, which best reflects the author's written style. For each author, a collection of paper fragments is formed. In our case, stylometric properties are determined for each fragment of the paper separately.

For stylometric properties, we used only the paper main text. We created two subsets for our experiments, one of them consist of 8 classes (8 authors, 1019

fragments), other – 32 classes (32 authors, 2633 fragments). These classes were selected randomly but sets are balance.

Groups of properties, which refer to the text statistical parameters and allow to determine the author's style with high accuracy are described in [3]. The authors of this paper created their own list of text statistical properties of Ukrainian, which are divided into 5 groups: average number of words in a sentence, average word length, average word frequency, punctuation (5 indicators); the number of words with length from 1 to 20 characters, the number of words with a word frequency from 1 to 8 times (28 indicators); frequency of using letters of the Ukrainian alphabet (33 indicators); frequency of using stopwords and pronouns (65 indicators); coefficients of language diversity (5 indicators) [4].

For our experiments of text classification, we took our corpus (2 subsets) and five groups of features: for separate groups and their combination. Weka software was used for classification task. Bayes Based Algorithms (Naive Bayes Multinomial, NBM), Support Vector Machine (SMO), Decision Trees (LMT, J48) were used as classification methods with the cross-validation parameter – 10 folds.

We conducted experiments for 32 authors (32 classes) and 8 authors (8 classes) and compared the results. Experiments showed that for 1-3, 1-4 and 1-5 groups of properties, the classification indicators are similar, despite the increase in the number of features. The best result (F-measure) of 32 classes we received for the SMO method (0.586) and LTM (0.614) for 1-5 groups of properties. The best result (F-measure) of 8 classes we received for the SMO method (0,794) and LTM (0,806) for 1-5 groups of properties.

According to the experiments, we obtained for 8 classes an average increase in the value of Correctly Classified Instances – 20%, MIN increase in the value Correctly Classified Instances – 15%, MAX increase in the value Correctly Classified Instances – 28%. The result of the experiments demonstrated the usefulness of the proposed language-independent stylometric properties indicators for text authorship attribution.

References:

1. B. Alhijawi, S. Hriez, A. Awajan, Text-based authorship identification - A survey. Paper presented at the 5th International Symposium on Innovation in Information and Communication Technology, ISIICT 2018. 2018, pp. 1-7. doi:10.1109/ISIICT.2018.8613287.

2. M. AlSallal, R. Iqbal, V. Palade, S. Amin, V. Chang, An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, Vol. 96., 2019 pp. 700-712. doi:10.1016/j.future.2017.11.023.

3. S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, M. Veinovic, Automated language-independent authorship verification (for Indo-European languages). Journal of the Association for Information Science and Technology, Vol. 70.8, 2019, pp. 858-871. doi:10.1002/asi.24163.

4. V. A. Vysotska, V. V. Pasichnyk, Yu. M. Shcherbyna, T. V. Shestakevych. Matematychna lingvistyka. Knyha 1. Kvantytatyvna linhvistyka, Lviv, Novyi Svit–2000, 2012.

Kyrylo Kadomskyi
Assistant Professor
Taras Shevchenko National University of Kyiv

EVALUATING DEEP LEARNING MODELS FOR ANOMALY DETECTION IN AN INDUSTRIAL TRANSPORTING SYSTEM

Abstract. In Cyber-Physical Production Systems (CPPS), the task of anomaly detection is commonly solved by model-based methods. While these methods have proven effective in some use cases, their performance can drop dramatically in other systems. In this study the problem of representativeness of evaluation of such methods is addressed. The CPPS data is used, on which existing models have proven ineffective. The perspective of applying deep learning approach to constructing a process model in such systems is investigated.

Keywords: anomaly detection, autoencoder, model evaluation, cyber-physical production systems, industrial IoT.

Modern industrial plants demonstrate both increasing pressure for efficiency, and new possibilities to use growing set of sensors to facilitate automation. In this context, diagnosis of complex production processes has gained new attention due to research agendas such as Cyberphysical Production Systems (CPPS) [1]. One of the most important goals is self-diagnosis, which includes identification of anomalous system behavior, suboptimal energy consumption, or wear in CPPS.

The model-based diagnosis is a commonly used approach, in which a dynamic process model captures spatio-temporal features of the system's behavior. Considering the challenges of CPPS agendas, precise mathematical or expert modeling is infeasible in most cases. Thus, novel dynamic modelling techniques are being developed for learning the model from system observations [1, 2].

Applying these techniques in CPPS poses two main challenges:

1. While showing good results in certain applications, existing models yield relatively poor performance in other similar use cases [2, 3, 4]. The hypothesis is that this effect is due to limited nature and fixed structure of spatio-temporal features learned by the model, which are imposed by the structure of the model itself. Then the informativeness of learned features will vary in different physical systems, which can explain the observed effect. To meet this challenge, more generalized models are required.
2. As previous studies suggest, results of model evaluation in some systems may not be representative [2]. Thus, new evaluation criteria are required for representative comparison and benchmarking of the models.

In this study the two mentioned challenges are addressed. Deep Learning models, such as autoencoders, are applied to remove the first limitation by automatically selecting the most relevant features and structure to represent the data. To address the second problem, two robustness criteria are proposed for representative model evaluation: reconstructed variation rate and reconstruction's sensitivity to anomalies.

These criteria are assessed from the statistical distributions of model's response to normal and anomalous data.

A set of autoencoder architectures were modelled and evaluated on the HRSS dataset [5], which has proven challenging for applying novel dynamic models [2, 4]. Proposed robustness criteria were used in conjunction with traditional performance metrics, aiming for accurate benchmarking of the two approaches. This in turn provides the possibility to assess the limits of model-based anomaly detection in given class of CPPS.

Each model is trained in unsupervised manner to reconstruct normal time series, targeting for minimal reconstruction loss. Then the trained model is used to reconstruct unseen time series with anomalies, where the reconstruction error is expected to peak at anomalous intervals. To evaluate the model, the distributions of reconstruction error in normal and anomalous intervals are analyzed for being statistically distinguishable. Finally, from these error distributions, a decision-rule classifier for anomaly detection is built in a supervised mode. This method detects anomalies with time step precision, and most of evaluated models can be applied in real time.

6 LSTM and 3 ConvNet architectures were empirically tested. A grid search approach was applied to each architecture in order to select optimal model's hyper-parameters. It was shown that increasing model complexity, both in LSTM and convolution-based models, allows to increase anomaly detection performance, but has a strong robustness tradeoff. This indicates that model evaluation in CPPS of this class cannot rely completely on performance metrics. Considering both performance measure and proposed robustness criteria, a single LSTM model is selected for HRSS data. Comparing to the baseline efficiency [2, 4], an increase by 102% in anomaly detection score and an increase by 121% in recall are achieved.

References:

1. O. Niggemann, C. Frey, Data-driven anomaly detection in cyber-physical production systems, AT – Automatisierungstechnik, 2015, vol. 63, issue 10. doi: 10.1515/auto-2015-0060.

2. A. von Birgelen, O. Niggemann, Using self-organizing maps to learn hybrid timed automata in absence of discrete events, in: Proceedings of the 22nd IEEE international conference on Emerging Technologies and Factory Automation, ETFA, Limassol, Cyprus, 2017, pp. 1–8.

3. N. Hranisavljevic, O. Niggemann, A. Maier, A novel anomaly detection algorithm for hybrid production systems based on deep learning and timed automata, in: Proceedings of the 27th international workshop on Principles of Diagnosis, DX-2016, Denver, Colorado, 2016.

4. M. Cerliani. Predictive maintenance with LSTM siamese network, 2019. URL: https://towardsdatascience.com/predictive-maintenance-with-lstm-siamese-network-51ee7df29767.

5. Physical factory / demonstrators IMPROVE, 2016. URL: http://improve-vfof.eu/background/physical-factory-demonstrators.

[1] **Iryna Nazarchuk**
Student
[2] **Hanna Krasovska**
PhD in Engineering Science, Associate Professor
[3] **Oleh Ilarionov**
PhD in Engineering Science, Associate Professor
[1,2,3] *Taras Shevchenko National University of Kyiv*

INTELLECTUAL AGENT FOR SENTIMENT ANALYSIS ON MOVIE REVIEWS

Abstract. Movie reviews collection from IMDb website was used as the main dataset and evaluation of the text sentiment by negative/positive attitude was performed. The research proposes bag-of-words and tf-idf models. The core of the intellectual agent system model was trained using such classification algorithms - Logistic Regression, SGD, Random Forest and a Deep Learning model in a form of feedforward neural network. The comparative analysis of the accuracy scores was held and model with the highest AUC rate was chosen for the system integration.

Keywords: Motion pictures, Sentiment analysis, Feature extraction, Classification algorithms, Data mining, Opinion mining, Movie review, Information Retrieval, Deep learning, Neural networks, Classifier.

The modern stage of human development involves a rapid increase of information generation. With the spread of Internet and web platforms, where every user is given an opportunity to express their opinions regarding any type of product or service, as well as a movie or book, it is an urgent and essential task to handle these huge amounts of data to determine the attitude of users towards a particular object.

In this paper, the definition of the sentiment or polarity of reviews for films is being examined. By constructing an intellectual agent that automatically extracts and identifies opinion within the text, given as an input, in particular, movie reviews (comments or posts in social networks) in order to determine the mood of the audience in relation to a particular movie.

In the process of analyzing already existing systems used for automated classification of the text sentiment such systems as RapidMiner [1], GATE [2], Google Cloud Prediction API [3] revealed the instability of the accuracy while determining the emotional expression of the submitted movie reviews. These systems are currently in high demand and gain huge popularity in market researches to analyze reviews for commercial products [4]. The results of the semantic analysis using the above-mentioned systems have shown that the accuracy rate of sentiment classification for input text fragment is 87% when using the RapidMiner system, 92% utilizing the Google Cloud Prediction API service, 90% with the help of GATE software application. Also, while using the above-mentioned systems in multiple iterations, the accuracy of the classification has undergone significant fluctuations of classification metrics scores. According to the author, the main disadvantage of these systems is their

general-purpose focus, which does not take into account the semantic subtleties inherent in such a genre as movie reviews.

The purpose of this study is to develop a software product to increase the effectiveness of the sentiment analysis of the opinion expressed in text and ensure the stability of classification metrics, regardless of the nature of the input information.

To build the classification model, a collection of movie reviews data from the IMDb web site [5] was used for machine learning model training. The polarity of the reviews was based on the type of review (positive, negative and partly positive or partially negative) chosen by the author. In order to apply these data as an input parameter for the classification algorithms, the pre-processing phase was applied, which included such procedures as stemming, tokenization, stop words elimination and lemmatization. In order to form a training dataset for machine learning models from the texts of movie reviews, a set of features was constructed, having a numerical value, that is, the procedure of vectorizing the text was carried out. As a way to represent vectorized text in a form of features, models bag-of-words [6] and tf-idf [7] were used.

For the purpose of training, three classification algorithms were selected: Logistic Regression [8, 9], SGD [10], and Random Forest [11]. Alternatively, to find the best model that performs a sentiment analysis to determine the polarity of the movie review with high accuracy, training of the feedforward neural network with a different configuration of hyperparameters and layered architecture was performed.

To compare the results of the trained models' accuracy of the classification, the AUC [12] metric was used, in the assessment of the ROC curves for each model (fig. 1).

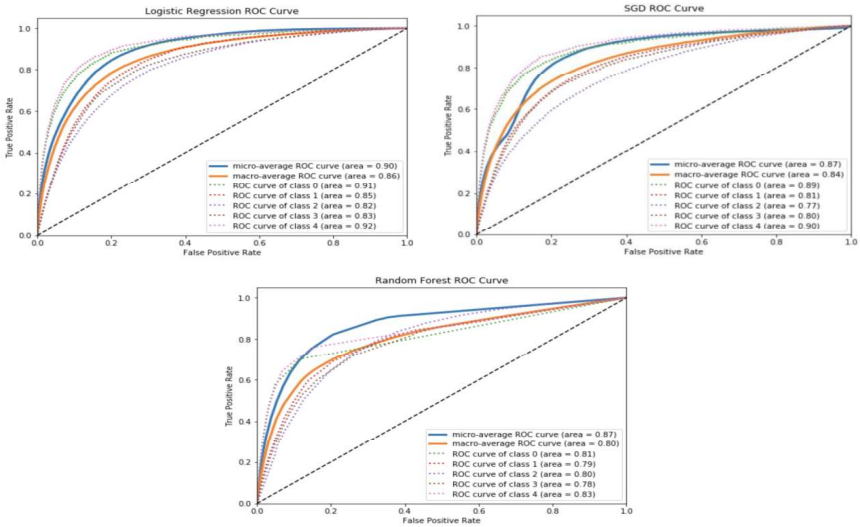

Figure 1 – ROC-curve graphs for logistic regression, stochastic gradient and random forest models

The received results showed that the highest value of the algorithmic accuracy was held by the neural network model with the AUC metric equal to 0.95, that is, the accuracy rate of the input movie review sentiment classification is 95%, using this software. On the basis of this, it can be concluded that the prototype system developed in the framework of this study is the best option for application with the purpose of sentiment analysis of movie reviews.

The prototype, developed in this research can be used as a basis for creating commercial software or as integration into existing systems. The novelty of this work is to use the ensemble of Machine Learning methods to achieve high accuracy of the text data classification.

Further improvement of the algorithm for analyzing the sentiment of the text is possible using deeper levels of natural language processing (syntactic and morphological). Also, it is necessary to investigate the processing of bipolar words, double objections, irony and sarcasm in the text.

References:

1. RapidMiner. Lightning Fast Data Science for Teams. Retrieved May 10, 2019, from https://rapidminer.com/.
2. GATE: a full-lifecycle open source solution for text processing. Retrieved May 1, 2019, from https://gate.ac.uk/overview.html.
3. Cloud machine learning engine. Retrieved April 29, 2019, from https://cloud.google.com/ml-engine/.
4. Pascual F. (2018, August 2). Why and How Companies Should Use Sentiment Analysis. Retrieved from https://www.northeastern.edu/levelblog/2018/08/02/companies-use-sentiment-analysis/.
5. IMDb datasets. Retrieved May 1, 2019, from https://www.imdb.com/interfaces/.
6. Wallach H.M. Topic modeling: beyond bag-of-words. Proc. 23rd Intern. Conf. on Machine learning. ACM, 2006, pp. 977–984.
7. Salton, G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.
8. Kleinbaum D.G., Logistic regression. A self-learning text, Springer-Verlag, 1994.
9. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. N.-Y.: Wiley; 2000. 375 p.
10. Roy R. (2018, April 10). ML | Stochastic Gradient Descent (SGD). Retrieved May 10, 2019, from https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/.
11. Biau G. Analysis of a random forests model. The Journal of Machine Learning Research, 98888:1063–1095, 2012.
12. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Kluwer Acad. Publ; 2004. 38 p.

[1] Tatiana Neskorodieva
Candidate of Technical Sciences (Ph.D.), Head of the Computer Science and Information Technology Department, Associate Professor
[2] Eugene Fedorov
Doctor of Technical Sciences, Professor of the Department of Robotics and Specialized Computer Systems, Associate Professor
[1] *Vasyl' Stus Donetsk National University*
[2] *Cherkasy State Technological University*

AUTOMATIC ANALYSIS METHOD OF AUDIT DATA BASED ON NEURAL NETWORK MAPPING

Abstract. The urgent task of increasing the audit efficiency was solved by automating the mapping of audit indicators by forward-only counterpropagating neural network. A learning algorithm based on k-means has been created, intended for implementation on a GPU using CUDA technology, which increases the speed of identifying parameters of a neural network model.

Keywords: audit, mapping by neural network, forward-only counterpropagating neural network, sequences of payment and supply of raw materials.

Currently, the analytical procedures used during the audit are based on data mining techniques [1]. The aim of the work is to increase the efficiency of automatic data analysis in the audit DSS by means of a neural network mapping of sets of audit indicators in order to identify systematic misstatements that lead to misstatement of reporting. It is assumed that the audit indicators are noisy with Gaussian noise, which in turn simulates random accounting errors (as opposed to systematic ones).

For the achievement of the aim it is necessary to solve the following tasks:

- generate vectors of indicators for objects of sequences of payment and supply of raw materials;
- choose a neural network model for mapping audit indicators (which are noisy with Gaussian noise, which in turn simulates random accounting errors (as opposed to systematic ones, which lead to distortion of reporting));
- choose a criterion for evaluating the effectiveness of a neural network model;
- propose a method for training a neural network model in batch mode;
- propose an algorithm for training a neural network model in batch mode for implementation on a GPU;
- perform numerical studies

Choosing a neural network model for mapping audit sets. In the work, the Forward-only Counterpropagating Neural Network (FOCPNN), which is a non-recurrent static two-layer ANN, was chosen as a neural network [2]. FOCPNN output is linear.

FOCPNN advantages:

1. Unlike most ANNs are used to reconstruct another sample using hetero-associative memory.
2. Unlike bidirectional associative memory and the Boltzmann machine, it works with real data.
3. Unlike a full counterpropagating neural network, it has less computational complexity (it does not perform additional reconstruction of the original sample).

FOCPNN model performing mapping of each input sample $\boldsymbol{x} = (x_1, \ldots, x_{N^x})$ to output sample $\boldsymbol{y} = (w_{i^*1}^{(2)}, \ldots, w_{i^*N^y}^{(2)})$, is represented as

$$i^* = arg\ \min_i z_i,\ z_i = \sqrt{\sum_{k=1}^{N^x}(x_k - w_{ki}^{(1)})^2},\ i \in \overline{1, N^{(1)}}, \quad (1)$$

where $w_{ki}^{(1)}$ – connection weight from the k-th element of the input sample to the i-th neuron,

$w_{i^*j}^{(2)}$ – connection weight from the neuron-winner i^* to j-th element of output sample,

$N^{(1)}$ – the number of neurons in the hidden layer.

Criterion choice for assessing the effectiveness of a neural network model for mapping audit sets. In this work for training model FOCPNN was chosen target function, that indicates selection of the vector of parameter values $W = (w_{11}^{(1)}, \ldots, w_{N^xN^{(1)}}^{(1)}, w_{11}^{(2)}, \ldots, w_{N^{(1)}N^y}^{(2)})$, which deliver the minimum mean square error (difference between the model sample and the test sample)

$$F = \frac{1}{PN^y}\sum_{\mu=1}^{P}\left\|\boldsymbol{y}_\mu - \boldsymbol{d}_\mu\right\|^2 \to \min_W, \quad (2)$$

where $\boldsymbol{y}_\mu - \mu$ -th, output sample according to the model

$\boldsymbol{d}_\mu - \mu$ -th test output sample.

Training method for neural network model in batch mode. The disadvantage of FOCPNN is that it does not have a batch learning mode, which leads to reducing of the learning speed. For FOCPNN was used concurrent training (combination of training with and without a teacher). This work proposes training FOCPNN in batch mode.

First phase (training of the hidden layer) (steps 1-6).

The first phase allows you to calculate the weights of the hidden layer $w_{ki}^{(1)}$ and consists of the following blocks (Fig 1).

Second phase (training the output layer) (steps 7-12). The second phase allows you to calculate the weights of the output layer $w_{ij}^{(2)}$ and consists of the following blocks (Figure 2).

Algorithm for training neuron network model in batch mode for implementation on GPU. For the proposed method of training FOCPNN on audit data example, examines the algorithm for implementation on a GPU with usage of CUDA parallel processing technology.

Numerical research. The results of the comparison of the proposed method using GPU and the traditional FOCPNN training method are presented in Table 1.

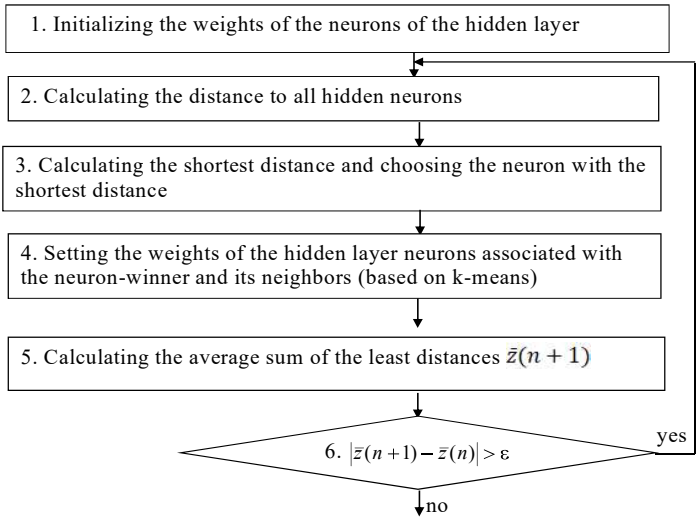

Figure 1. The sequence of steps in training method of FOCPNN in batch mode (the first phase)

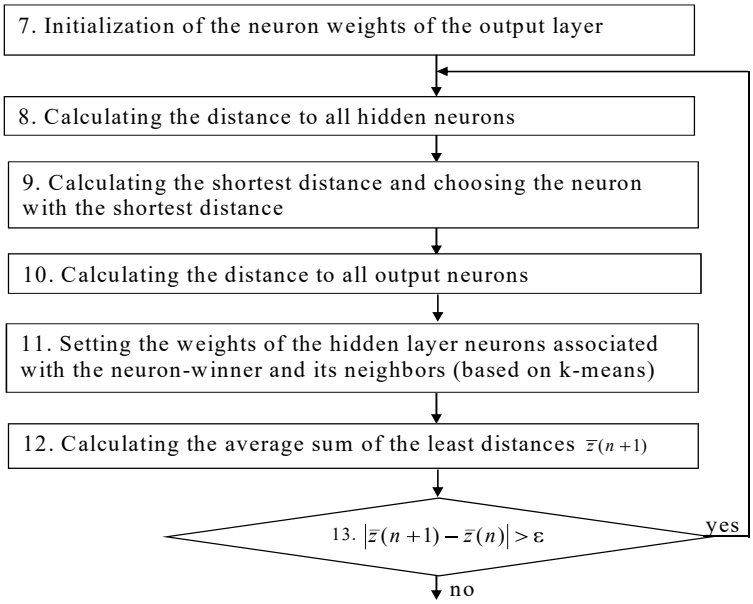

Figure 2. Sequence of procedures for the FOCPNN training method in batch mode (second phase)

Table 1
Comparison of the computational complexity of the proposed and traditional training methods of FOCPNN

Feature	Method	
	proposed	traditional
Computational complexity	$O(n_1^{max} + n_2^{max})$	$O(PN^{(1)}n_1^{max} + (PN^{(1)} + P)n_2^{max})$

Evaluation of computational complexity of the proposed method using the GPU, and the traditional method of teaching FOCPNN were based on the number of calculation distances, computing of which is the most consuming part of method. Moreover, n_1^{max} – the maximum number of iterations of the first training phase, n_2^{max} – the maximum number of iterations of the second training phase, $N^{(1)}$ – the number of neurons in the hidden layer, P – the power of the training set.

Discussion. The traditional FOCPNN learning method does not provide support for batch mode, which increases computational complexity (Table 1). Proposed method eliminates this flaw and allows for approximate increase of learning rate in $PN^{(1)}$.

Conclusion

1. The urgent task of increasing the effectiveness of audit in the context of large volumes of analyzed data and limited verification time was solved by automating the formation of generalized features of audit sets and their mapping by means of a forward-only counterpropagating neural network.

2. For increased learning rate of forward-only counterpropagating neural network, was developed a method based on the k-means rule for training in batch mode. The proposed method provides: approximately increase learning rate in $PN^{(1)}$, where $N^{(1)}$ is the number of neurons in the hidden layer and $P$is the power of the learning set.

3. Created a learning algorithm based on k-means, intended for implementation on a GPU using CUDA technology.

4. The proposed method of training based on the k-means rule can be used to intellectualize the DSS audit.

Prospects for further research is the study of the proposed method for a wide class of artificial intelligence tasks, as well as the creation of a method for mapping audit features to solve audit problems.

References:

1. T.V. Neskorodieva. Postanovka elementarnykh zadach audytu peredumovy polozhen bukhhalterskoho obliku v informatsiinii tekhnolohii systemy pidtrymky rishen (Formulation of elementary tasks of audit subsystems of accounting provisions precondition IT DSS). Modern information systems. 3(1), 48-54, 2019. doi:10.20998/2522-9052.2019.1.08

2. A. Fischer, C. Igel. Training Restricted Boltzmann Machines: An Introduction Pattern Recognition, vol. 47. pp. 25-39, 2014.

[1] Yurii Samokhvalov
Doctor of Technical Sciences, Professor, Professor of the Department of Intellectual Technologies.
[2] Bohdan Bondarenko
Master student of the Intellectual Technologies Department
[1, 2] Taras Shevchenko National University of Kyiv

USE OF NEURAL NETWORKS IN INFORMATION RETRIEVAL SYSTEMS

Abstract. The issues of using neural networks in modern information retrieval systems are considered. A comparison of classical methods of information retrieval and neural networks is performed. The efficiency of neural networks application for information retrieval is substantiated.

Keywords: information retrieval systems, neural networks, information retrieval, DSSM network, semantic search.

Machine learning plays an important role in many modern information retrieval systems. Recently, such systems have begun to use deep learning. The rapid pace of modern deep learning research has given rise to many different approaches in information retrieval. The use of neural networks has greatly increased the efficiency of information retrieval systems.

Many experts, such as Tom Kenter, Christophe Van Gysel and others, have dealt with the use of neural networks for information retrieval. A large number of international conferences, seminars and round tables are devoted to this topic. However, the use of neural networks in information retrieval systems needs further development.

Information retrieval systems evaluated the document's relevance to the query based on word matches. As the amount of information increased, there was a need to sort it by relevance. Classic algorithms for estimating the importance of a word in the context of a document, such as TF-IDF and Okapi BM25, have appeared [1].

Classical methods of information retrieval are able to study complex relationships in text, but they use only a fixed set of features designed for search queries and documents [1]. This shortcoming is absent in neural networks, the deep hierarchical structure of which allows to analyze data and generate connections without human intervention.

Neural networks are actively used in image search. Using a network with a specially constructed architecture, image is converted into a vector in N-dimensional space. The query, which can be either text or image, is also converted to a vector. To calculate relevance, the two resulting vectors are compared with each other. The closer one vector is to another, the more the image matches the query.

Using neural networks to search for text is a complex and resource-intensive process, but it justifies itself by showing higher efficiency than with classical methods. The network must find hidden semantic relations to process the request for documents at the semantic level, while keyword-based comparisons often fail. It has to represent

the text of the request and the text of the document title in the form of vectors, scalar multiplication of which would reflect the relevance of the document to the request. To do this, the network has to generate similar vectors for semantically close texts.

A significant breakthrough in the use of neural networks in information retrieval was made by experts from Microsoft Research. They introduced their network for modeling semantic similarity between two text strings, which was called DSSM (Deep Semantic Similarity Model) [2]. The text of the request and the document title are fed to the model input. They are divided into trigrams. This allows representing the text as a vector of several thousand elements that have a value of 0 or 1. Trigrams present in the input text take the value of 1, otherwise - 0. The input vectors are processed by three following hidden layers that have a size of 300 - 300 - 128 neurons respectively. Thus, hidden layers convert the input vector into a vector of hidden semantic features [3]. The output of the model is the result of scalar multiplication of these vectors for the query and the title of the document. The goal of training in this case is to maximize the output value of the similarity function.

The results of the DSSM neural network were compared with other methods of information retrieval according to the NDCG metric, which is used to measure the quality of data ordering in search systems. The neural network has showed the best result in this comparison [2]. At the same time, the performance of the network with such architecture can be significantly improved, for example, by adding to the input layer not only trigrams, but whole words and phrases.

Besides DSSM, other neural network architectures exist and are being developed to search for semantic relations in a text, for instance, MT-DNN, UniLM, and others [4]. An improved version of DSSM is used in various search engines and shows high performance results.

To sum up, the ability of neural networks to find deep semantic connections allows us to build high quality data models. Deep neural networks can learn to process images and are being used in image search. Simple similarity functions, such as the cosine of similarity, can be applied to data generated by neural networks to detect semantically similar words, sentences, paragraphs, and so on. In addition, neural networks can optimize an information retrieval system.

References:

1. Thiago Akio Nakamura, Pedro H. Calais, Davi de Castro Reis, André Paim Lemos. An anatomy for neural search engines. Information Sciences, Elsevier, vol. 480. April 2019

2. Po-Sen Huang, Xiaodong He, Jianfeng Gao. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. At ACM International CIKM. October 2013.

3. Nguyen, Gia-Hung and Tamine, Lynda and Soulier, Laure and Souf, Nathalie. Toward a Deep Neural Approach for Knowledge-Based IR. 2016.

4. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Jianfeng Gao. Unified Language Model Pre-training for Natural Language Understanding and Generation. NeurIPS 2019. 2019.

[1] **Serhiy Semerikov**
Doctor in pedagogy, Professor (Full)
[2] **Hanna Kucherova**
Doctor in Economics, Professor
[3] **Vita Los**
PhD in Economical Sciences, Associate professor
[4] **Dmytro Ocheretin**
PhD in Economical Sciences, Associate professor
[1] *Kryvyi Rih State Pedagogical University*
[2] *Classic Private University*
[3,4] *Zaporizhzhia National University*

NEURAL NETWORK ANALYTICS AND FORECASTING THE COUNTRY'S BUSINESS CLIMATE IN CONDITIONS OF THE CORONAVIRUS DISEASE (COVID-19)

Abstract. The paper proposes an approach to modeling the business climate of the country, which is based on the principles of information transparency, and makes it possible to assess the development trends of the studied indicator in conditions of the COVID-19. This approach has been tested on the example of Ukraine. The results obtained make it possible to analyze the cyclical development of the country's economy with high accuracy and reliability even under quarantine restrictions.

Keywords: Business climate, Business confidence index, Correlation analysis, Socio-economic indicators, Taxonomic model, Neural network model, COVID-19.

The dynamism of changes in the business climate of the countries of the world is accompanied by the increasing uncertainty of the external environment and internal disturbances of socio-economic systems. This is a reaction to new conditions of functioning and development, the emergence of which is due to the global pandemic and quarantine restrictions. The sensitivity of the business climate to such changes is high, therefore, the trends in the indicators that characterize it require system monitoring, thorough and multidimensional data analysis, and increased forecasting accuracy without time delay. This ensures that proactive management decisions are made on time in the context of the impact of COVID-19, which determines the goal and the task of this research.

One of the key indicators that determine the country's business climate is the business confidence index (BCI). The assessment of the business climate is based on the results of generalizing the opinions of business entities about their expectations of the dynamics of changes in production, demand, reserves, the general socio-economic state in the country. Therefore, the results of surveys of business entities, which underlie the formation of the BCI, determine the subjectivity, vagueness, and poor structuredness of the constructed index, which well-known researchers are trying to overcome. Despite the obvious subjectivity of the methodic approach to assessing the business climate of countries, scientists have repeatedly proved the close relationship of

the series of its values with the dynamics of macroeconomic indicators.

To solve the problem of predicting trends in the business climate of countries as a tool for strategic analysis, a wide range of forecasting tools is actively used. The paper proposes an approach to modeling the business climate of the country, which is based on the principles of information transparency, and makes it possible to assess the development trends of the studied indicator in conditions of the COVID-19.

The authors' previous research was based on statistical methods, however, the popularity and efficiency of neural network technologies proved the expediency of their application to solving problems of forecasting the country's business climate. The authors proposed to predict the business confidence index (BCI) using a methodological approach, which includes the step-by-step construction of taxonomic and neural network models.

As a result of using the methodological approach, a time series of quarterly values of the business confidence index in Ukraine was predicted for the period 2008-2020. The forecast was based on socio-economic indicators selected by their closeness to the business confidence index, namely: Retail sales, Industrial production, Steel production, Export, Imports and GDP annual growth rate. The forecast value of the composite index of business activity is obtained as follows:

$$\begin{aligned}\overline{BCI_i} &= W_1 \cdot RS_i + W_2 \cdot IP_i + W_3 \cdot SP_i + W_4 \times \\ &\times Exports_i + W_5 \cdot \mathrm{Im}\ ports_i + W_6 \times \\ &\times GDP\_AGR_i = 0,218 \cdot RS_i + 0,176 \times \\ &\times IP_i + 0,128 \cdot SP_i + 0,096 \cdot Exports_i + \\ &+ 0,096 \cdot \mathrm{Im}\ ports_i + 0,286 \cdot GDP\_AGR_i\end{aligned} \quad (1)$$

The quarterly values of socio-economic indicators for the past thirteen years (2008-2020) were taken as input data. The results of taxonomic analysis established that the GDP annual growth rate and retail sales have the greatest impact on the business confidence index. A forecast has been built for the trend of changes in the business confidence index (forecast accuracy of 89.38%), which proves the similarity of development trends in the country's business climate.

In addition, the most important thing is that the tendency of the studied indicators is identical, in particular, during the period of the emergence of crisis phenomena (beginning of 2009, end of 2014, beginning of 2015, period of the COVID-19 in second quarter of 2020), the decrease in the level of indicators is similar, which suggests that there is a real possibility of using the alternative to business confidence index, which calculated by the taxonomic method of in order to predict the business climate in conditions of limited information transparency.

Having determined the predicted value of the business confidence index (BCI) using a taxonomic model in accordance with the proposed methodology, the next step is forecasting using neural network technologies. An artificial neural network consists of one hidden layer, which contains two neurons, and one output layer (business confidence index). The number of variables in the input layer corresponds to the

number of selected economic indicators for modelling, i.e. six. Thus, to predict business confidence index, used the neural network of the type [6–2–1].

The activation function of the hidden layer is the sigmoid function. This type of function is often used for modeling and the outgoing values of such a function continuously fill the range from 0 to 1. The learning algorithm is the back-propagation error algorithm (Back-Propagation) with a learning rate of 0.1. The difference between the reference and the real output of the network is less than 0.05 (learning rate). The number of learning iterations is 10000.

Formation and analyzing a neural network model were carried out on the basis of the analytical platform Deductor Studio Academic 5.3, which allows you to perform all the steps of data mining from their loading and visualization to building and evaluating the quality of finished models. The time period for analysis is 50 values (first quarter of 2008 - second quarter of 2020). The training set consists of 88% of the data (44 values, time period between first quarter of 2008 and fourth quarter of 2018), and the test set – 12% of data (6 values, time period between first quarter of 2019 and second quarter of 2020).

The constructed neural network model with training capabilities showed the best results in the accuracy and quality of the forecast (forecast accuracy of 96.22%). A decrease in the business confidence index is predicted in third quarter 2020 (will be 87.65). The sharp decrease in the dynamics of the indicator in the studied forecast period is also explained by the influence of the negative consequences of COVID-19 and the introduction of quarantine restrictions in the country and the world.

The article examines the risks of deteriorating the business climate in Ukraine, as a result of which such preconditions as: the weakness of the judicial system, corruption, political and economic instability, the growth of tax pressure, changes in legislation, the slowdown and curtailment of reforms are identified. The situation due to the introduction of prolonged restrictive measures due to COVID-19 was worsened. Insufficient attention has been established in Ukraine to the issues of the ecological system's influence on the formation of the country's business climate, which requires a separate research.

The results obtained make it possible to analyze the cyclical development of the country's economy with high accuracy and reliability even under quarantine restrictions.

The effectiveness of the proposed alternative approach is manifested in saving costs for generating input data for assessing the country's business climate by using official statistics instead of survey results, the subjectivity of which is much higher. In general, the implemented alternative approach is unified, can serve as the basis for further deepening the methodological provisions for studying the business climate of countries with high accuracy and reliability of the results. The prospect of the research is to determine the impact of COVID-19 and introduction of quarantine restrictions on the value and dynamics of the business climate in other countries.

The problem for the implementation of an alternative approach remains limited access to key statistics, which is the result of a policy of ensuring information transparency in different countries.

References:

1. S. Arslankaya, V. Öz. Sakarya, Time Series Analysis on Sales Quantity in an Automotive Company and Estimation by Artificial Neural Networks. University Journal of Science 22, 1482-1492 (2018). doi: 10.16984/saufenbilder.456518.

2. D. Ocheretin, V. Los, H. Kucherova, O. Bilska, An alternative approach to modeling the country's business climate in conditions of limited information. E3SWC 166 (2020): 13024. URL: https://www.e3s-conferences.org/articles/e3sconf/abs/2020/26/e3sconf_icsf2020_13024/e3sconf_icsf2020_13024.html.

3. L.A. El'shin, Mechanisms for the identification of business cycles of regional economic systems based on cross-correlation analysis. Regional Economics: Theory and Practice 15(8), 1540-1551 (2017). doi: 10.24891/re.15.8.1540.

4. V. Los, D. Ocheretin, H. Kucherova, O. Bilska, Neural network technology forecasting the country's business climate, in: Hryhoruk, P., Khrushch, N. (eds.), Proceedings of the 6th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019) 95, pp. 320-324. Atlantis Press (2019). DOI: 10.2991/smtesm-19.2019.62.

5. M. R. Safiullin, L.A. El'shin, A.I. Shakirova, Evaluation of business and economic activity as a short-term forecasting tool. Herald of the Russian Academy of Sciences 82(4), 623-627 (2012). doi: 10.1134/S1019331612040053

6. S. Feuerriegela, J. Gordon, News-based forecasts of macroeconomic indicators: a semantic path model for interpretable predictions. European Journal of Operational Research 272(1), 162-175 (2019). doi: 10.1016/j.ejor.2018.05.068.

7. H. F. Mendonca, A. F. G. Almeida, Importance of credibility for business confidence: evidence from an emerging economy. Empirical Economics (2018). doi: 10.1007/s00181-018-1533-5.

8. H. Sakaji, R. Kuramoto, H. Matsushima, K. Izumi, T. Shimada, K. Sunakawa, Financial Text Data Analytics Framework for Business Confidence Indices and Inter-Industry Relations, in: Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pp. 40-46. Macao, China (2019).

9. V. Los, D. Ocheretin, Construction of business confidence index based on a system of economic indicators, in: Semerikov, S., Soloviev, V., Kibalnyk, L., Chernyak, O., Danylchuk, H. (eds.) SHS Web of Conference. The 8th International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2 2019), vol.65, pp. 1-6. SHS Web of Conferences (2019). doi: 10.1051/shsconf/20196506003.

Marianna Sharkadi
PhD, associated professor
Uzhhorod National University

NEURO-FUZZY MODELING OF LEVEL ASSESSMENT IN THE SYSTEM OF FINANCIAL-ECONOMIC SECURITY

Abstract. The solution of the actual problem of determining the level of financial-economic security for companies through the prism of neuro-phase modeling is presented. In this study, it is proposed to use a multilayer neural network, each layer of which solves a number of problems. The proposed approach will make it possible to determine the level of financial security of the company at different times of its operation. The developed model allows each company to use its own set of financial indicators to determine the level of security. Each layer of the neural network is an autonomous unit that allows you to develop a network.

Keywords: security level, neural network, fuzzy modeling.

Introduction

The use of information technology in various fields of human activity is accompanied by the development of intelligent systems that use the connection of knowledge in the general case with the outside world. The solution to any problem is related to specific subject areas, which are usually badly or poorly structured. During the design and development of an intelligent system, knowledge undergoes a similar transformation of data - from more generalized sets to narrower, specific to a given subject area. In the development of intelligent systems, knowledge of the specific subject area for which the system is developed is rarely complete and reliable.

One of the most promising and active areas of applied research in the field of management and decision-making in poorly structured systems is fuzzy modeling. The fuzzy modeling methodology specifies the methodology of system modeling in relation to the process of construction and application of fuzzy models of complex systems. Every year, the range of fuzzy models and methods expands to cover various new areas. The essence of fuzzy mathematical modeling is that the elements of the study are not numbers, but some fuzzy sets or combinations thereof. At the heart of this approach is not traditional logic, but logic with fuzzy truth, fuzzy connections and fuzzy inference rules.

A significant number of important problems in supporting management decisions that arise in various areas of human activity, is reduced to the task of assessing various kinds of phenomena and processes. When designing and managing a complex socio-economic system, a problem arises when a person is unable to give accurate and then practical values of judgments about their behavior.

The existence of any state in today's globalized world depends on its economic security, which is one of the important components of national security as a whole. One of the main segments of economic security, which significantly affects its level, is the financial segment, which is the set of financial indicators of the economic entity, which are combined into a global indicator. Forecasting this indicator is a complex analytical

and computational process and requires a detailed study of development trends and prediction of the impact of the studied factor's components on the level of the state's economic security.

The urgency of the work lies in the development and study of models and methods for obtaining multicriteria assessment using neuro-fuzzy technologies, which is currently undisclosed sufficiently.

Formulation of the problem

For an economic entity, economic security is considered as a state in which its economic development and stable activity is ensured, the protection of its financial and material resources is guaranteed. Ensuring financial security involves planning, forecasting and anticipating many factors of the internal and external environment. At the same time, a systematic, comprehensive approach based on the effective use of appropriate information and analytical support, logic and modeling to involve the modern mathematical apparatus is extremely important.

The general statement of the problem (task) can be presented as follows. Let a set of quantitative and qualitative indicators of its functioning be known for a certain subject of economic management, as well as the history of these indicators for certain periods of time is known. There is a task to provide an assessment of the entity's economic security level.

We formulate the statement of the evaluation problem as follows. Suppose we have at the entrance some object of study, which is evaluated by many indicators $K=(K_1,K_2,...,K_m)$. Indicators K can be a whole system of criteria and models. Each indicator is a quantitative estimate, which can be obtained, for example, using financial reporting models [1].

Based on a set of estimates $K=(K_1,K_2,...,K_m)$, it is necessary to establish the level of financial security of the object.

To solve the formulated problem , a model of a neuro-fuzzy network is proposed (Figure 1), which consists of a set of successive layers, each of which solves a number of specific classes of problems[2].

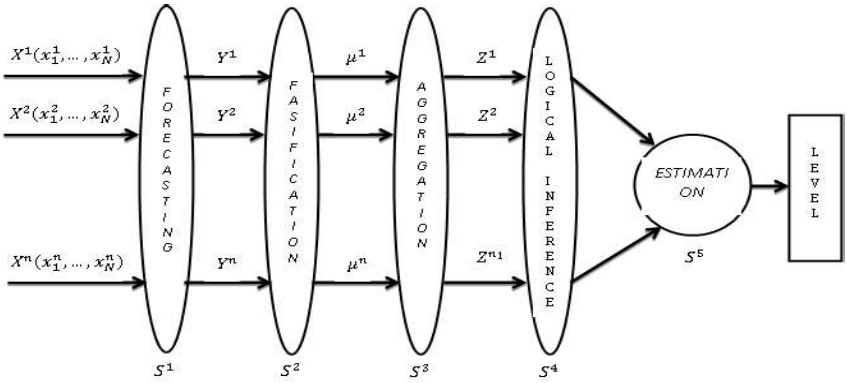

Figure1: The structure of the neuro-fuzzy network.

Conclusions

A study of the current task of determining the assessment of the level of economic security for economic entities of the socio-economic system and the state as a whole, taking into account key management indicators. The following results were obtained for the first time.

The structure of a multilayer neural artificial network with a fuzzy mathematical model for determining the security level assessment in the financial and economic system is proposed. On the basis of this network the information technology is developed, which allows to forecast step by step on the basis of input statistical information on financial indicators the values of these indicators for future periods, to determine their efficiency through membership functions, to obtain aggregated fuzzy estimates on certain groups of criteria [3]. Conclusion to obtain a fuzzy integrated assessment of the safety level, which with the help of defasification methods reduce to a clear value.

References

1. Walsh Karyan. Key Management Indicators: A Complete Guide to Working with Critical Numbers Managing Your Business, 4th ed., Companion Group, Kyiv, K, 2010.

2. Yu. P. Zaichenko. Fuzzy models and methods in intelligent systems [Text]: textbook. Manual, Slovo, Kyiv, K, 2008.

3. M.M. Malyar. Models and methods of multicriteria limited-rational choice: Monograph, RA "OUTDOOR-SHARK", Uzhhorod, 2016.

[1] **Petro Soroka**
PhD in Physical and Mathematical Sciences, Associate Professor,
Associate Professor of the Intellectual Technologies Department
[2] **Serhii Krasnovidov**
Master student of the Intellectual Technologies Department
[1,2] *Taras Shevchenko National University of Kyiv*

BUSINESS ANALYTICS INFORMATION TECHNOLOGIES FOR ANALYSIS OF THE ACTIVITY OF A COMMERCIAL ORGANIZATION

Abstract. This paper considers modern business analytics system approaches for analysis of a commercial organization's activity, including data mining and market basket analysis.

Keywords: business analysis, business analytics, data mining, market basket analysis.

Current economic situation, given the global crisis, does not allow large and small enterprises to operate at full capacity. In such conditions, it's important for each commercial organization to maintain its profitability and competitiveness in the market. Business analytics systems help to cope with this task allowing managers to make informed decisions reducing significant amount of their time in finding and analyzing the necessary information.

Nowadays, one of the most effective tools for managing a company is business analysis. It allows to get a full view of your organization and the prospects for its development. Business analysis is a set of tasks and techniques used both to understand the activities of the organization as a whole and to obtain effective management decisions [1].

In our time, the amount of data has reached such a mark that even a group of people is not able to analyze them on their own. However, large raw data sets often contain knowledge that can be used in decision making. For their processing and analysis Data Mining technology is used extracting previously unknown, non-trivial, practically useful and interpretable knowledge necessary for decision-making [2].

Consider some methods of Data Mining: sequence, classification, regression, association. The sequence is to find a temporal pattern between events, i.e. such a relationship that if event X occurs, then after some time event Y will occur too. For example, after purchasing a car, the driver will most likely take out an insurance policy, then buy a first aid kit, spare wheel etc.

Classification helps to identify features that define a group of certain objects. The purpose of regression is to find a function of given variables that would determine the range of valid values. A prerequisite for such an analysis is the relationship between variables [3].

The task of the association is to identify patterns between related events, i.e. rules of the type «event Y follows event X», or association rules. Analysis of purchases and goods sold together is often called market basket analysis.

Market basket analysis focuses on the study of a large set of information in search of trends, patterns and relationships that contribute to effective decision making [4]. A market basket is a set of goods purchased by a customer. The most popular algorithm that solves this problem is still the Apriori algorithm, which has been repeatedly improved by various researchers. There are now a large number of software products and commercial technologies, such as Oracle Data Miner, Tableau and Deductor Studio, that effectively solve this problem.

Deductor Academic software was used in the research to form a data warehouse, analyze the activities of a commercial organization – a shop selling auto parts, and generate its annual reports. Initially, the data cleansing removed the records of accidental transactions (there were about 20 such records) and completed some records replacing 'auto part code' nulls with a position code from the same department of similar price. After that, an OLAP cube was formed by product groups. The products that bring 80% of the profits are automotive oils, automotive chemicals, cosmetics and small spare parts. Friday was found as the busiest day of week in the store.

To identify sales increasing techniques, there were also found the association rules of the most popular pairs of goods pre-configured with a minimum support of 10% and a minimum probability of 80%. As a result, 4 rules were found: 2 of them had already been implemented in the form of respective departments' planning (trivial rules), the other 2 were recommended to the store administration.

Conclusions. Modern business analysis technologies allow a commercial organization to consolidate knowledge about its activities, make OLAP-slices of sales, manage store load and design its own data warehouses. Association rules allow to determine its most popular sets of goods and see the options of redistribution of less popular goods. Therefore, based on the analysis, the company can optimize its business processes, which will lead to increasing sales.

References

1. Franks, B. (2014). *Taming The Big Data Tidal Wave*. Hoboken, NJ: John Wiley & Sons, Inc.

2. BABOK Guide 2.0. Retrieved from: https://www.iiba.org

3. Rouaud, M. (2013). *Probability, Statistics and Estimation (Short Edition)*. Retrieved from: http://www.incertitudes.fr/book.pdf

4. Babin D. V. (2004). *Genetic algorithm for solving the problem of market basket analysis* (Institute for Artificial Intelligence Problems). Retrieved from http://www.iai.dn.ua/public/JournalAI_2006_4/Razdel3/06_Babin.pdf

[1] **Petro Soroka**
PhD in Physical and Mathematical Sciences, Associate Professor,
Associate Professor of the Intellectual Technologies Department
[2] **Roman Savchenko**
Master student of the Intellectual Technologies Department
[1,2] *Taras Shevchenko National University of Kyiv*

MACHINE LEARNING METHODS FOR SPORT RESULT PREDICTION

Abstract. The purpose of this paper is to consider various approaches of utilizing machine learning methods for sport result prediction.

Keywords: machine learning, API, betting market, hierarchical scoring structure, match, set, game, point.

As a result of exponentially growing computing capabilities and data availability machine learning now is being applied to almost all areas of life, from financial institutions and medicine to self-driving vehicles. At the same time, its utilizing for sport result prediction and the related betting market is given relatively less attention. More traditional statistical approaches still dominate this area. In addition, one of the main areas of investigation has been the football market, and tennis has been in the background, although it is one of the most popular sports. The potential profit, as well as scientific interest, encourages to search for effective methods of predicting result of tennis matches.

Most modern approaches to tennis prediction use a hierarchical scoring structure to determine the probability that a player would win a match. Assuming the points have independent uniform distribution, the expressions only need the probability that two players will win a point on their own serve. From these basic statistics, it becomes possible to deduce the probability that a player would win a game, then a set, and finally a match.

Barnett [1], O'malley [2], and Knottenbelt [3] in their researches defined hierarchical models for calculating the probability of winning a point on serve using only matches with players' common opponents instead of all previous opponents. This reduces bias due to differences in the level of opponents. Madurskaya [4] improved the model by using different winning probabilities in different sets and allowing the model to reflect how a player's performance changes during a match.

This mathematical approach represents the level of players using only one value (points won), it does not take into account a large number of extremely important parameters. For example, a player's susceptibility to a particular opponent's game strategy, time after the last injury or fatigue accumulated from a previous match. In addition, there are such important characteristics of the match itself as location,

weather, surface, etc. Given the huge amount of diverse historical data, an alternative approach to predicting the results of tennis matches can be based on machine learning. Player parameters and features of the match itself, combined with its outcome, can form a dataset for machine learning application. Supervised learning algorithms can be applied to determine the function of predicting the results of future matches.

The paper examines not common approach to predicting tennis matches - predicting the amount of total games played in the second set based on the result of the first set. For this purpose, a set of historical data was gathered and pre-processed. The goal was to classify matches into two classes: with the total of games in the second set less than 9.5 or more than 9.5 games. To solve it, the following machine learning methods were used: k-nearest neighbors, support vector machine, multilayer perceptron. The return on investment when tested on real data ranged from -17.16% to 18.67%, which is close to the results of existing researches.

For the convenience of applying prediction result in practice an automatic notification system was developed. Console app based on C#.NET in combination with the Telegram API allows to collect information about current matches from open resources in real time, analyze it and send notifications about potentially profitable matches to the Telegram messenger.

Conclusion. In this paper, machine learning methods were applied to predicting the result of tennis matches. Also were found ways to improve result by using more complete and informative data for training as well as optimizing the configuration of the neural network. This indicates the significant potential of machine learning for predicting the results of sport competitions.

References

1. T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics, 16:113–120, 2005.

2. J. A. O'Malley. Probability Formulas and Statistical Analysis in Tennis. Journal of Quantitative Analysis in Sports, 4(2), 2008.

3. W. J. Knottenbelt, D. Spanias, and A. M. Madurska. A common-opponent stochastic model for predicting the outcome of professional tennis matches. Computers and Mathematics with Applications, 64:3820–3827, 2012.

4. A. M. Madurska. A Set-By-Set Analysis Method for Predicting the Outcome of Professional Singles Tennis Matches. Technical report, Imperial College London, London, 2012.

[1] **Bohdan Sus**
PhD, Assist. Prof. of Nanophysics of Condensed Matter Dept.
[2] **Ilona Revenchuk**
PhD, Assoc. Prof. of Software Engineering Dept.
[3] **Oleksandr Bauzha**
PhD, Assoc. Prof. of Computer Engineering Dept.
[1,3] Taras Shevchenko National University of Kyiv
[2] Kharkiv National University of Radio Electronics

MODEL OF IMPLEMENTATION VIRTUAL LABORATORY WORK FOR SUPPORTING EDUCATIONAL PROCESS

Abstract. Due with the current pandemic, online education is becoming very popular and it is one of the good ways to get access to educational resources without a threat to health. The article discusses Model of Implementation Online Educational Systems for Supporting Educational Process at the organization (schools, universities etc.). E-learning laboratory work are mainly remote and should be integrated with Learning Management System for more effective learning process.

Keywords: Learning Management System, E-learning, Virtual Laboratory Work

The proliferation of new technologies and internet tools is fundamentally changing the way we live and work. The lifelong learning sector is no exception with technology having a major impact on teaching and learning. This in turn is affecting the skills needs of the learning delivery workforce. From Moodle to Edmodo to Inquisiqr3, there's a lot of tools we can use to manage, track, and deliver educational courses and training programs in our universities.

All Supporting Technology that could be used to deliver educational courses and programs could be classified to several groups by their role in the educational process:

- Online Education Systems - OES (Online Education, E-learning, OES, Integrated OES, Standards Specifications);
- Content Creation Tools -CCT (CCT, Authoring Tools, Assessment Tools, Learning Content Management Systems (LCMS), Learning Objects);
- Learning Management System - LMS (LMS, Learning Platform, Virtual Learning Environment (VLE), Learning Service, Provider (LSP));
- Management System (MS) (Student Management System, Enterprise Resource Planning System, Human Resource Information System, Knowledge Management System, Competency Management System).

There are many terms for online education. Some of them are: virtual education, Internet based education, web-based education, and education via computer-mediated communication. The Web-edu project uses a definition of online education that is based on [1] definition of distance education. In general, online education is characterized by:

- the separation of teachers and learners which distinguishes it from face-to-face education;

- the influence of an educational organization which distinguishes it from self-study and private tutoring;
- the use of a computer network to present or distribute some educational content;
- the provision of two-way communication via a computer network so that students may benefit from communication with each other, teachers, and staff.

E-learning is here defined [1] as interactive learning in which the learning content is available online and provides automatic feedback to the student's learning activities. Online communication with real people may or may not be included, but the focus of e-learning is usually more on the learning content than on communication between learners and tutors.

Very often the term e-learning is using as a synonym for online education. E-learning includes a set of applications and processes, as instance Web-based learning, computer-based learning, virtual classrooms and laboratory, digital collaboration. It includes the delivery of information or content via Internet.

Therefore online education is covering all systems that support it.

Beside content, integrated learning system - ILS includes a tools for making assessments, notes, report creation, and user files that help to evaluate and identify a monitoring learning progress via students learning needs and competences.

CCT are the tools that course designers and teachers use to create the content in online education courses. The CCT are used to develop learning material (text, slides, graphics, pictures, animations, simulations, assessments, audio, video etc.)

Authoring tools could be regarded as a subset of CCT. A software application, used by non-programmers, that utilizes a metaphor (book, or flow chart) to create on-line courses [2].

The LCMS is a computer application provide procedures to manage workflow in a collaborative in common environment such as for creation text and publishing, editing and modifying content, organizing, deleting as well as maintenance from a central interface. It use for creation teaching contents and materials for educational process in online or as blended learning.

The learning object is a reusable, media-independent chunk of information used as a modular building block for e-learning content. Learning objects are most effective when organized by a meta data classification system and stored in a data repository such as an LCMS.

The content-management systems focused in general more attention to the creation, developing, and managing content for online courses, and less - to the control of students experience.

Institutions use LMS software for support of organization of learning process, namely: to plan educational process, implement it, the evaluation educational process, support of access for the students and teachers, monitor the student learning.

Beside this, software provides course preparation, educational content and resources according to the standards or templates. It centralizes the delivery and tracking of student activities during education, as instance discussion, collaboration, to

do tasks, assessments. All these activities will control by the virtual environment of LMS that provides a user data protection.

Each LMS is different and it give to users different possibilities for implementation such as content-oriented, activity-oriented, network-oriented, linear, and branching. Some systems realize asynchronous instruction, while others are providing synchronous instruction [3].

There are many different types of LMS, or LCMS, for manage learning process and course content delivery. That why one is the important question is how to choose effective LMS or LCMS according to the organization needs. The main rule is to know how LMS will be deliver training materials to students, and then compare your organization needs with LMS opportunities by the LMS functionality.

There are many projects for the development of virtual laboratory work (VLW) in natural science. Laboratory work for cloud electronic learning environments with algorithms and methods for protecting information between devices using combined communication channels and embedded systems is discussed in [4-6].

To develop methods of virtual laboratory work wide use of interactivity with the implementation of interdisciplinary methods to learning is required. In the case of laboratory work, the essence of these approaches is to expand ways of virtual laboratory work on every step, the presence of self-control and methods of evaluation of outcomes. During laboratory work, students have to perform a task of conscious choice and means of experiment process. This may be the choice from the possible list of virtual devices and conditions of the experiment. An adaptive model of the virtual electronic laboratory work is presented on fig.1.

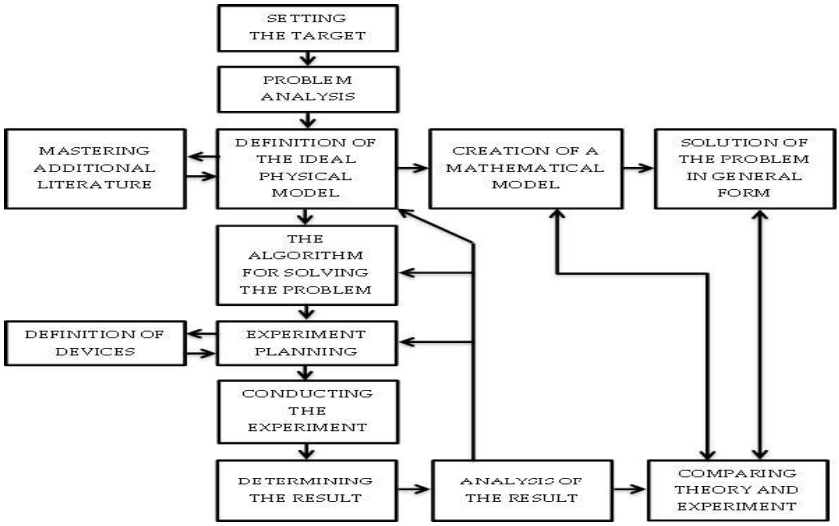

Figure 1 - Block diagram of the algorithm of functioning of electronic laboratory work

Methods of visual intercommunication, methods of modeling and data acquisition have been also realized for virtual laboratory development. Tasks are performed in real-time, with the illustration of complex processes and interactive analysis of additional data.

At any step of the process, the student can investigate the data from the current set of measurements.

Each element of the virtual device can interchange data with other elements of the VLW and external programs by the means of network technologies.

VLW can also be a good tool to get introduced with original laboratory devices with limited access.

Conclusion. VLW as a as custom eLearning implementation and design solution for virtual class are based on real experimental setups. Integration process of Moodle with VLW and variety of other programs with functional together as a system in further very useful to meet various needs of a website for eLearning, including developing courses using authoring tools and make it available through Moodle with the internal course booking system.

References:

1. Desmond J. Concepts: Problems in defining the field of distance education, 2009. URL: https://www.tandfonline.com/doi/abs/10.1080/08923648809526619.

2. D. Sewart, D. Keegan, B. Holmberg, Distance education: International perspectives, 2020. URL: https://books.google.com.ua/books?hl=en&lr=&id=OujyDwAAQBAJ&oi=fnd&pg=PT7&ots=Tl2D99Oz_J&sig=PCh7apz6xYcx4INPJegCFm9eiZE&redir_esc=y#v=onepage&q&f=false

3. LMS Selection Guide 2020. URL: https://mindflash.com/resource-center?postType=literature

4. Tmienova, N., Sus, B.: Hardware data encryption complex based on programmable microcontrollers. In: CEUR Workshop Proceedings, pp. 199–208 (2018). URL:http://ceur-ws.org/Vol-2318/paper17.pdf.

5. Bauzha, O., Sus, B., Zagorodnyuk, S., Stuchynska, N.: Electrocardiogram Measurement Complex Based on Microcontrollers and Wireless Networks. In: International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T, pp. 345-349. (2019)

6. Sus, B., Tmienova, N., Revenchuk, I., Vialkova, V.: Development of Virtual Laboratory Works for Technical and Computer Sciences. In: Damaševičius, R. and Vasiljevienė, G. (eds.) Information and Software Technologies. pp. 383–394. Springer International Publishing, Cham (2019). URL:https://doi.org/10.1007/978-3-030-30275-7_29.

[1]**Nataliia Tmienova**
PhD in Physical and Mathematical Sciences, Associate Professor of the Intelligent Technologies Department, Associate Professor
[2]**Oleksandra Dulich**
Master student of the Intelligent Technologies Department
[1,2] *Taras Shevchenko National University of Kyiv*

AUTOMATIC QUESTION GENERATION SYSTEM FOR UKRAINIAN-LANGUAGE TEXTS

Abstract. The problem of automatic generation of questions for Ukrainian-language texts is considered. The problems that arise when generating questions for Ukrainian-language sentences are presented, the analysis and comparison of the used methods are given. The main result is a software module for automatic question generation.

Keywords: natural language processing, text processing, tokenization, stemming, POS-tagging, automatic question generation, question generation methods, rule-based methods, text corpus.

We live in a time when the amount of information produced by humanity is greater than ever, and the amount of this data is growing every day. However, significant benefits can only be obtained from this information if this data is properly processed and analyzed. On the other hand, the number of tasks that humanity trusts computers to solve is growing at an incredible tempo. More and more processes are being automated. Thus, there are problems with computer analysis and natural language synthesis, and the need to improve natural language processing methods is growing and enjoys inexhaustible interest. This is especially relevant for Ukrainian realities, since, unfortunately, there is a lack of tools for processing the Ukrainian language, such as libraries for programming languages, marked corpora, dictionaries, and thesaurus.

Automatic question generation in natural language processing is one of the most urgent tasks of Computational Linguistics [1]. Systems with similar functionality are usually used in the field of education to test students' knowledge, namely when compiling questions on theoretical material [2]. Thus, the implementation of a system for automatic question generation is a promising area of work and can be used in the development of chatbots, compiling tests for online courses or distance learning, which are now in high demand during the pandemic.

The main problem of automatic generation of questions for Ukrainian-language texts is that before the stage of direct question generation, you need to perform a number of transformations with the input text in order for the computer to understand human speech at first, and only then generate a question. Performing these actions for the English language is easier due to the availability of appropriate tools for processing the English language. Meanwhile for the Ukrainian language, even at the preliminary stage of processing, a few problems appear, because the quality of understanding the language depends on many factors.

Thus, there is a need for pre-processing of the Ukrainian-language text: it is necessary to have the text processed at the semantic and syntactic levels, and for this, it is urgent to implement such tasks of graphematic and morphological analysis as determining the boundaries of sentences, words tokenization, lemmatization or stemming, stop words extraction and parts of speech tagging. It should also be noted that sometimes there is a problem in finding new ways to define key parts of a sentence when well-known algorithms for determining keywords.

Existing tools for preprocessing Ukrainian-language text, such as sentence boundary detection, tokenization, and parts of speech tagging, did not meet the expected results. So, to solve the problem of sentence boundary detection and tokenization, it was decided to resort to the NLTK library [3], namely, tools created for English and Russian, which were partially improved and adapted for the correct processing of Ukrainian-language texts. This library also provides the ability to add Ukrainian stop words, the step of getting rid of which is also quite important when analyzing text.

The morphological analyzer pymorphy2 was used for parts of speech tagging [4]. Work on the tagging ordinary words is performed by pymorphy2 tools based on the Russian-language OpenCorpora corpus, so it is not always possible to get the correct processing result for the Ukrainian language texts.

It was decided to use a text corpus-based generation method for question generation for the implementation of the software module. Unlike other methods, this one has no restrictions on the number of question types and format of text data. It can be divided into 3 stages: text preprocessing, sentence filtering, and question generation itself, which has a rule-based implementation. This approach is based on the development and extension of existing rules, so the system does not require a massive training corpus compared to the machine learning approach.

The result of implementing this method is a software module for automatic question generation, which receives text for input, performs text preprocessing, determines the keywords and key sentences for which questions are generated. The used rule-based approach is flexible and easy to debug. It also does not require a massive training corpus and shows high accuracy. It should be noted that adapting text processing tools of other languages to Ukrainian-language texts raises the possibility of errors during analysis, which is a disadvantage of this module, along with the need to improve and expand the existing rules for generating questions.

References

1. A.M. Kurtasov, A. N. Shvetsov. Program for generating educational tests based on a semantic approach // Proceedings of the International scientific and methodological conference “Informatization of engineering education” - INFORINO-2012. Moscow: publishing house of MEI. P. 71-74 (in Russian).

2. I. L. Bratchikov. Generation of test tasks in expert training systems // Bulletin of the peoples’ friendship University of Russia. Series: Informatization of education. 2012. № 2. P. 47-60 (in Russian).

3. Bird, Steven, Edward Loper and Ewan Klein, Natural Language Processing with Python. O’Reilly Media Inc., 2009.

4. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332, 2015.

[1] **Dmytro Yakymenko**
Postgraduate student (applicant)
[2] **Iryna Tregubenko**
PhD, Associate Professor
[1,2] *Cherkasy State Technological University*

MODIFIED METHOD OF CONSTRUCTION OF INFORMATION IMAGE OF ELECTRONIC TEXT DOCUMENTS BY MEANS OF INTELLECTUAL DATA ANALYSIS

Abstract. The problems of analysis and processing of electronic documents are investigated in the work. The existing methods of analysis of text documents are analyzed. The study identified ways to improve existing methods. The article presents a mathematical model of the method.

Keywords: data mining, Kohonen maps, information retrieval image of documents, TF-IDF, array of values, cluster.

1. Formulation of the problem

The rapid increase in the number of electronic documents that is currently observed clearly shows that traditional mechanisms for processing electronic documents are not able to cope with the needs of users. This trend is noticeable both on the Internet and in corporate networks.

That is why there is a need for methods that will provide a quick and easy distribution of documents by category or by keyword.

Thus, we can identify the main problems associated with increasing the amount of information:

— the rapid growth of information contained on the Internet is the cause of more and more difficulties in finding the necessary documents and organizing them in the form of structured repositories;

— most technologies for working with text documents are focused on the organization of convenient work with information for humans, but there are virtually no opportunities to convey the semantic content of the text, ie there is no semantic indexing;

— unstructured information is a significant part of modern electronic text documents.

Therefore, there is a need to develop new modern methodologies for data processing and analysis. Data mining has become such a new methodologydata mining [1].

Data mining is the processing of information and the identification of patterns and trends that help make decisions. The principles of data mining have been known for many years, but with the advent of large amounts of data, they have become even more widespread. Large amounts of information have led to an explosive increase in the popularity of broader methods of data mining, because information has become much more, and it by its very nature and content is becoming more diverse and

extensive. When working with large data sets, relatively simple and straightforward statistics are no longer enough.

The reasons for the popularity of IAD are as follows:

— rapid accumulation of data (the account is already on exabytes);

— general computerization of business processes;

— penetration of the Internet into all spheres of activity;

— progress in the field of information technologies: improvement of DBMS and data warehouse;

— progress in the field of production technologies: rapid growth of computer productivity, storage volumes [2].

2. Presenting main material

After analyzing the analogue, it was decided that there is a problem in the method that needs to be solved.

The method proposes to modify the method of data clustering, based on the Kohonen method. It is planned to modify the method of selecting the data to be analyzed.

To increase productivity, it is proposed to introduce normalization of data sampling, within the specified limits. This will ensure faster data sampling, as data for analysis will be more carefully selected.

The developed algorithm for forming images of documents is based on a statistical approach to the analysis of texts in natural language. It is proposed to form the image of each document in the form of a multidimensional vector of normalized and weighted single words (signs) found in the text of this document. The dimension of such a vector will be equal to the number of unique features in the document collection.

The proposed method of forming images of PD consists of the following main stages:

$$\Phi D = < \Phi P, \Phi DP, \Phi R > \quad (1)$$

where ΦP - a way to remove features from the texts of documents; FDP - a way to display documents in the space of their features; FR - algorithm for reducing the space of document features [3].

The method of removing the signs of FP is to perform the following operations: lexical analysis of the text (removal of markup, punctuation, numbers, conversion of all letters to uppercase, etc.), removal of stop words, ie commonly used words that do not have an independent meaning , for example, prepositions, conjunctions, particles and pronouns; morphological analysis.

We propose to use such an approach to morphological analysis as the selection of pseudo-bases of words. As a result of this analysis, words from the text are reduced to a special type, and in the future, words that have the same special form (pseudo-basis) are considered as one feature. As a result of extracting features by the FP method, it is possible to obtain NP - a dimensional set of features (pseudo-words) of the document collection P, which is also called the general dictionary of features of the document collection.

The method of displaying documents in the space of their features FDP is based on the procedure of weighing features. Weighing of features of documents is offered to carry out by means of traditional technique tf * idf which is independent of existence of a training set, considers frequency of occurrence of a term, both in the separate document, and in all collection as a whole.

The need to develop an algorithm for reducing the feature space of FR documents is due to the fact that high-dimensional and sparse document vectors in the feature space are not sufficiently clear orientation so that automatic methods by calculating the distance between them could make an unambiguous conclusion about their relationship or difference. To solve this problem in modern information retrieval systems, forced reduction of the feature space by the DF criterion is used. The algorithm of forced reduction according to the DF criterion removes from the general dictionary of features P of the document collection all those features whose document frequency is above the threshold value DF maxτ and below the threshold value DF min τ .

The input of the neural network is fed a set of document vectors in the form of a matrix of size N by M, where N is the number of documents that are clustered, and M is the number of unique terms in the collection of documents that are clustered. At the intersection of columns and rows are the weights of the j-th term in the i-th document, calculated by the method tf * idf.

The basic algorithm for learning the Kohonen network is as follows:

Step 1 Initialize the weight matrix with small random values (on the interval [0, 1]).

Step 2 Randomly select a vector from the source set.

Step 3 For each output neuron j calculate the distance between its weight vector w_i and the output vector x:

$$d_j = \sqrt{\sum_{i=1}^{n}(w_{ij} - x_j)^2} \qquad (2)$$

Step 4 Find the original winning neuron jmin with the minimum distance min (dj).

Step 5 For the original winning neuron jmin and for its neighbors in the vicinity, the weight vectors are updated as a rule:

$$W_{ij}(t+1) = W_{ij}(t) + e(t) * h(t,j,m) * (x_i - W_{ij}(t) \qquad (3)$$

where w_{ij} (t) is the value of the weighting factor of the input neuron i and the output neuron j at time t; h (t, j, m) - the value of the neighborhood function with the central neuron of the source layer m for the neuron of the source layer j at time t; e (t) is the coefficient of learning speed at time t; x_i is the output of the neuron of the first layer numbered i.

Step 6 Repeat the steps from step 2 for all elements of the source set.

The training cycle lasts until the system reaches the desired state. The following can be used as criteria for stopping the learning process:

- topological ordering of the feature map (weight matrix);
- weight changes become insignificant;
- the network output is stabilized, ie the output vectors do not pass between

cluster elements;

- the limit value of the error on the map has been reached;
- passed a given number of epochs.

Conclusions

As a result of the research, a modified method of forming an information retrieval image of an electronic document was developed. The modified method provides faster and better data selection for the formation of IPO. The method of constructing an information retrieval image of electronic documents by means of data mining provides stable and fast work with documents.

In the future it is planned to refine the method and make changes to it. Since this topic is currently in high demand, it is possible to make amendments to the method that will improve and speed up the work of this method.

References:

1. Korneev VV Databases. Intelligent information processing. / Korneev VV, Gareev AF, Vasyutin SV, Reich VV -M .: Nolidzh, 2001. 496 p.

2. Barsstyan AA Data analysis technologies: Data Mining, Visual Mining, OLAP./ Barsgyan AA, Kupriyanov MS, Stepanenko VV, Kholod II - СПб: БХВ-Пстербург, 2007. 275 с.

3. Oksanich IG Intellectual analysis of an array of text documents based on Text Mining technology / IG Oksanich, DM Piskunov, DP Chernysh // Information processing systems. - 2013. - Vip. 2. - P. 139-143.

CYBERSPACE PROTECTION TECHNOLOGIES

[1] **Serhii Buchyk**
Doctor of Technical Sciences, Professor at the Department of Cyber Security and Information Protection
[2] **Yaroslav Andrushchenko**
Student
[1,2] *Taras Shevchenko National University of Kyev*

SEARCHING FOR A POTENTIAL CRIMINAL USING WIRELESS INTERNET NETWORKS AS ONE OF THE TARGETS OF STATE SECURITY

Abstract. The article discusses one of the possible ways of finding potential criminals using wireless Internet networks in order to strengthen the security of facilities of particular importance and the country in general. It is also described how modern digital and network technologies can be used to develop a comprehensive search system and control a person's cell phone.

Keywords: wireless network security, 'man-in-the-middle', cybercriminals, cyber special service, data interception, deauthers.

In the context of the hybrid war, in which Ukraine is currently engaged, the question of ensuring security and integrity of the country is beyond doubt. Many countries, for their territorial integrity, are investing not only in the armed forces, in the traditional sense of this term, but also in cyber-weapons. The advent of cyber-weapons also makes it necessary to develop means of countering these weapons. That is why governments fund and support projects that protect the country in cyberspace.

Detection and prevention are the main objectives of the country's special services. That is why they must use modern equipment and technology, by means of which, they can monitor and, if necessary, influence the work of information systems or individual devices. But how to set up this solution? And what would that take? It's impossible to track every single person in the country. People generate a lot of information, and there are physically not enough specialists to process that information. And furthermore, the Constitution protects the human rights of privacy. Special services have the right to interfere in the life of a person only if suspicion has been made and a court order has been executed. And only after that you can legally interfere with a certain person's life.

You can't defend yourself if you can't attack. That is why it is not uncommon for special services to use the help of cybercriminals to solve cybercrime. They also train their specialists, the so-called "white hackers" who work for the country. However, such specialists are not of much use without special tools. The authorities seek help from developers who create tools to search for and deanonymize criminals.

One of those tools that could help special services is a tool that can scan mobile phones with an enabled Wi-Fi adapter. This type of solution should consist of two parts. The first part is a server that processes, stores and transmits information, and the second part is an access point that emits the work of a Wi-Fi router. Such Wi-Fi points are placed in strategically important and lively places such as airports, bus stations,

train stations and so on. They're also placed in shopping and entertainment centers - all the places where there are a lot of people. In such public places, there is usually free Wi-Fi that people gladly use. Also, these places are visited by criminals, as they can be "invisible" among the crowd. And it is this solution that would allow us to detect the presence of such attackers.

As mentioned earlier, the solution should consist of a server and an access point. However, there should be two access points, so that a person can be found by x and y coordinates. Each point runs at frequencies of 2.4 Ghz and 5.0 Ghz. These are the open frequencies on which the Wi-Fi protocol works according to IEEE 802.11[1] standard. These points are different from others because they have a much higher power, their signal power reaches several kilometers. The points cover 360 degrees around, but their antennas have a clear direction, which is why they have to be positioned in the direction of these antennas. Otherwise, their efficiency will be reduced. These antennas must be reprogrammed with special software that is based on the Linux kernel with long-term support distributions such as Debian, or its descendant - Ubuntu. However, the specific user (operator) "shell" should be developed which is oriented to the Cisco command-line interface (CLI). These access points should be connected to the Internet and directly to the mini-server on which the scanned information will be stored and the database will be supplemented.

Next up is the server. There should be two types of server. The first one is a cell phone - a small computer that will be connected directly to the point which stores, processes and transmits the information to the data center, where a more powerful server, operated by the information security administrator will already be. The mobile server will also update and configure the access point.

For easy operation, a graphical interface (GUI), operated by a cybersecurity specialist, should be developed. With this interface, the operator should be able to process all incoming information. Scanning information should be displayed on his/her monitor and have certain filters to make the work easier. He/she should also be able to control, configure and reconfigure the version of the access point. In case of sensor problems, the operator should be able to connect to it directly, and with debugging commands that are specially designed to test the work of the sensor, reconfigure it or gather information which will be forwarded to the developers for further software correction and updating.

The question then is: What is this all about? With such a solution, the operator will monitor and, if necessary, conduct an attack on the mobile phone in order to capture the attacker's phone. In that case, two phases of intervention would be carried out. The first one is known as the Man-in-the-Middle attack[2]. The essence of the attack is that it interferes with the transmission of information in such a way, that from the target victim the information passes through the attacker and then to the router and vice versa. When the victim requests the information, the router first directs it to the attacker because it considers him to be the end-user and then the attacker redirects it to the user.

This is a description of the classical method, but the solution will use a more effective method. The information, that is most often transmitted, is encrypted and

cannot be retrieved. In case of using this solution, the public router, which is located in the mall, will be replaced by the router of the cybersecurity operator. This all will be done thanks to these sensors. That is why these access points will work exclusively with the cell phone of the criminal and not with the cell phone of ordinary citizens. The operator chooses his target; the system first calculates the frequency at which it operates - 2.4Ghz or 5.0Ghz, then finds the exact channel using the phone. The sensor then sends deauthers[4]. Deauthers are images that are sent to the target's phone and report that it has been disabled. At this time, the point is fixed to a specific channel at a certain frequency and creates an access point with an exact copy of the SSID[3]. When the criminal's phone tries to reconnect the Wi-Fi, the access point is no longer public, but artificially created by the system. The victim won't notice any change in work, because it happens pretty quickly. And once the target is caught, we can analyze and alter the traffic. That's how phase one ends. Next, phase number two - infection of the target's phone and retrieval of information from the device - begins. The solution starts sending packets and switching the phone so that it can "tell" about itself. This enables the installation of malicious software for spying purposes, the ability to redirect the phone to phishing sites in order to obtain confidential data or download programs.

All actions described above are illegal and can be carried out exclusively by special services, the ones that have the permission of the court. The operator's target won't be random. The MAC address of the target's phone must be approved by the court, and permission for obtaining the information from the target phone must be granted. Otherwise, all obtained information in the course of an investigation cannot be considered in court, because it has been obtained illegally.

The theoretical material discussed above is all about improving cybercrime investigations and the work of the special services of the country.

References:

1. IEEE 802.11-2012 - IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications

2. Man in the middle (MITM) attack. - https://www.imperva.com/learn/application-security/man-in-the-middle-attack-mitm/

3. What Is an SSID, or Service Set Identifier? - https://www.howtogeek.com/334935/what-is-an-ssid-or-service-set-identifier/

4. Wi-Fi deauthentication attack. - https://en.wikipedia.org/wiki/Wi-Fi_deauthentication_attack

[1] Serhii Buchyk
Doctor of Technical Sciences, Professor at the Department of Cyber Security and Information Protection, Faculty of Information Technologies
[2] Yaroslav Symonychenko
Graduate student at the Department of Telecommunication and Radio Electronic Systems, Faculty of Aeronautics, Electronics and Telecommunications
[3] Anna Symonychenko
Student at the Department of Cyber Security and Information Protection, Faculty of Information Technologies
[1,3] Taras Shevchenko National University of Kyiv
[2] National Aviation University

THE METHOD OF DETECTION OF HIDDEN INFORMATION USING STEGANOGRAPHIC METHODS

Abstract. In this work it is proposed the method of detection of hidden information using steganographic methods. Abstracts focus on the method of detection of data in a digital image that is hidden using steganographic methods. The proposed method of analysis is based on detecting presence of hidden data embedded in a digital image using bit modification methods of color components of an image based on the RGB color model.

Keywords: steganographic system, steganocontainer, steganographic analysis, detection of hidden information, software steganographic tools, information protection, linguistic group, linguistic ratio parameter.

Currently, the development of methods of computer steganographic analysis is an urgent task.

The realization of steganographic analysis allows to conduct research of digital messages for the purpose to find the fact of presence of hidden data and possible decoding of a hidden message [1].

The aim of this work is the realization of the detection method of presence of hidden message using a digital image while saving it as a digital image file.

This method analyzes the distribution of the number of single bits in conditional blocks of bit planes of color components of a digital image with the purpose to detect a hidden message.

This method provides the ability to determine the type of hidden data, namely, belonging to a linguistic group, if a message is text information.

This approach is carried out by comparing the distribution of the number of single bits in conditional blocks of bit planes of color components of an image and symbols codes of text message at their binary representation.

Most often, steganographic tools are as steganographic means that can be used to hide and decode the hidden data. So, given the functionality of modern steganographic software [2]:

- the most common type of supported computer files is files with graphic information (an image file of BMP format with a digital image);
- the most common type of hidden data is text information;
- as the method of hiding data it is used computer steganographic methods that

support the direct replacement of least significant parts of an image (extra information) with bits of hidden data.

Considering that the most common type of hidden data is text information, we analyzed the distribution of number of single bits (further— DNB) for codes of message symbols of texts in English, Russian, Ukrainian.

To conduct this research we follow the steps according to the developed algorithm [3]:

Stage 1. Make representation of each message symbol in proper binary code (8 bits each).

Stage 2. Make the division of the formed bit representation into the blocks that corresponds to the bit digit of each binary representation (from 0 to 7).

Stage 3. Make calculation of the bit number with a value «1» in each bit block.

We chose 3 messages of different length in each language and make the representation of their DNB (distribution of number of single bits).

We used 24-bit digital bitmap images to research the detection method of hidden data in a digital image.

An image was embedded by the method of modification of a lower bit in the blue color component when using the RGB color model.

We conduct the realization of the proposed method of analysis of detection of hidden data by following these steps:

Stage 1. Decode each color component of the researched image using the RGB color model.

Stage 2. Convert the matrix of the blue color image component into a vector-column formed from the color gradation values of the blue component, and convert the value of the vector into binary.

Stage 3. Split each column of bit plane into eight conditional blocks that are formed by counting bits with a value «1» as follows: each conditional block (0…7) of size 1×8 contains the values of the number of single bits the proper column of the binary representation matrix (0…7) with a shift step «+8».

Stage 4. Get a matrix of conditional blocks (further MCB), of size 8×8, the columns correspond to the number of a bit plane of the image components, and rows— to the conditional block number with the count of distribution of the number of single bits in each block.

It is a graphic representation of the values of the formed matrix of conditional blocks.

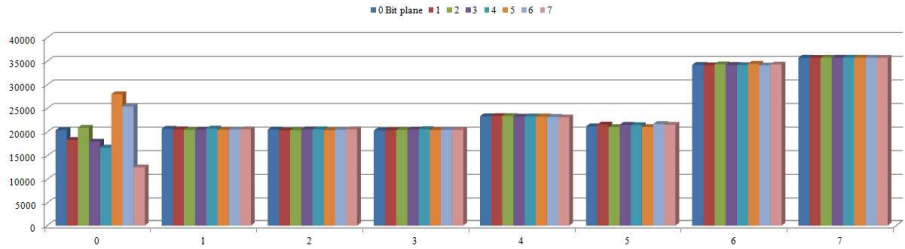

Figure 1 - Graphic representation of the values of the formed matrix of conditional blocks

To evaluate and compare DNB matching in bit image planes of an image and codes of symbols of a text message we compare them based on the Pearson correlation coefficient indicator.

Make the determination of the Pearson correlation coefficient DNB of 5 bit planes of blue color image component and symbols codes of the text message.

We used 24-bit digital bitmap images of the following classes to research the detection method of hidden data in a digital image [4-6]:

class 1 – images with few colors (4-16) and large areas filled with the same color;

class 2 – images with smooth color transitions built on a computer;

class 3 – photorealistic images;

class 4 – photorealistic images with overlay business graphics.

The degree of image modification when embedding an English text message to the blue color component of each image.

For more detailed research we embed the English message into each image. Determine the value of Pearson correlation coefficient of DNB of zero bit planes of blue color component of certain images and symbols codes of text message.

We proposed the method of steganoanalysis of the detection of presence of hidden data formed using steganographic system. This method allows to research a digital image for presence of embedded hidden text message using the method of modification of lower bit of color component of RGB model and determine language of a hidden message.

The detection of presence of hidden data in the image is conducted by comparing DNB MCB of the bit planes of the image color components and symbol codes of the message at their binary representation.

References:

1. Barannik, V. and Shulgin, S., (2016), The method of increasing accessibility of the dynamic video information resource. Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET): 13th Intern. conf., (Lviv-Slavske, Ukraine, febr. 23–26, 2016). Lviv-Slavske: 2016. pp. 621-623.

2. O. Yudin, Y. Symonychenko and A. Symonychenko, "The Method of Detection of Hidden Information in a Digital Image Using Steganographic Methods of Analysis", 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 2019, pp. 262-266, doi: 10.1109/ATIT49449.2019.9030479.

3. A. N. Alimpiev, V. V. Barannik, and S. A. Sidchenko, "The method of cryptocompression presentation of videoinformation resources in a generalized structurally positioned space," Telecommunications and Radio Engineering, vol. 76, no. 6, 2017, pp. 521–534.

4. V. Barannik, A. Bekirov, A. Lekakh, and D. Barannik, "A steganographic method based on the modification of regions of the image with different saturation," Proceeding of the XIVth International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Lviv-Slavsko, Ukraine, 2018, pp. 542-545.

5. M.Grundmann, V.Kwatra, M.Han, I.Essa, "Efficient hierarchical graph- based video segmentation," Preceedings of the IEEE Conference on Computer Vision and Pattern, San Francisco, 2010, pp. 85-91.

6. P.Gurzhiy, ,B.Gorodetsky, ,O.Yudin, ,Y.Ryabukha, "The Method of Adaptive Counteraction to Viral Attacks, Taking into Account Their Masking in Infocommunication Systems", Proceedings of the 3rd International Conference on Advanced Information and Communications Technologies (AICT), Lviv, Ukraine, 2019, pp.423-426.

Antonina Kashtalian
Ph.D., Associate Professor
Khmelnitsky National University

HONEYPOTS MODELS IN COMPUTER NETWORKS ACCORDING TO MALICIOUS ATTACKS TYPES

Abstract. The paper proposes honeypots models based on typical computer attacks and architectural features of baits developed taking into account the architecture of a distributed system with baits. The main features are configuring different types of honeypots and their integration with other components in multilevel security system. Typical features of honeypots and methods of system analysis and decision theory for solving problems are analyzed with using baits by types of attacks and organizing the interaction of components of a multilevel system.

Keywords: honeypot, malicious actions, computer attack detection, forecasting

Computer networks connected to Internet become objects for malicious actions [1, 2]. Detection of malicious actions and protection against them are used in systems of different types [3]. A promising way of computer networks protection is using separated honeypots and honeynets and their integration with other protection systems. The work goal is to develop honeypots models according to their architecture features, utilization features and attacks types on networks. Honeypots perform functions of information collection and analysis about malicious actions in networks [4, 5].

Models characteristics are grouped and generalized for model creating. With this purpose honeypots operation levels are taken into account with its functions: detection, analysis, reaction, execution. The honeypot model, taking into account the functions from which it is formed, is set in the way $M_P = \langle P, \Omega_1 \rangle$, where P is the set of a honeypot functions; Ω_1 is the set of predicates on the set P.

It is necessary to specify features of a certain type of an attack for its detection and its type determination. A honeypot should provide: ports and services that are attacked; collection, storage and processing data of network traffic of a honeypot; interaction with a honeynet.

Ports and services deployed on a honeypot depend on attack type or types which a honeypot intercepts. Collection and storage of traffic data may be implemented in the same way for different types of attacks.

Honeypots architectures are considered, taking into account typical attacks in corporate computer networks.

The attack ‘mailbomb’ is aimed at e-mail box or e-mail server. For detection of the attack ‘mailbomb’ it is necessary to develop SMTP-service for simulating e-mail server (fig. 1). Data monitored by the honeypot: IP-address/ IP-prefix of e-mail source; domain name/ URL/ URL type of e-mail source; user or users ID; received e-mail volume. On the basis of these data the analysis of activity on honeypot SMPT service is performed, threshold values of e-mail or e-mails volume and their received speed are

determined. Threshold values are used for defining e-mail traffic as malicious and one that requires analysis.

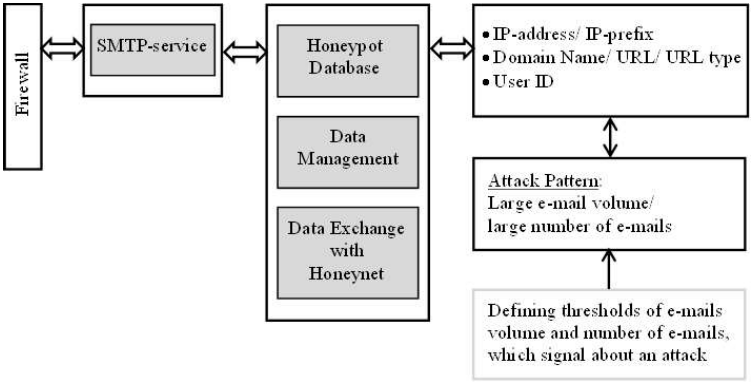

Figure 1 – The honeypot designed for an attack ‘mailbomb’

For detection of the attack ‘neptune’ the honeypot should contain the service, which works by TCP protocol, for example HNNP and FTP services. Data monitored by this type of honeypot: IP-address/ IP-prefix of the request source; domain name/ URL/ URL type of the request source; the number of half-open requests. As the result of analysis the network traffic is classified as not malicious/ malicious and threshold values of requests number and their received speed are defined.

For detecting the attack ‘portsweep’ the honeypot should contain significant number of ports, which may be both physical and virtual and have to correspond to real services ports. Data collected by the honeypot in the case of attack ‘portsweep’: IP-address/ IP-prefix; domain name/ URL/ URL type; existence of ports access. The analysis of time series of ports access provides a way to classify an activity as malicious.

The developed honeypots model for typical attacks can be applied for creating the network of honeypots inside a corporate computer network. This allows them to be used depending on the types of attacks. It is necessary to use the methods of decision theory for the organization of functioning such honeypots network with the purpose of determining a dynamic architecture of honeynet.

The dynamic honeynet allows collecting different types of attacks and analyzing intruders. The analysis provides finding similar intruders, anomaly activity detection, and forecasting intruder activity with traditional statistical and state of the art methods like deep learning.

Finding similar intruders is performed on the basis of clustering their activity, id est, time series belonged intruders. Clustering is executed with different methods, among them there are clustering on the basis of informational criterions, clustering on the basis of Gaussian mixture models, clustering on the basis of hidden Markov models, clustering on the basis of neural network models etc.

Forecasting intruders activity provides early detection of malicious actions and determination of attack probability and their characteristics on the basis of current state and previously detected patterns. The analysis of intruders actions allows to get the patterns of attacks.

The developed models of honeypots with considering their architectural features and typical attacks features are basis for creating systems of wrong attacks objects integrated in the general security system of corporate computer network. This improves security level including through analysis of information collected in a honeypot. When organizing the detection and interaction of distributed system components, it is necessary to involve the methods of system analysis and decision theory which allow enhancing the result of work of the entire system.

Conducted research of mentioned methods allows separating important ones from them for their use in the developed system. The result of experimental investigation allows performing the creation of the honeypot network on the basis of a distributed multilevel system.

References:

1. Savenko O.S Research of methods of antiviral diagnostics of computer networks / O.S. Savenko, S.M. Lysenko // Visnyk of Khmelnytsky National University. Technical sciences. - 2007. - № 2, v. 2. - P. 120–126. (in Ukrainian)

2. Savenko O.S., Klots Y.P, Mostoviy S.V. Research and analysis of process blocking in a computer system // Visnyk of Khmelnytsky National University. - 2007. - № 3, Volume 1.- P.248-251. (in Ukrainian)

3. Sidiroglou S./ Composite Hybrid Techniques for Defending Against Targeted Attacks// S. Sidiroglou, A.D. Keromytis. Part of the Advanced in Information Security book series (ADIS, volume 27), 2007, 213-229pp.

4. Sochor Tomas/ Study of Internet Threats and Attach Methods Using Honeypots and Honeynets// Tomas Sochor, Matej Zuzcak - Springer International Publishing Switzerland 2014, A. Kwiecie´n, P. Gaj, and P. Stera (Eds.): CN 2014, CCIS 431, pp. 118–127, 2014.

5. Sokol Pavol/ Data Collection and Data Analysis in Honeypots and Honeynets// Pavol Sokol, Patrik Pekarčík, Tomáš Bajtoš. http://spi.unob.cz/papers/2015/2015-19.pdf [Access 18.04.2020].

[1] **Danil Koltsov**
[2] **Ivan Parkhomenko**
PhD in Technical Sciences, Associate Professor Department of Cybersecurity and Information Protection
[1,2] *Taras Shevchenko National University of Kyiv*

TRAVERSAL UTILITIES FOR NAT

Abstract. The question of why you can't connect to the server via NAT is considered. Requirements to traversal NAT. Options for this to be done.

Keywords: Traversal NAT, P2P applications, STUN, TURN, ICE.

1. INTRODUCTION

A big problem for many applications is the inability to establish a connection at all. This is especially true for P2P applications such as VoIP, messengers, and file sharing, which often need to act as both a client and a server to provide two-way direct communication.

If NAT is available, the internal client does not know about its public IP address. It knows its internal IP address, and NAT devices overwrite the output port and address in each TCP/UDP packet and the output IP address inside the IP packet. However, if the client transmits its private IP address a part of its application data to an equal external environment outside its private network, then the connection will fail.

Therefore, if you want to share peer-to-peer code outside your private network, the application must first detect its public IP address. Another packet that arrives at the public IP address of the Nat device must also have a destination port and an entry in the NAT table that can translate it to the internal IP address of the destination host and a tuple of ports.

To eliminate this discrepancy in NAT, there are methods for bypassing STUN, TURN, and ICE, which are used to establish end-to-end communication between peer members on both sides.

2. SESSION TRAVERSAL UTILITIES FOR NAT (RFC 5389)

Session bypass utilities for NAT STUN (RFC 5389) is a protocol that allows the host application to detect the presence of a network address translator on the network and, if available, get a dedicated public IP address and a tuple port for the current connection. To do this, the protocol requires the help of a well-known third-party stun server, which must be located on a public network.

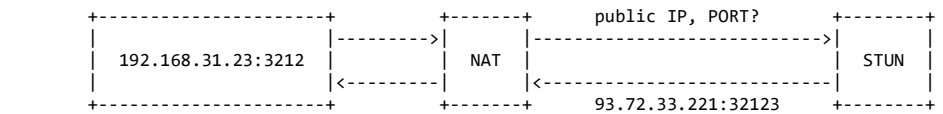

Figure 1 – STUN request for public IP and Port

Assuming that the IP address of the stun server is known (via DNS detection or at a manually specified address), the application first sends a request to bind to the stun server. In its turn, the stun server responds with a response that contains the public IP address and client port that is visible from the public network.

This process has several problems.

- The app detects its public IP and port packet, and can then use this information as part of its app data when communicating with its members.
- An outgoing binding request to the stun server sets NAT routing records along the path, so that incoming packets arriving at the public IP address and Port tuple can now find their way back to the host application on the internal network.
- The STUN protocol defines a simple ping saving mechanism to avoid waiting times for NAT routing records.

With this mechanism, when two peer-to-peer partners want to communicate with each other, they first send binding requests to their respective STUN servers, and after both parties successfully respond, they can use the established public tuples of IP and ports to exchange data.

3. TRAVERSAL USING RELAYS AROUND NAT (RFC 5766)

However, in practice, STUN is not sufficient to work with all NAT topologies and network configurations. In some cases, UDP may be blocked by a firewall or other network device - a common scenario for many corporate networks. To solve this problem when the STUN fails, we can use the relay bypass protocol around Nat (TURN) (RFC 5766) as a backup option that can work over UDP and switch to TCP if all else fails.

The key word in TURN is, of course,"relays". The protocol relies on the availability and availability of a public repeater to transmit data between members.

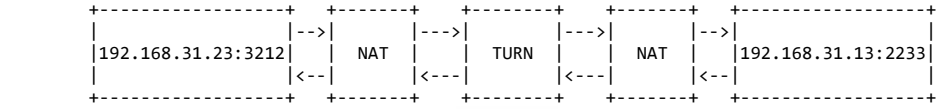

Figure 2 – TURN

- Both clients start their connections by sending a distribution request to the same turn server, followed by permission approval.
- Once reconciliation is complete, both peers communicate by sending their data to the TURN server, which then passes it to another peer.

4. INTERACTIVE CONNECTIVITY ESTABLISHMENT (RFC 5245)

ICE (RFC 5245) is a protocol and set of methods that aim to establish the most efficient tunnel between participants, provide direct connection where possible, use stun negotiations where necessary, and finally return to TURN if all else fails.

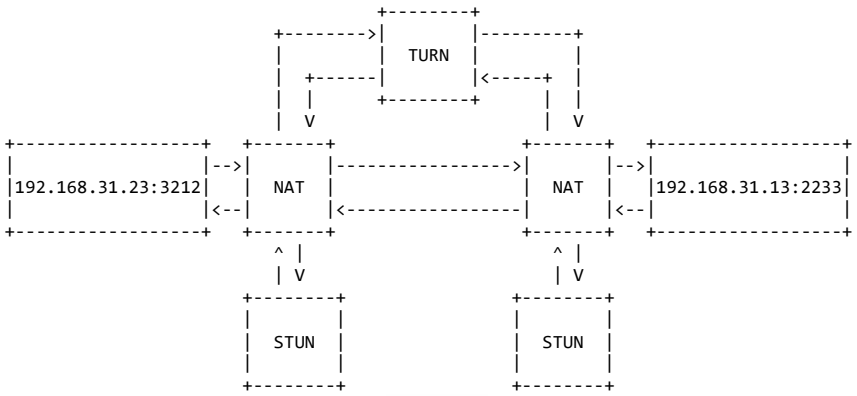

Figure 3 – ICE

5. References

Recommendations [1, 2, 3]
RFC 5389 Session Traversal Utilities for NAT
RFC 5766 Traversal Using Relays around NAT
RFC 5245 Interactive Connectivity Establishment

[1] **Nataliia Lukova-Chuiko**
Doctor of Technical Science, Professor of the Department of Cybersecurity and Information Protection
[2] **Alexander Bystrov**
Student
[1,2] *Taras Shevchenko National University of Kyiv*

ADVICE ON SELECTING AN INTRUSION DETECTION SYSTEM FOR SMALL AND MEDIUM-SIZED BUSINESSES

Abstract. This article provides recommendations for choosing an intrusion detection system for small and medium-sized businesses. These recommendations can be applied in practice by employees of the information security department of the enterprise. These recommendations can significantly increase the level of information security of the enterprise and minimize possible losses in the future.

Keywords: Open-Source, SIEM, Intrusion Detection, Monitoring

In the modern information field, there is an acute problem of information security. This problem includes both protection from attacks and their detection at an early stage in order to minimize the consequences of an attack and respond to it in time. One of the main tools to detect attacks is SIEM-systems, which, if properly configured, reduce the response time to an attack to a minimum. Typically, efficient systems are expensive, and their implementation and support are heavily funded, which can be afforded by an Enterprise or Medium-Business companies. For small companies, SIEM implementation and maintenance becomes an impossible burden, as usually the means to purchase an SIEM system are bigger than possible losses from a successful attack, which can paralyze the whole business for an indefinite period of time. The way out of this situation are open source applications that can be adapted to штекгішщт detection systems and perform their functions.

This application can be Osquery, the software was developed in 2014 by Facebook. This application is distributed under MIT license, so it can be used for commercial purposes, without any usage fees [1]. The software uses a client architecture, so the application must be installed on each server from which data is to be received. The essence of this application is that the operating system is perceived as a relational database with tables displaying information about the system. These data can be conveniently obtained with the help of SQL language and record both in the log file and sent to a remote syslog-server, which will accumulate data from different endpoints to present a complete picture of the infrastructure security. One example of using this application is the detection of one of the most popular tactics for launching malware, namely deleting the executable file after the process has already been created, making it difficult to detect the attack. So to search for these processes, you need to form a simple SQL query - "SELECT * FROM processes WHERE on_disk = 0", after executing this query, Osquery will display all the processes whose executable files have been removed from the file system. Thus, with the help of this software it is possible to

create many markers that will trigger the presence of malicious software in the system. All settings for information security markers are recorded in a configuration file, which also contains settings for information output (syslog-server, log file, etc.) and the frequency of marker checks. Thus, these configuration files are conveniently distributed between a large number of endpoints using automation tools such as Ansible.

The problem with Osquery is that the application does not provide a user-friendly interface to analyze all incidents that have been collected from different parts of the infrastructure, so to use it effectively you need to implement an application to analyze and display this data. This application is ELK-Stack which is also distributed under MIT license, so it can be used for commercial purposes [2]. This software consists of three components, each of which is responsible for different purposes and tasks. Elasticsearch is the main component of the system, which accumulates data and analyzes them. Kibana - is responsible for managing the components in the web browser interface and building charts, maps and other graphical means of displaying information using data stored in Elasticsearch. The last component used in the system is Logstash. This component is responsible for converting data to provide them in a format understandable to Elasticsearch for further analysis. Together, all these components form a powerful stack of technologies that can be used to analyze and visualize information from different sources. In addition, this software can be used not only for analysis and displaying information security events, but also for its intended purpose, namely as software for analysis and aggregation of log files and collection of metrics for information system operation. It may also be noted that the data is stored in a document-centric database, with effective indexing and mechanisms for managing this data using the API and Index Lifecycle Management. The whole set of tasks for which ELK-Stack can be adapted makes it attractive for implementation in companies with small IT departments and limited financing.

In combination, these software tools provide an opportunity to constantly monitor the entire infrastructure and identify possible attacks, which will minimize losses from them. In addition, the cost of implementing this software package is minimal, as only an information security officer with knowledge of the software is needed. At the same time, this software complex can be developed together with the company that has implemented it and scaled up together with information systems of enterprises.

References:

1. Osquery license. Available: https://github.com/osquery/osquery/blob/master/LICENSE-Apache-2.0
2. ELK license. Available: https://github.com/elastic/elasticsearch/blob/master/LICENSE.txt

[1] **Nataliia Lukova-Chuiko**
Doctor of Technical Science, Professor
[2] **Andriy Fesenko**
Candidate of Technical Science, Associate Professor
[3] **Hanna Papirna**
Master's degree student
[4] **Sergiy Gnatyuk**
Doctor of Science in Engineering, Associate Professor
[1,2,3] *Taras Shevchenko National University of Kyiv*
[4] *National Aviation University*

THREAT HUNTING AS A METHOD OF PROTECTION AGAINST CYBER THREATS

Abstract. The article presents the structuring of a new approach to countering cyber threats - Threat Hunting. The article proposes a functional diagram for an application of this approach in practice by specialists in the field of cybersecurity.

Keywords: Threat Hunting, indicators of compromise, proactive cybersecurity.

1. Introduction

To date, most information security threats are known, and can be defended by traditional means of protection such as antivirus, firewalls, and so on. Such threats include spam, denial-of-service attacks, viruses, rootkits, and other classic malware. The remaining minority of threats are unknown and the most dangerous. They are difficult to detect and even more so to protect against them. Examples of such threats are encryption viruses, crypto miners, etc. This situation has led to the development of means of protection against cyber threats in the direction of developing new technology that would be able to counteract the most serious and complex threats.

Proactive threat search or Threat Hunting (hereinafter - TH) is the latest way to counter cyberattacks, which through proactive and iterative search, allows to detect complex threats that traditional means of protection are not even able to notice. It should be noted that TH is not a specific software or hardware product and is not a passive activity. Proactive threat search is, first of all, mainly a manual process with elements of automation, in which the analyst, based on his knowledge and skills, checks large amounts of information for indicators of compromise, according to a predetermined hypothesis of the presence of a threat.

2. The order of Threat Hunting conduct

According to the approach of the leading American company in the field of cybersecurity and big data analytics - Sqrrl, in general, the whole process of TH can be reduced to four main stages, which are repeated cyclically [1].

On the first stage of the hypothesis creation begins with asking questions about how an attacker can gain access to an organization's network. Then these questions need to be divided into specific and measurable hypotheses that determine what threats may be present in the network and how they can be identified [2].

Once the observations have led to the development of hypotheses, they should be tested during the stage of research. In general, it is possible to identify four types of techniques that can be used by specialists in TH at this stage: search, clustering, stack counting or accumulation and machine learning [2].

A fairly effective method at this stage is Linked data - a method of publishing structured data that allows to link them and seek confirmation of hypotheses using semantic queries. Related data analysis is particularly effective in presenting the data needed to solve hypotheses in an understandable form, and is therefore an important component of the TH. Linked data can even help prioritize and direct visualization, making it easier to search large datasets and use more powerful analytics. Methods of analysis of both source and related data should be used to identify patterns in disparate data sets, to detect the actions of attackers [1].

Today, there are many information security technologies that can provide assistance in the process of TH. However, the authors of this paper tend to narrow the set of technologies to the next most necessary: SIEM (Security Information and Event Management) systems, EDR (Endpoint Detection and Response) systems and NTA (Network Traffic Analysis) systems. As experts in the field of cybersecurity, the authors of this article note that the above systems are the technical basis for the construction of a modern SOC (Security Operations Center).

SOC is a specialized center for monitoring and prompt response to information security incidents. Such a center is a group of information security experts who are responsible for continuous monitoring and analysis of the security of the organization, using a combination of technological solutions and acting within well-structured processes. It is important to note that most often TH processes seek to implement organizations that already have their own SOC or use such services through outsourcing.

The third stage allows to reveal new harmful patterns of behavior and tactics, techniques and procedures (hereinafter - TTP). The gap in the detection of violations arises from the ability of attackers to evade the mechanisms of detection. As detection capabilities continue to evolve and expand, cybercriminals will find new ways to evade these measures. Thus, over time, TTPs of attackers will evolve to ensure that they can evade detection and act unnoticed in the IT environment.

It is due to this stage of the previous TH procedures that the TTP frameworks are filled and improved.

MITRE ATT&CK is a structure that describes the methodologies used by attackers during cyberattacks. It is presented in the form of a matrix consisting of eleven tactics, each of which contains a list of related techniques [3].

The fourth stage of the cycle forms the basis for informing and enriching automated analytics. This can be done in a variety of ways, including developing a default search for regular execution, creating new analytics or even providing feedback to a controlled machine learning algorithm [1].

One of the main mistakes of organizations initiating the TH process to their overall information security strategy is that they do not define metrics for assessing the TH either because of the difficulty of defining such indicators or because they believe

that because threat detection must be a flexible process, indicators cannot be identified.

However, there are useful metrics that can measure the performance of the TH process to help improve it, as well as help build a business case for further investment (financial and time) in staff training and tools. The following is an approximate set of metrics that can be used [2]:

— graph of trends and / or comparisons: number of incidents detected proactively (compared to reactively); the number of proactively identified vulnerabilities (compared to vulnerability assessments); waiting time (delta) in the detection of incidents proactively (compared to those detected reactively);

— percentage: data coverage (data types and asset coverage).

— pie chart: number of hypotheses on MITRE ATT&CK tactics; number of TH procedures according to MITRE ATT&CK tactics; number of incidents under MITRE ATT&CK tactics.

— service level: the percentage of successful THs that led to a new analytical conclusion or detection rule; sensitivity and specificity of analytics or rules obtained as a result of TH (number of true and false positive results).

3. Developing Threat Hunting process diagram

The "Hunting loop" by Sqrrl [1] provides a working and stable cycle of actions for experts. However, in the context of organizations where there are information security departments wishing to implement TH process, the cycle needs to be detailed and supplemented with initial data, as well as a connection with the classic incident management process. The authors of the paper propose the improvement of the original cycle, described on the Figure 1 and its integration in the whole information security management system of organizations.

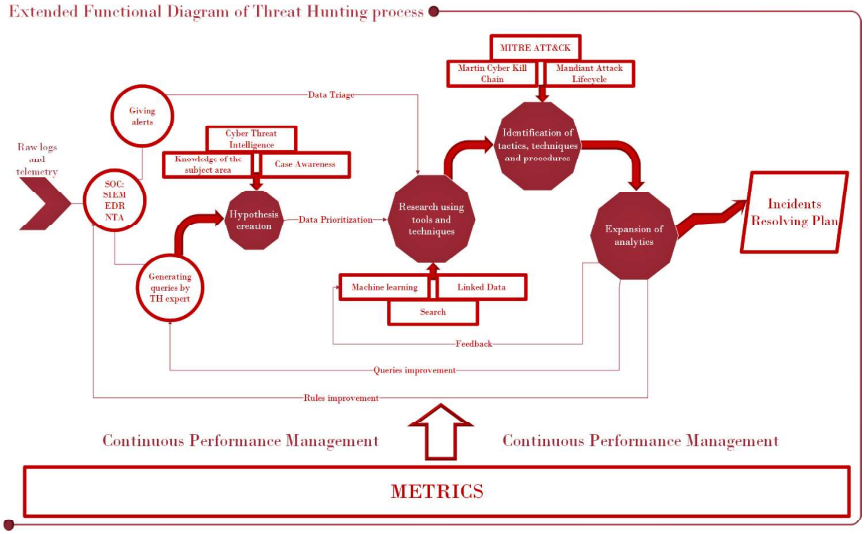

Figure 1 – Extended functional diagram of Threat Hunting process

The proposed extended functional diagram of the TH process includes:

— raw logs and telemetry from network and infrastructure assets of the organization;

— means of the initial analytics and information processing – SOC: SIEM, EDR, NTA;

— alerts and queries conducted by TH expert using initial analytics means;

— stage of the hypothesis creation;

— preliminary data triage and prioritization before conducting the research stage;

— stage of the research using tools and techniques; stage of the identification of the TTPs enriched by the trusted frameworks in the field;

— detailed outputs of the expansion of analytics stage;

— incidents resolving plan as an ultimate goal of the incident management process;

— metrics for the continuous performance management of the TH process.

Existing SOC tools, such as the SIEM platform, in the context of the second stage of TH, can be used to query data, from basic search to more advanced methods, and visualization can help identify anomalies and unusual patterns of behavior.

It can be said that the planning and implementation of the TH procedure in the daily process of information security can afford organizations with a high level of maturity of security processes, which already have established procedures and technologies for threat prevention and are ready to move to a higher level - the level of proactive threat response.

Thus, we can conclude that the TH cycle is a simple but effective process that can radically improve the level of security of the organization. This procedure is most effective when used in conjunction with traditional security systems, complementing the measures and tools to detect cyber threats that already exist in most organizations. The ultimate goal of TH should always be to go through the four-stage loop as efficiently as possible.

References:

1. White paper: A Framework for Cyber Threat Hunting, Sqrrl Data, 2018. URL: https://www.threathunting.net/files/framework-for-threat-hunting-whitepaper.pdf.

2. Detecting the Unknown: A Guide to Threat Hunting, Home Office Digital, Data and Technology, version 2.0, 2019. URL: https://hodigital.blog.gov.uk/wp-content/uploads/sites/161/2020/03/ Detecting-the-Unknown-A-Guide-to-Threat-Hunting-v2.0.pdf.

3. MITRE ATT&CK Framework, 2020. URL: https://attack.mitre.org/tactics/enterprise/.

[1] **Nataliia Lukova-Chuiko**
Doctor of Technical Science, Professor of the Department of Cybersecurity and Information Protection
[2] **Victoriia Klochko**
Student
[1,2] *Taras Shevchenko National University of Kyiv*

COLLECTIVE DEFENSE OF CORPORATE NETWORKS AGAINST COMPUTER ATTACKS

Abstract. These days, content analysis of text information is used to prevent threats, along with the analysis of the network traffic characteristics, the behavior of corporate networks and their security policy. Existing systems of text analysis and modeling include different kinds of search engines and information-analytical systems. They are capable of solving such tasks as classification of documents by its subject matter, author identification, detection of plagiarism, modeling representations of the knowledge about the subject area and the content of text.

Keywords: corporate network, attacks, SDA, System Monitoring Unit, cybersecurity, collective protection.

In the modern world, problems related to the use and spread of malicious software, information attacks and other types of cyber threats, which have received the general name “cybercrime” are becoming more and more relevant. Sophisticated threats require an innovative approach. Collective defense empowers organization to stay ahead of evolving threats to better defend network through real-time sharing and collaboration across industries and sectors.

Earlier, the main efforts of developers were pointed to create effective detection algorithms. These detection algorithms have used different mathematical basis: statistical methods, methods of automata theory, methods of interacting sequential processes calculus, methods of mathematical logics, neural networks, fuzzy logics, and other formalisms.

Some detection algorithms, in particular algorithms on basis of neural networks have cyberspace-adaptive properties. However, the rapid dynamics of the environment change (the variety of network structures, the variety of types of attacks, etc.) often reduced efforts of designers to zero.

A number of freely distributed and commercial systems of defense from attacks (SDA) was developed and became widely accepted in the field of corporate computer networks building [1].

The analysis of the structure of circulating packages in the corporate network is the essence of the analysis at the network layer of protection in SDA. As a rule, the package flags, the port addresses for network nodes, the time intervals between specific events and so on are analyzed here.

The package contains the information about the sender, which is often represented as a DNS-address. This information is definitely of a great value as it can

clearly point at the source of the attack. However, the truth of address information about the source of the attack is often questionable, since it can be easily corrected by the sender of the package. For some protocols, such as mail, the address of the attacker may also be obviously stated. However, as in the previous case, the address of the sender can easily be changed.

As a result, there is a need to allocate one more level of realization of the protective methods – the level of the global network.

At this level the information, which is contained in the text documents on web-sites, global network portals, social networks or other legitimate objects of the information space can be analyzed and both the sources of attacks and their information characteristics can be indirectly identified.

The concept of a text document here is multivalued: it is text information from websites and portals, and emails, and program codes that are entered into the computing environment of the victim's computer. In any case, this level is characterized by, on the one hand, methods used in intelligence activities, including business or competitive intelligence, and, on the other hand, methods of text processing.

IT professionals very often have problems with viruses and other malware. Actual threats include spreading spam, phishing, network attacks on enterprise infrastructure, including tar- get and DDoS attacks, where use potentially dangerous software vulnerabilities.

These and other similar examples show a close relationship between cybersecurity systems and word processing systems: when detecting spam, data loss, detecting and tracking potentially dangerous messages, etc.

This field of the research is actively evolving lately. From one side, it is connected with intellectual property protection, from another, it is connected with the necessity of cyber threats prevention, which arises because of the malware usage. In the latter case, it is hard to overestimate the possible damage, which can be caused to control systems by the key infrastructure, including to the military targets. Because there are new kinds of malware being created all over the globe, there is a necessity of the identification of the malicious code creators and bringing them to justice.

Processing, careful analysis and synthesis of information collected from Internet resources is made using content and/or rapid analysis methods, bibliometric and/or cluster analysis, as well as expert and/or situational methods [2].

However, a tight time limit for the search, collection, extraction and processing of information circulating in the global information space of the Internet, its accumulation, classification by certain attributes, further analysis, synthesis, compilation and making it accessible to the concerned users, as well as transformation into synthesized conclusions and recommendations necessitates some arrangements. First, the automation of all measures in the complex of risks monitoring system associated with these processes. Second, the configuration of SDAs subordinate to the System Monitoring Units of corporate networks according to their risk vectors.

The development of a corporate networks protection model with a collective System Monitoring Unit defense module, methods for detecting and identifying computer attacks with help of content analysis of the global information space and the

architecture of SDA, related to it, will provide a basis for the synthesis of a reliable and high-performance adaptive cyber threats detection systems and will shorten the detection time of the computer attacks of the new generation.

Further improvement of the security and stability in functioning of the information and telecommunication systems of corporate networks in the conditions of massive influence of computer attacks requires an increase in the probability of detection of new computer attacks and a decrease in the recognition time for the signs of known attacks [3].

To solve this problem, it is not enough to use only traditional methods that utilize identification characteristics of network traffic and information about the work of corporate networks and security devices. The processing of data sets of the body of network packages, content of Internet pages, information from social networks is very valuable in this area.

Calculations of risks from various attacks require the identification of sources of attacks on indirect grounds, determining their inclinations to attacks or undesirable influences of one kind or another, determining the characteristics of attack activity, calculating predictive activity indicators based on time series analysis, and the like other [4].

This protection becomes possible or by configuring the corporate network SDA to prepare the activation of attack detection algorithms. Given the temporary limitations of the attack detection process, such actions should be performed based on predictions of the activity of potential attack sources, the detection of which is the task of the global network security level of the corporate network.

References:

1. Chi, S.-D., Park, J. S., Jung, K.-C. & Lee, J.-S., Network security modeling and cyber-attack simulation methodology. Lecture Notes in Computer Science. Springer-Verlag 2001.

2. Martovitsky V., Ruban I., Lukova-Chuiko N., Kortyak E., Kruglikov Y., Model monitoring network infrastructure based on standard FIPA, 2020.

3. Ruban I., Martovitsky V., Lukova-Chuiko N., Approach to Classifying the State of a Network Based on Statistical Parameters for Detecting Anomalies in the Information Structure of a Computing System (2018), 302-309.

4. Shannon, C. E., A mathematical theory of communication. Bell System Technical (1948), Vols. 27.

[1] **Volodymyr Nakonechnyi**
Doctor of Technical Sciences, Professor of the Department of Cybersecurity and Information Protection
[2] **Maksym Bondarenko**
Master`s degree, Student
[1,2] Kyiv National University of Taras Shevchenko, Ukraine

APPLICATION OF BIOMETRIC METHODS OF USER IDENTIFICATION IN INFORMATION AND COMMUNICATION SYSTEMS

Abstract. Biometrics are physical or behavioral human characteristics to that can be used to digitally identify a person to grant access to systems, devices or data. In this publication it is described why this topic is an important part of information security and corporate information security as well; moreover, using of multimodal biometric identification system was proposed for stronger security.

Keywords: biometry, IT, security, cybersecurity, characteristics, error, identification, recognition, reliability.

The relentless expansion of the scope of computer information processing and computer telecommunications is attracting more and more people to the field of information technology, which increases the risks of information threats and their implementation. Despite the extensive technological capabilities of protection, today the number of crimes and fraud is growing with every minute.

One of the most common protection technologies is biometric protection system. It is the most convenient because it does not require storing complex passwords or carrying special identifiers (keys, cards, etc.), and it will be enough to just say the code word, put your finger or hand, or put the face to scan to access.

There is a limited number of characteristic personality attributes that can be used to identify a person [1]:

- what the person owns (identification mark, key or plastic card);
- features of behavior (language, handwriting, the nature of the keyboard);
- some physical characteristics (fingerprints, hand shape, blood vessel pattern).

These properties are used in everyday practice when people communicate with each other to identify visitors, messages, etc. On their basis special automatic devices are created and methods of identification of the person are developed.

A common characteristic that is used to compare different methods and techniques of biometric identification are statistical indicators [2]: error of the first kind (do not let into the system of "local") and second kind error (let into the system of someone else).

It is very difficult to sort and compare statistical and dynamic biometric methods according to the first kind of errors, as they are different for the same methods due to

the equipment for which they are implemented [3].

According to the indicators of errors of the second kind, the general sorting of biometric authentication methods looks like this (from best to worst) [1]:

- DNA.
- Iris, retina.
- Fingerprint, facial thermogram, palm shape.
- The shape of the face, the placement of veins on the palm and hand.
- Signature.
- Keyboard handwriting.
- Voice

Based on the analysis of modern biometric systems of human recognition, it was proposed to use a multimodal (bimodal) system of identification, which consists of two characteristics: face and voice.

Multimodal biometric identification system is a multi-factorial identification of a person, which consists of two main static components [3]:

1) identification with the image of a person;

2) identification with a passphrase.

Face identification is performed in real time mode at the moment of raising or approaching the device with the camera. Three images are enough for registration and identification.

Voice identification is based on the use of a static password phrase. At the stage of phrase registration it is necessary to repeat several times, in this way the maximum reliability is reached and the variability of the utterance is estimated.

References:

1. Біометрія як універсальний спосіб ідентифікації людини [Електронний ресурс]. – Режим доступу: http://bablyukh.clan.su/publ/1-1-0-4

2. DigitalPersona Fingerprint Identity Solutions for Identity Protection, Security and Compliance [Електронний ресурс]. – Режим доступу: http://www.digitalpersona.com.

3. Identix – Protecting and Securing Personal Identities and Assets [Електронний ресурс]. – Режим доступу: http://www.l1id.com/pages/17.

[1] **Volodymyr Nakonechnyi**
Doctor of Technical Sciences, Professor of the Department of Cybersecurity and Information Protection
[2] **Ivan Voitenko**
Master's degree, Student
[1,2] *Taras Shevchenko National University of Kyiv*

COMPARATIVE CHARACTERISTICS OF ALGORITHMS TO IMPROVE SPAM PREVENTION MECHANISM

Abstract. E-mail spam is one of the main problems of the modern Internet, causing financial damage to companies and irritates individual users. Among the approaches designed to prevent spam, filtering is essential and popular. [1] This article provides an overview of the state of comprehensive algorithms used in machine learning applications for spam filtering and evaluating and comparing different filtering methods. There is also a brief description of mechanisms that operates to prevent spam.

Keywords: IT, security, cybersecurity, characteristics, spam, e-mail, prevention, filter, algorithms.

Spam is any unwanted, unsolicited digital communication, often e-mail, that is sent in bulk. Today, there are about 900,000 active spam servers that send malicious mail to large companies and users. [2] Methods of protection against this type of attack exist, but they may not be sufficient to weed out unwanted e-mails. Therefore, this issue is an urgent problem today. Spam e-mails are classified into the following categories: [3]

- Advertising companies or people who advertise their products by sending e-mails.
- Advertising of illegal products - people or companies that advertise illegal products by sending e-mails.
- Anti-advertising - the type of ads that negatively affects a company's reputation, person or other communities.
- Phishing - a type of fraud in which the letter may have malicious code or a file. Usually, launching the program or going to a link from this type of message is accompanied by an infection of the workstation or withdrawal of funds from the user's bank card.

Bulk spam has a low cost per message for the sender. However, a considerable amount of useless messages causes apparent harm to recipients. First of all, we are talking about the time wasted on filtering out unnecessary mail and looking for necessary individual letters among it. [4] Very often, internet traffic is expensive, and the user has to pay for obviously unnecessary e-mails. The most significant harm comes from the ignorance of spam recipients who open spam e-mails, follow links allegedly sent by their friends, download viruses and, without suspecting that, spread them in society. It is providers who have to waste resources on redundant hardware and anti-spam systems. According to publicly available statistics, at least 80% of forwarded e-

mails are currently spam. [5] Mail servers cut most of it off at the time of receipt. But even the remaining small part is enough to complicate the lives of users. Providers incur additional costs due to the constant need to fight spammers. Since sales letters tend to be very different from regular correspondence, filtering them out of the incoming mail flow has become a standard method of dealing with them. Currently, this method is the main and most widely used one. [6]

Generally, there are two preventing methods from e-mail spam: [7]

- Automatic filtering - There is software for automatically detecting spam. It can be intended for end-users or use on servers. This software takes two main approaches.
    - The first is that the letter's content is analyzed, and a conclusion is made whether it is spam or not. A message classified as spam is separated from other correspondence: it can be marked, moved to another folder, deleted. Such software can run both on the server and the client's computer. In the latter case, the user does not see the filtered spam but continues to incur the costs associated with receiving it since the filtering software receives each message and only then decides whether to show it or not. However, the user risks not receiving a letter perceived by the filter as spam in this case.
    - The second approach uses various methods to identify the sender as a spammer without looking at the letter's body. This software can only run on a server that directly receives messages. With this approach, additional traffic is spent only by the server for communicating with spam mail programs and calls to other servers during the check.
- Manual filtration - many programs and mail services on the Internet allow users to set their filters. Such filters can consist of words or, less often, regular expressions, depending on the presence or absence of which the message gets or does not end up in the trash bin. However, such filtering is time-consuming and inflexible and requires a certain degree of familiarity with using computers. On the other hand, it allows you to filter out some spam effectively, and the user knows strictly which messages will be filtered out and why.

Automated filtering software uses statistical analysis of e-mail content to decide whether it is spam. In practice, Bayesian spam filtering methods are popular. [8] For these methods to work, preliminary "training" of the filters is required by sending them manually sorted letters to reveal periodic messages and spam's statistical features. E-mails have many attributes, such as headers, domain keys, MX records, etc. used by automatic spam filters to analyze them. [9] Therefore, the question of choosing a fast and efficient automated spam filter algorithm is essential. The following charts use the notation for recognition algorithms to be analyzed: [11]

- Bc –Bayes;
- Kn –Correlation;
- Mh – Mahalanobis distance;
- Th – Thermal agitation;
- Kp – Capon.

We described possible algorithms for using automated spam filters to enhance performance. Each of the algorithms has its advantages and disadvantages, among others [11]. The first figure shows the efficiency of algorithms when the number of email attributes is more than 5 (figure 1).

The second figure represents a graph of the efficiency of the algorithm when the number of system errors is more than 5 (figure 2).

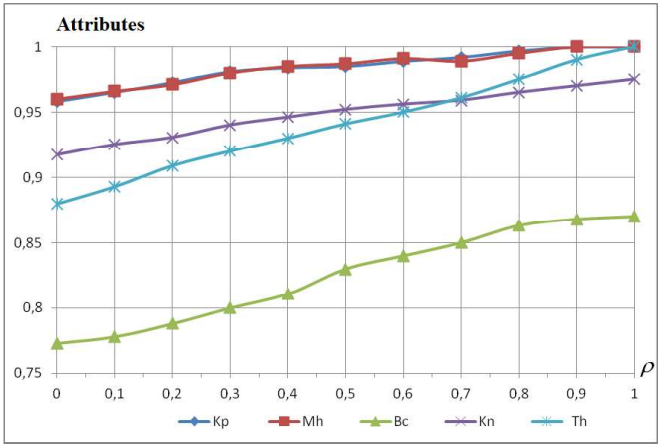

Figure 1 - Dependences on the coefficient with the number of e-mail attributes

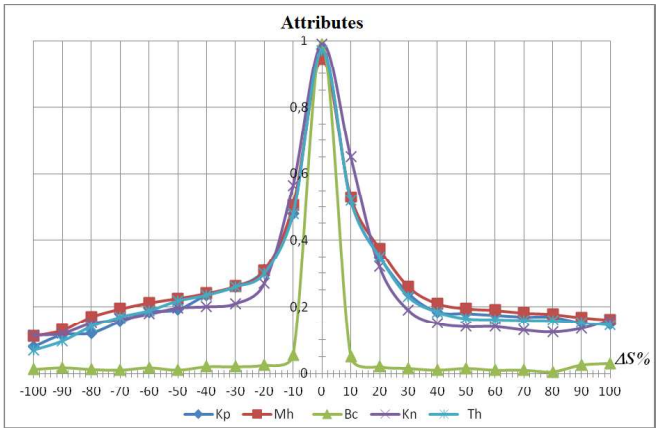

Figure 2 - Dependences on the coefficient with the number of false positive attributes

Mathematical modeling methods analyzed the effectiveness of detecting spam or malicious e-mails by the proposed algorithms. In this article, we confirm that Capon's algorithm is the most effective compared to other methods. Its use in preventing spam messages and improving the performance of the automated spam filter by 5%.

References:

1. Agrawal B, Kumar N, Molle M (2005) Controlling spam emails at the routers., Proceedings of the IEEE international conference on communications, ICC 2005, vol 3, pp 1588–1592

2. Albrecht K, Burri N, Wattenhofer R (2005) Spamato—an extendable spam filter system, Proceedings of second conference on email and anti-spam, CEAS'2005

3. Androutsopoulos I, Koutsias J, Chandrinos KV, Spyropoulos CD (2000a) An evaluation of naive bayesian anti-spam filtering, Potamias G, Moustakis V, van Someren M (eds) Proceedings of the workshop on machine learning in the new information age, 11th European conference on machine learning, ECML 2000, pp 9–17

4. Androutsopoulos I, Koutsias J, Chandrinos KV, Spyropoulos CD (2000b) An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages, Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '00. ACM Press, New York, NY, USA, pp 160–167. ISBN 1-58113-226-3. http://doi.acm.org/10.1145/345508.345569

5. Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos C, Stamatopoulos P (2000c) Learning to filter spam e-mail: a comparison of a naive bayesian and a memory-based approach, Zaragoza H, Gallinari P, Rajman M (eds) Proceedings of the workshop on machine learning and textual information access, 4th European conference on principles and practice of knowledge discovery in databases, PKDD 2000 pp 1–13

6. Androutsopoulos I, Paliouras G, Michelakis E (2004) Learning to filter unsolicited commercial e-mail (Technical Report 2004/2). NCSR "Demokritos". Revised version

7. Androutsopoulos I, Magirou E, Vassilakis D (2005) A game theoretic model of spam e-mailing, Proceedings of second conference on email and anti-spam, CEAS'2005

8. Aradhye H, Myers G, Herson J (2005) Image analysis for efficient categorization of image-based spam e-mail, Proceedings of eighth international conference on document analysis and recognition, ICDAR 2005, vol 2. IEEE Computer Society, pp 914–918.

9. Blanzieri E, Bryl A (2007) Evaluation of the highest probability svm nearest neighbor classifier with variable relative error cost, Proceedings of fourth conference on email and anti-spam, CEAS'2007. pp 5

10. Enrico Blanzieri & Anton Bryl (2009) - A survey of learning-based techniques of email spam filtering, Artificial Intelligence Review, pp 62-92.

11. Наконечний В.С Методи та засоби підвищення ефективності функціонування радіотехнічних систем розпізнавання багатопозиційного базування / В.С. Наконечний, С.В. Толюпа, В.А. Дружинін, Н.В. Лукова-Чуйко, І.І. Пархоменко: монографія - К.: 2019. - 237 с.

[1] **Anastasiia Nicheporuk**
Postgraduate Student at the Computer Engineering & System Programming Department
[2] **Oleg Savenko**
Doctor of Technical Sciences, Full Professor, Computer Engineering & System Programming Department
[3] **Andrii Kazantsev**
Postgraduate Student at the Computer Engineering & System Programming Department
[1-3] *Khmelnitsky National University*

THE ARCHITECTURE OF CNN MODEL FOR ANDROID MALWARE DETECTION

Humanity has made a significant leap forward in the development of the information technology industry and, in particular, mobile devices operated by the Android operating system. However, the advent of new mobile devices features has automatically created new vulnerabilities for them and has driven an increase in the amount of malware, which are trying to use them.

Considerable attention today is being paid to the problem of Android malware detection. Several solutions have been developed using an academic approach, utilizing features such as permissions, API calls, opcodes, strings, metadata or intents. Naive Bayes, J48, decision tree, C-means clustering methods are used as machine learning algorithms [1, 2]. However, the presented solutions are characterized by a fairly high level of complexity which prevents their use in real-time systems.

As a result of ours' research architecture of convolutional neural network (CNN) for Android malware detection have been proposed. Involvement of convolution layers creates an analogy with the human brain, allowing the identification of local features that are subsequently fed to the input of fully connected layers to form a membership degree of an input object to one of the predicted classes. In the field of pattern recognition, such features may be, for example, the presence of inclined lines at a certain angle. Another important advantage is that the weights in the convolution layer are locally connected and move throughout the feature map. This leads to involving much less of a number of weights compared to fully connected neural network architectures.

When designing architecture of neural network, we aimed to use it to detect malware. As an input data for convolutional neural network, we used the API method calls and a set of permissions for Android app. Application Program Interface (API) is a set of procedures that represent an intermediate layer for communicating applications between themselves and the Android kernel. In fact, no one high-level action doesn't take place without the participation of API invocation. Thus, by analyzing them, we could represent the behavior of the application through the sequence of API calls. For example, the sequence getDeviceId (), loadLibrary (), sendTextMessage () might be determined as the behavior of receiving and sending information. The detection process may then be

defined as a procedure of search the similarity of the program's behavior with the knowledge about the typical malware behaviors.

Except API calls, no less important attributes that can enhance behavior representation is a set of app's permissions. The permissions mechanism restricts access to certain components or functionalities of the application. All permissions used by the application are specified in the AndroidManifest.xml file. According to the results of previous studies, the distribution of permissions in malware and benign applications is differed. Thus, knowledge of attracting permissions may indicate a set of potential actions that will need to be granted. In order to implement convolutional neural network both type of data was represented in binary form.

The proposed neural network consists of two separate parallel convolutional branches, each of which processes its own type of data (see Fig. 1). As a result of convolution and max-pooling operations, the input data, i.e. behavioral patterns of Android app, is prepared for fully connected layers (FCL). In order to produce nonlinear decision making, there is one hidden layer between the first and third FCL. The result is provided by the last layer consisting of two neurons.

The proposed neural network architecture utilized an approach without convolutional and max-pooling layers alternating. This is due to the fact that after the next pair of CONV + POOL layers, the dimension of the data decreases, which leads to the loss of some information about the input object. In the proposed architecture, in each of the two sub-branch, the two convolution layers C_{11} and C_{12}, as well C_{21} and C_{22}, are placed one after the other, where the first convolution layers C_{11} and C_{21} highlight simple features that will be used by the layers C_{12} and C_{22} to represent higher-level behavior patterns. The input matrices with size $K_{11} \times D_a$ and $K_{21} \times D_p$, respectively, are feed to the convolutional layers C_{11} and C_{21}. The D_a and D_p are the dimension of input feature vector for API call and permissions respectively. For layers C_{12} and C_{22}, the size of the convolution kernel is $K_{12} \times 1$ and $K_{22} \times 1$, respectively. Following each pairs of convolution layers there is placed one aggregation layer, which reduce the dimension of each type of feature. In order to transform the data into a one-dimensional vector, each sub-branch uses a Flatten layer. After concatenation the data of both sub-branches, the resulting vector with size $F_1 + F_2$ is feeds to FCL. The output layer consists of two neurons that accumulate the probability that a suspicious Android application belongs to one of two classes – malware or benign. The quantitative indicators of the proposed convolution network are presented on the Table 1.

For all layers, except last, the ReLu activation function was selected. The neurons of the last layer were activated by a softmax function that simulates the probabilities of belonging suspicious app to one of the two classes. The neural network minimized the cross-entropy loss function. In order to reduce the impact of overfitting of the neural network between fully connected layers dropout regularization was used with parameter p = 0.5 (during testing, the dropout parameter was p = 1.0). The learning rate and the batches size were set at 0.001 and 64, respectively.

Table 1 – Quantitative indicators of the proposed convolution network

№	K_{11}	K_{12}	C_{11}	C_{12}	F_1	K_{21}	K_{22}	C_{21}	C_{22}	F_2
1.	*3*	*3*	*64*	*128*	*64*	*2*	*3*	*64*	*128*	*64*

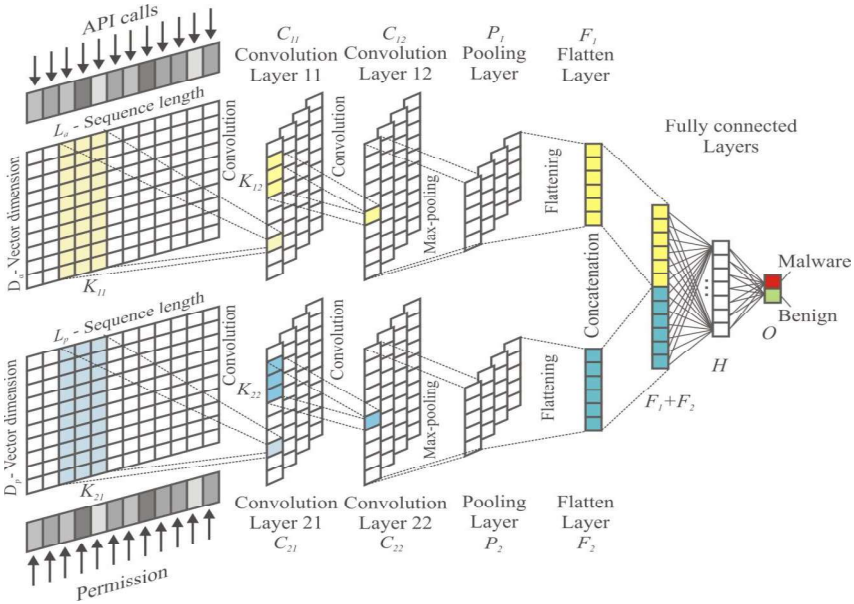

Figure 1 – The architecture of CNN model for Android malware detection

We propose architecture of convolution neural network which is the basis for Android malware detection. The Android malware detection process is based on the involvement of a neural network for training on a test sample, each instance of which is presented in the form of API calls and a set of permissions. The architecture of the proposed neural network consists of two separate parallel convolutional branches, each of which processes its own type of data. The outputs from both branches of the network are combined to form the input for fully connected layers, which determine the probabilities of belonging suspicious app to one of the classes – malware or benign.

References:

1. M.K. Alzaylaee, S. Yerima, S. Sezer DL-Droid: Deep learning based android malware detection using real devices, Computers & Security (2020) vol. 89 doi: 10.1016/j.cose.2019.101663

2. L. Wen, H. Yu, An Android malware detection system based on machine learning, Proceedings of the International Conference on Green Energy and Sustainable Development, Chongqing City, China, 2017, pp. 1-7 doi:10.1063/1.4992953

[1] **Ponomarov Serhii**
Student.
[2] **Natalia Lukova-Chuiko**
Doctor of Technical Science, Professor of the Department of Cybersecurity and Information Protection.
[1,2] *Taras Shevchenko National University of Kyiv*

BREACH AND ATTACK SIMULATION AS A NEW VECTOR OF INFORMATION SECURITY

Abstract. This article discusses the excellence of the increasingly popular automated testing tool over today's common network security testing tools. This approach allows you to automate the capabilities of the traditional penetration test, keep the condition of protection under constant control and monitor all key vectors of attacks simultaneously.

Keywords: BAS, Red Team, Penetration Test, IT, SaaS, cybersecurity

1. Introduction

Penetration test is one of the most common methods for assessing the level of protection and security strength, demonstrating methods for attacks and identifying existing security problems.

However, penetration test require a significant portion of human participation and are held with a certain frequency and in a short period of time. The results reflect the static image recorded at the time of the event.

Nowadays Breach and Attack Simulation (BAS) is the most popular product for simulating intrusions and attacks on the new, rapidly growing market of tools for automatic security testing in real time.

2. Analysis of how to test network security and their shortcomings

Penetration tests are performed manually by employees of some company or external consultants who try to assess the security of infrastructure of organization by breaking it down safely.

There are some drawbacks of the penetration test:

- The results depend on skills and experience of the tester and do not give a complete image as it is not possible to check all aspects of the system manually.
- A limited testing environment does not allow using of all features that real hackers could have.
- A tester is not able to check all known attack technologies.
- Penetration test results reflect the state of the system in a certain period of time. The high cost of such testing does not allow it to be conducted often.

Red Team testing is an imitation of a targeted cyber attack, which is becoming more and more popular. In addition to identifying critical vulnerabilities and the overall security assessment, a proactive approach provides valuable information about the ability of IT services to detect and block attacks directly during their implementation. [1]

The disadvantages of this approach are:

- Requires trained full-time or part-time staff.
- Imitations should be performed regularly.
- Test results can be difficult to compare as they can be performed under different rules and conditions.
- Significant resources are involved.
- Due to insufficient automation, it is important to repeat testing uniformly.
- It is difficult to assess the impact of changes in the IT environment and track the dynamics of protection effectiveness.

3. Breach and Attack Simulation as the way automatic testing

The BAS platform develops the idea of simulating targeted attacks and assesses the actual readiness of the organization to repel cyber attacks. This method allows to detect critical infrastructure vulnerabilities by conducting cyberattacks from several vectors as it would be done by real attackers. [2,4]

Penetration tests are carried out according to the patterns of real hacker groups, state cyber forces and even on behalf of imaginary unreliable employees.

The SaaS model allows you to run simulations at any time without affecting users or infrastructure[3].

Here are some reasons why BAS is singled out in the cybersecurity services market:

- Allows to automate the verification process.
- Decreases the influence of the human factor.
- IT infrastructure is a living organism and it changes regularly with new vulnerabilities and threats appearing every day, which leads to the need for constant security testing.
- Not every big enterprise can afford continuous penetration testing services or its own Red Team.

4. Conclusion

From the material mentioned above we can make a conclusion that BAS products allow companies to independently and continuously assess their own security, to check security mechanisms by simulating attacks in various directions. Thus, the use of BAS solutions for continuous security assessment can significantly increase the level of actual security of the IT infrastructure.

5. References:

1. List of Adversary Emulation Tools. Available**:** https://pentestit.com/adversary-emulation-tools-list/

2. What is breach and attack simulation. Available: https://blog.cymulate.com/what-is-breach-attack-simulation

3. H.A. Kuchuk, A.A. Kovalenko, N.V. Lukova-Chuyko Method to minimize the average delay of packets into virtual connection support network cloud services. Management systems, navigation and communication. - Poltava. National University «Yuri Kondratyuk Poltava Polytechnic», 2017. - Vol. 2 (42). - P. 117-120.

4. Ruban, V. Martovytskyi, N. Lukova-Chuiko approach to classification of network condition on the basis of statistical parameters for detection of anomalies in the information structure of the computing system. Cybernetics and Systems Analysis. 2018.Vol. 54, No. 2. P. 142-150.

[1] Volodymyr Rusyn
PhD, Assistant Professor of Department of Radio Engineering and Information Security
[2] Aceng Sambas
Lecturer of Department of Mechanical Engineering
[1] Yuriy Fedkovych Chernivtsi National University, Ukraine
[2] Universitas Muhammadiyah Tasikmalaya, Indonesia

SIMPLE AUTONOMOUS SECURITY SYSTEM BASED ON THE FINGERPRINT SCANNER MODULE AND ARDUINO PLATFORM: A STUDY CASE

The basic premise of biometric authentication (the term is derived from the Greek word “bio” meaning life and “metric” meaning to measure) is that every person is unique and each individual can be identified by his or her intrinsic or behavior traits. Biometric technology is able to recognize a person on the basis of the unique features of their face, fingerprint, signature, DNA or iris pattern and then impart a secure and convenient method for authentication purposes [1-7].

Biometrics is therefore the measurement and statistical analysis of a person’s physical and behavioral characteristics. For example, voice recognition systems work by measuring the characteristics of a person’s speech as air is expelled through their lungs, across the larynx and out through their nose and mouth.

The speech verification software will compare these characteristics with data already stored on the server and if the two voiceprints are sufficiently similar, the biometric security system will then declare it a match. In this paper, we proposed a simple autonomous security system that based on the fingerprint scanner module and Arduino Uno.

Arduino board was designed in the Ivrea Interaction Design Institute intended for students without a background in electronics and programming concept. This board started altering to adapt to new requirements and challenges, separating its present from simple 8-bit boards to products for IoT (Internet of Things) applications, 3D printing, wearable, and embedded surroundings [8]. All boards are entirely open-source, allowing users to build them separately and finally adapt them to their exact needs. Over the years the Arduino boards has been used to build thousands of projects, from daily objects to compound scientific instruments.

The Arduino integrated development environment is an environment in which an Arduino board can be programmed. A written program or code is called SKETCH. In this work, the Arduino IDE is used as an environment in which the Arduino Uno program is written, compiled and uploaded on the Arduino board. The Arduino can be

connected to a computer through the USB port and programmed using a language similar to C++. The connection scheme is quite simple and it's rather difficult to make a mistake (Figure 1).

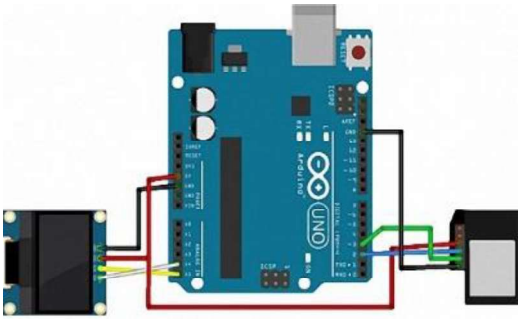

Figure 1 – The connection scheme

References:

1. Elmir, Y., Ghazaoui, O., Boukenni F.: Multimodal Biometrics System's Resistance to noise (Fingerprint and Voice). CEUR Workshop Proceedings **942**, 25-28 (2012)
2. Xia S., Liu Y., Yuan G., Zhu M., Wang Z., Indoor fingerprint positioning based on Wi-Fi: an overview, ISPRS Int. J. Geo-Inf. 6, 135, (2017)
3. Suining H., Gary Chan S. H., Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons, IEEE Commun. Surv Tut, 18, 466-490 (2016)
4. G. Li, C. Busch and B. Yang, "A novel approach used for measuring fingerprint orientation of arch fingerprint," 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2014, pp. 1309-1314. doi: 10.1109/MIPRO.2014.6859770
5. S. P. Sandip and P. H. Zope, "Selective review of fingerprint enhancement, classification and matching techniques," 2015 IEEE Bombay Section Symposium (IBSS), Mumbai, 2015, pp. 1-6. doi: 10.1109/IBSS.2015.7456656
6. Caso G., De Nardis L., On the Applicability of Multi-wall Multi-floor Propagation Models to WiFi Fingerprinting Indoor Positioning, Future Access Enablers for Ubiquitous and Intelligent Infrastructures: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, (2015)
7. Tomas Trainys, Algimantas Venčkauskas. Encryption Keys Generation Based on Bio-Cryptography Finger Vein Method. CEUR Workshop Proceedings **2145**, 106-111 (2018)
8. Rusyn V, Subbotin S., Sambas A.: Analysis and Experimental Realization of the Logistic Map Using Arduino Pro Mini. CEUR Workshop Proceedings **2608**, 300-310 (2020).

[1] **Anastasiia Shved**
Bachelor's degree, Student
[2] **Sergii Buchyk**
Doctor of Technical Sciences, Professor
[1,2] *Taras Shevchenko National University of Kyiv*

BASIC APPROACHES TO PERSONAL DATA PROTECTION IN CLIENT RELATIONSHIP MANAGEMENT SYSTEM

Abstract. Customer Relationship Management (CRM) systems based on the storage of customers' personal data and their processing in the system. Ensuring proper protection of this data is the most important stage in the development, implementation and enforcement of this system. Modern CRM systems are created using various data protection methods, which will be outlined in this work.

Keywords: Customer Relationship Management system, personal data protection, CRM system architecture, logging, access delimitation.

The business processes of any commercial organization are aimed at ensuring its main business goals: making a profit, providing services or goods. It is not a secret that most of a company's profits depend on customers, that why the loss of them means the loss of most of the revenue. Therefore, attracting new customers, awakening existing ones – all this forces companies to spend a lot of resources on their marketing companies. It is also important to know everyone's needs, to be able to offer exactly what person interested. Only this approach will provide an effectively customers' attract. Small businesses, can save the entire customer base, for example, in Excel. For companies with a large amount of customer data and complex business processes, this will not be enough. In addition, maintaining a large customer base in such programs can lead to loss or damage some data, because of fact that they don't provide any protection of information, and all security is assigned to access to the user's workplace.

The Customer Relationship Management (CRM) system allows storing customers' personal data in a convenient way, providing complex connections between them. In addition, modern CRM systems allow automating many processes of the enterprise, which will work out depending on the characteristics of each client and his needs. [1] Such automation can significantly speed up the customer service process – in today's pandemic realities this is a very big advantage – the speed of work minimizes the manifestations of large queues and crowds of people in one room. Great opportunities for the use of CRM systems determine their growing popularity and necessity.

However, usually, no matter how extensive the capabilities of modern CRM systems, companies seek to customize it to their needs. Customers who use CRM systems, in most cases, pay attention to expanding its capabilities, and security issues are left to the developers of the platform. But it should be noted that the storage of customers' personal data in one place is the main advantage, but at the same time the disadvantage of Customer Relationship Management systems – having access to the

system and its main components provides an access to customers' personal data. Therefore, it is necessary to understand the basic capabilities of protection methods in CRM systems in order to build the right vector for the development of personal data protection system in the CRM system.

On the example of the architecture of CRM system Creatio we will consider the main approaches to data protection in the CRM systems. [2]

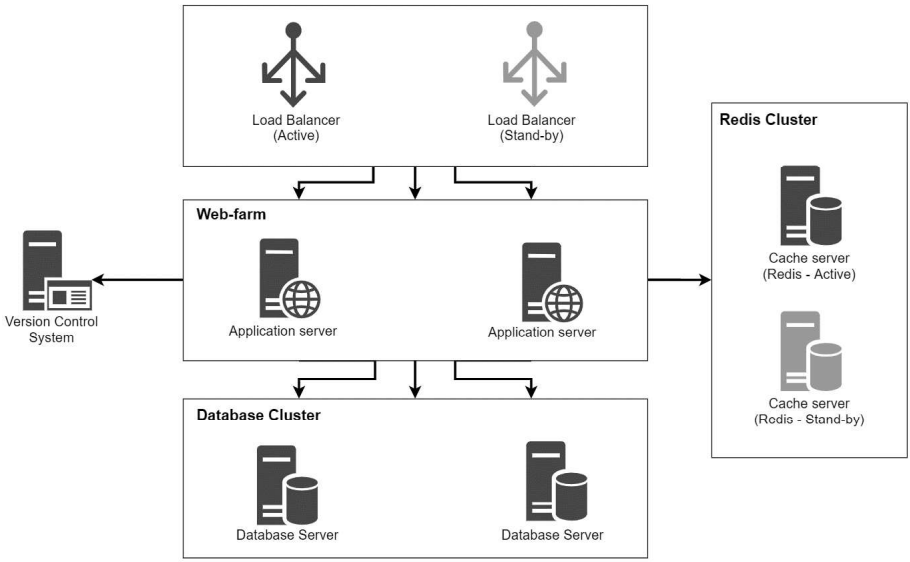

Figure 1 – Customer Relationship Management system architecture

According to the architecture of the presented CRM system, it is necessary to provide protection at the following levels: authorization to the system, work with personal data and security of data transfer between the main components of the system.

Access to any secure system begins with entering login and password – this allows to identify the user who plans to work in the system and to authorize the user in the system. An effective way to protect data at the application server level is to integrate CRM systems with security systems of the database management system (DBMS). With this approach, a separate record is created for each user in the system in the corresponding database table, and all rights are distributed at the database level.

Anyone in the enterprise's local network can get an access to the application by following the link. In case of the databases, only an employee with the role of system administrator, database administrator or security administrator can get an access to database tables, which minimizes the risk of login and password theft. In addition, user passwords are usually encrypted with a 128-bit key and stored in the database in an unreadable form. [3]

As mentioned above, companies seek to expand the functionality of the CRM system to their needs. This uses a variety of web services, which in turn connect using the SSL protocol that allows using public key encryption to authenticate and encrypt client-server connections. [4]

Virtually any CRM implementation project is carried out with the initial configuration of access rights for individual users or groups of users within the system. System administrators can configure access for the entire table, as well as for individual records or fields in the table. Rights are distributed at the database level, so it is important to protect database servers most effectively, because gaining direct access to the database gives an access to all information in the system.

Another good tool for recording system changes is logging, which allows recording the fact of adding, changing or deleting data, the time of these actions and the users who make them. In addition, this tool allows tracking the actions that may preceded the threat. The collection of statistics can prevent the realization of the threat of copying data, or their change: one suspicious action can be explained by human factors, two – by coincidence, three similar actions are the reason to pay more attention to these actions.

After analyzing the approaches to data protection in CRM systems, we can conclude that the basic methods of personal data protection meet the basic needs of the enterprise. However, to achieve a higher level of protection, it is necessary to refine the protection system, using, for example, two-factor authorization by phone number, allocation of Demilitarized Zone, development of additional processes for tracking employees and preventing third parties from accessing system components. All of these methods will be considered in next papers.

References:

1. Efromeeva Elena Valentinovna, Lelaev Magomed Isaevich, Efromeev Nikolay Maksimovich The relevance of the implementation of CRM systems // Problems of Science. 2016. No. 8 (50)

2. Creatio Customer Relationship Management system documentation. URL: https://academy.terrasoft.ua/documentation

3. Shpilevoy DI Regulation of data access in Oracle DBMS // Izvestiya SFU. Technical science. 2008. No. 8

4. Bekker Mikhail Yakovlevich, Terentyev Andrey Olegovich, Gatchin Yuri Armenakovich, Karmanovsky Nikolay Sergeevich Using digital certificates and SSL / TLS protocols for data encryption in cloud computing // Scientific and technical bulletin of information technologies, mechanics and optics. 2011. No. 4 (74)

[1] **Lada Slipachuk**
PhD student
[2] **Serhii Toliupa**
Doctor of technical sciences, professor
[1,2] *Taras Shevchenko National University of Kyiv*

SYNTHESIS FEATURES OF FUNCTIONAL MODEL OF INTEGRATED INDUSTRY MANAGEMENT SYSTEM OF NATIONAL CYBERSECURITY

As of today, a number of issues concerning constructing functional models of industry integrated automated IT, providing general industry management based on integrated automated IT, and ensuring industry integration in functional modeling is still completely uncovered.

The national cybersecurity sector is a strategically important but problematic area for the state. Gradually developing and improving, the national cybersecurity system looks for and builds new trend-perspective mechanisms for its improvement:

— by solving the problems of good governance;

— based on the development and implementation of new management tools and technologies.

Therefore, providing computer-integrated management of the national cybersecurity sector based on an integrated industry-specific information control system is of strategic importance for the subject area of IT [1].

Setting the research objective. The problems outlined above determined the objective of our scientific research, which is the substantiation of features of a functional type of model creation of a branch integrated system of management of the sector.

The topic of the presented research is quite relevant at the state level because:

— it is linked closely to the common state needs and industry problems;

— it is aimed at the creation of a functional model of a sectoral integrated management system, which is a modern tool for improving the management of the national cybersecurity sector.

Analysis of recent research and publications, which started the elaboration of the problem.

Peculiarities of engineering of integrated management systems of the national cybersecurity industry were studied by S. Toliupa, V. Nakonechnyi, L. Slipachuk and other scientists [1-5].

Methodological apparatus of the research. In order to build the model a system-integrated approach for structural and functional filling of the model was used.

Summary of the main research material.

The format of representation of the genesis essence of the functional model was decided to be fixed by the following argumentation:

1. Due to a pragmatic and rational approach when designing the model, it was decided to dwell on the functional type of model.

2. The functional model type:

— is at most applied, informative, dynamic, effective, and flexible;

— meets the requirements that are advanced to its functional capacity of the system;

— clears up functioning mechanisms [2].

At the engineering stage, functional modeling of industry-integrated control systems required:

— an in-depth study of the structural and functional features of the industry macro-object that needs management.

— adequate consideration of the needs, tasks, goals, objectives and specifics of the national cybersecurity sector as a problem area.

The design process of functional model engineering:

— provided a high degree of scientific and methodological reasoning for research and design work;

— was focused on significant aspects concerning the specifics of the national cybersecurity sector;

— separated everything secondary, minor and insignificant, which does not affect the achievement of the goal;

— singled and took into account only the functional aspects [3].

Let us consider in more detail the typological features of the model [2].

1. The functional model was based on certain functional aspects that not only provide operational efficiency of the control system but also allow displaying:

— functions performed by each component of IT as a control system;

— processes occurring in the system;

— inputs to the system and outputs from the system;

— the behavior, mode of action, and properties of IT as a control system;

— information flows and control signals;

— closed loops of work cycles;

— functional load of working modules;

— dynamics of everything that occurs within the model;

— all interconnections and interoperability of parts in the functioning of IT as a control system;

— functional properties of a control system;

— functional and industry capabilities embedded in the simulated system [4].

2. Architectural basis of the system is working modules, as software subsystems, united by a single task in the software and technology chain.

3. Working modules as working subsystems have defining characteristics. They:

— perform specific tasks that go beyond goals and objectives;

— implement the requirements that are imposed on them;

— carry multifunctional workloads (manage modes, communications, data exchange, processes, informatization, analytical data processing, managerial decision-making);

— play an applied role as regards the integrated provision of centralized

management of the national cybersecurity sector;

— work according to their own principles of functioning [5].

This feature of the structure allowed to ensure the functional industry and management capacity of the system.

Obtained scientific result and value of the research. The presented paper discloses:

— genesis essence of the functional type of the model of the branch integrated system of management of the national cybersecurity sector;

— features of the creation of a functional type of model of branch integrated system of management of the national cybersecurity sector;

— elements of scientific research aimed at solving the problems related to the management of the national cybersecurity system on the basis of a functional model of an integrated sectoral information system for management;

— reasons for choosing a functional type of model of the integrated sectoral information system for management.

Conclusions. The materials of this article prove that the functional type of the model of the integrated branch information system of management of the national cybersecurity sector is the most appropriate choice.

Prospects of further research. Further research advisable to focus on the creation and implementation of integrated sectoral management information systems and their models for other sectors of the military defense or industrial complexes of the state.

References:

1. Slipachuk, Lada. (2020). Relevance of the integrated information management system as a subject of management of the national cyber security sector of Ukraine. The synergetic concept. Journal of Physics: Conference Series. 1454. 012002. 10.1088/1742-6596/1454/1/012002.

2. Slipachuk, Lada & Nakonechnyi, Volodymyr. (2019). Typology of the Model of Integrated Sectoral Information System of the National Cyber Security Management. doi: 271-276. 10.1109/ATIT49449.2019.9030595.

3. Lada Slipachuk. Design principles of the functional model of the integrated industrial MIS by the national cybersecurity sector // International science journal "Polish Science Journal" (ISSUE 3(24), 2020) – Warsaw: Sp. z.o. o. "iScience", 2020. – pp. 39-46. - ISBN 978-83-949403-4-8.

4. S. Toliupa, L. Slipachuk Applied aspects of the integrated industry information system management model for the national cybersecurity sector // 2rd International Scientific-Practical Conference on Problems of Cyber Security of Information and Telecommunication Systems (PCSITS) – 2019, pp. 305-308.

5. Lada Slipachuk. Functional principles of the model of the integrated industrial MIS by the national cybersecurity sector // Actual scientific researches in the modern world. Scientific publications Journal – No. 3 (59) Part 1, 2020. - pp. 75-85. - ISSN 2524-0986.

[1] **Mykola Stetsiuk**
Postgraduate Student at the Computer Engineering & System Programming Department
[2] **Andrii Nicheporuk**
Candidate of Technical Sciences (PhD), Associate Professor at the Computer Engineering & System Programming Department
[3] **Bogdan Savenko**
Master Student at the Computer Engineering & System Programming Department
[1-3] *Khmelnitsky National University*

ENSURING THE FAULT TOLERANCE AND SURVIVABILITY OF SPECIALIZED INFORMATION TECHNOLOGIES IN CORPORATE COMPUTER NETWORKS UNDER THE INFLUENCE OF MALICIOUS SOFTWARE

Fault tolerance and survivability [1] define one goal – to ensure high efficiency of IT, which is achieved in different ways. One of its parameters is the time of unavailability, i.e. the time when the system is unable to perform its functions within the requirements for it. For different systems, this time is different and ranges from zero to a certain, still acceptable value. For specialized IT, which operate in corporate computer networks and perform the function of information support in such a highly specialized subject area as financial and economic activities in various fields of application, this parameter is much higher than zero, but the requirements for such IT are also quite high, especially constant growth of their quantitative parameters of functioning (increase in the number of users, complexity of information flows and volumes of processed data) and work in the conditions of influence of malicious software.

The survivability of the developed IT is provided by: redundancy of the server part of IT with territorial diversity of the main and backup server, the feature of redundancy is that the server function, at a critical moment, takes over the mirror SQL server, which in normal mode provides FTP-server; redundancy of client software, the feature of which is that the reserve is not a dedicated computer, and the performance reserve of individual client computers, which, according to the redundancy plan, is installed software of the client, which is in critical the moment will be used as a regular, without losing the functionality of IT.

Fault tolerance of its client part of IT is ensured by performing a set of measures, which includes in addition to traditional hardware redundancy and functional redundancy: organization of automatic updating of system and application software of client PCs by monitoring its relevance with a given frequency; algorithms of procedures that implement critical functions of the client part of IT, with the inclusion of a non-trivial (intelligent) error handling unit, which is performed in parallel with the procedure itself; use of non-trivial data editors, which include in their algorithm an interactive procedure that eliminates uncontrolled manipulation of database data by the operator; implementation of critical for the use of resources, calculation procedures with the ability to quickly select the place of their execution, which prevents overloading of the IT hardware platform.

Maximization of criteria for fault tolerance and survivability in IT configurations based on the client-server architecture can be achieved for each of the parts of the system separately. Then the function $f_1(S_i), i = 1, 2, \dots, n$ determination of fault tolerance in computer systems in quantitative form will look like:

$$f_1(S_i) = \frac{T_{f_1(S_i),1}}{T_{f_1(S_i),1} - (T_{f_1(S_i),2} + T_{f_1(S_i),3})}, \quad (1)$$

where i is the number of components of specialized IT, $i = 1, 2, \dots, n$, $T_{f_1(S_i),1}$ – time between adjacent failures; $T_{f_1(S_i),2}$ – the time required to detect the failure and find a way around it; $T_{f_1(S_i),3}$ – time required to recover IT after failure.

The task of structural survivability analysis requires the definition of: the system architecture required to fulfill the purpose of IT operation at some point or time when adverse effects on the system occur. We define the function $f_2(S_i)$,, in which $i = 1, 2, \dots, n$, the definition of survivability in quantitative units in computer networks is presented as follows:

$$f_1(S_i) = \frac{T_{f_1(S_i),1}}{T_{f_1(S_i),1} - (T_{f_1(S_i),2} + T_{f_1(S_i),3})}, \quad (2)$$

where $T_{f_2(S_i),1}$– the time of operation of IT processes in standard mode, $T_{f_2(S_i),2}$ – the time spent on the processes of survivability, $i = 1, 2, \dots, n$.

This definition of the survivability function makes it possible to display the standard mode of operation with a unit value, and if there is a need to ensure survivability and in the case of a much longer time than the standard mode of operation, the function value will display a quantitative ordinal value.

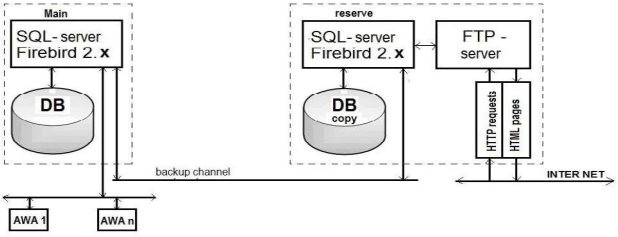

Figure 1 – The scheme of ensuring the survivability and fault tolerance of the server part of the IS.

An important area of further research to improve the effectiveness of IT is to develop a method to ensure effective protection of information directly in the structure of IT and computational processes that take place under the influence of malware.

References:

1. DSTU 3396.2-97 Zaxy`st informaciyi. Texnichny`j zaxy`st informaciyi. Terminy` ta vy`znachennya [DSTU 3396.2-97 Information protection. Technical protection of information. Terms and definitions]

[1]**Serhii Toliupa**
Doctor of Technical Sciences, professor
[2]**Mykola Brailovskyi**
Candidate of Technical Sciences, Associate Professor
[3]**Ivan Parkhomenko**
Candidate of Technical Sciences, Associate Professor
[4]**Bohdan Zhurakovskyi**
Doctor of Technical Sciences, professor
[1,2,3]Taras Shevchenko National University of Kyiv
[4]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

SAFETY OF CRITICAL FUNCTIONS INFRASTRUCTURE

Abstract. The questions of conceptual and regulatory bases of critical infrastructure's information security are considered, and the critical infrastructure objects cyberpower level assessment under the conditions of external cybernetic influence is carried out. The strategy, that allows us to obtain a quantitative assessment of the critical infrastructure protection level from the risk of external cybernetic influence and to establish requirements for the cybernetic security systems formation with a set of measures aimed at increasing the security level of these objects, is proposed.

Keywords: critical infrastructure, objects of critical infrastructure, threats to critical infrastructure, protection of critical infrastructure, cyberspace, cyberattack, cybersecurity, cyberpower.

Introduction

In many countries, the concept of critical infrastructure is being implemented, which allows us to focus on systems, networks, and individual objects, the destruction or disruption of which will have serious negative consequences for national security. As the world experience shows, the process of establishing a legislative framework in the field of critical infrastructure protection is rather laborious and long-lasting. Legislation on the protection of critical infrastructures in different countries is often uncoordinated, in addition, there are certain problems with the mechanisms of assigning objects to critical infrastructure. Each country defines critical infrastructure by considering its specificity, the criticality of individual sectors, and the importance of certain services for the society and the state's security [1].

Analysis of known research and problem statement.

Some countries have tried to define the critical infrastructure and develop a strategy for its protection. The list of vital (critical) infrastructures is different for each country and is determined according to its traditions, social and political beliefs, as well as geographical and historical characteristics of each country [2].

After researching scientific publications and analytical materials related to the international experience of the formation and implementation of the critical infrastructure protection system, it can be concluded that the organization of measures

concerning critical infrastructure protection in different countries is implemented in different ways. In some countries, an organizational model is defined and structured, and measures are focused and systematic, while other countries' models have a non-systematic nature when measures are carried out in an informal manner [3].

The purpose and objectives of the research

To propose a strategy that will provide a quantitative assessment of the critical infrastructure objects protection level from the risk of external cybernetic influence and to establish requirements for the formation of cybernetic security systems with a set of measures aimed at increasing the protection level of these objects.

1. Usually, critical infrastructure includes life support systems (water and heat supply) of megalopolises, high-speed and government communication channels, central authorities, power and transport trunk networks, oil and gas pipelines, seaports, emergency response services, and emergency services to the population, high-tech enterprises and military-industrial complex enterprises [4].

That is why the adoption of measures to formulate a cybernetic security policy at the state level should be a priority task of the political leadership in any country.

2. Cybernetic security has become one of the most important components of any country's national security. The maintenance of a country's optimal state is impossible without the development of a national system based on a tolerant attitude to the norms and international law principles, the protection of the primary values determined by the current legislation, and also the national interests of safeguarding cyberspace.

3. Relying on the experience of scientists, it is impossible to disagree with the fact that the national system of cybernetic security should be the engine of its subjects' interaction in order to unite the special services, state, and law enforcement authorities, which regulate the field of telecommunications and information security.

The main goal of governance in this area should be the development and application of all possible methods for timely detection, cessation, and prevention of cybernetic nature threats.

Tasks such as assessing the protection level of Critical Infrastructure Objects against the risk of third-party cyber impact belong to the multi-criteria class. For their collegial solution under conditions of uncertainty and conflict among the existing methods of mathematical modeling, methods of formation and research of generalized quality indicators using graph analytic and similar approaches, expert methods for solving complex tasks of evaluation and choice of any objects, including special purpose objects, as well as analysis and situations forecast with a large number of significant factors, the most rational and determinant are the expert methods [5]. They provide an opportunity to explore more deeply the phenomena that significantly affect the protection level of both the state as a whole, as well as its individual information and cyberinfrastructure objects against the influence of internal and external cybernetic interventions and threats, to identify the most important and significant in these processes, without omitting those details and interconnections, without which the model of the problem under study cannot be constructed.

Conclusions

Thus, the proposed strategy will provide an opportunity to get a quantification of the OCI protection level from the risk of the outside cybernetic influence, to set the organizational requirements of their own cybernetic security systems, and to work out measures aimed at increasing their effectiveness. The reason for such actions can be the detection of deviations from the normal mode of IP, IT systems and networks, functioning, as well as respective software and hardware.

References

1. Slipachuk, L., Toliupa, S., & Nakonechnyi, V. (2019). The Process of the Critical Infrastructure Cyber Security Management using the Integrated System of the National Cyber Security Sector Management in Ukraine. In 2019 3rd International Conference on Advanced Information and Communications Technologies, AICT 2019 - Proceedings (pp. 451–454). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/AIACT.2019.8847877

2. Toliupa, S., Parkhomenko, I., & Shvedova, H. (2019). Security and regulatory aspects of the critical infrastructure objects functioning and cyberpower level assessment. In 2019 3rd International Conference on Advanced Information and Communications Technologies, AICT 2019 - Proceedings (pp. 463–468). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/AIACT.2019.8847746.

3. Dovgan, O.D. Critical infrastructure as an object of protection against cybernetic attacks / O. D. Dovgan // Information security: challenges and threats of the present: materials of the scientific and practical conference., April 5, 2013 - K.: P. 17-20.

4. Makhutov N., Petrov V., Reznikov D. Analysis of Critical Facilities and Infrastructures in Russia /2nd International Disaster and Risk Conference, Davos, 2017. P. 118–119.

5. Pederson P., Dudenhoeffer D., Hartley S., Permann M. Critical Infrastructure Interdependency Modelling: A Survey of U.S. and International Research / Idaho National Laboratory Critical Infrastructure Protection Division Idaho Falls, Idaho 83415. 2016.

[1]**Serhii Toliupa**
Doctor of Technical Sciences, professor
[2]**Serhii Buchyk**
Doctor of Technical Sciences, professor
[3]**Yanina Shestak**
Candidate of Technical Sciences
[4]**Andrew Kulko**
Master of Cyber Security
[1,2,3,4]*Taras Shevchenko National University of Kyiv*

CYBERATTACK DETECTION SYSTEMS BASED ON THE SIGNATURE METHOD

Abstract. The presence of important information in the functioning of the systems and objects of critical national infrastructures enables its usage by the negatively-minded elements and groupings for the implementation of unlawful actions in the cyber space by violating the integrity, availability and confidentiality of information, and inflicting damage on information resources and information systems. The purpose of the material is to develop a system for recognizing cyber threats based on signature analysis, which would reduce the time of detection of an attack of a cyber defense system while the number and complexity of cyber attacks are increasing.

Keywords: cyberspace, cyber attack, decision-making system, cyber intrusion.

Introduction

Possibilities of the cyberspace, rapid development and implementation of leading-edge information and telecommunication technologies provide the unprecedented opportunities for accumulation of data and its usage. The presence of important information in the functioning of the systems and objects of critical national infrastructures enables its usage by the negatively-minded elements and groupings for the implementation of unlawful actions in the cyber space by violating the integrity, availability and confidentiality of information [1]. The purpose of the is to develop a system for recognizing cyber threats based on signature analysis, which would reduce the time of detection of an attack of a cyber defense system while the number and complexity of cyber attacks are increasing.

Main part

In general, modern systems of intrusion and cyber attacks detection are far from ergonomic and effective solutions, according to the security. But the improvement of efficiency should be considered not only in the sphere of detection of improper activities on the infrastructure of secure information objects, but also according to everyday exploitation of these measures and to the saving of computing power and information resources of an owner of a security system.

The most widespread cyber threats to information resources can be considered as potentially possible cases of natural, technical or human-induced nature, which may lead to unwanted effects on the information system, as well as on the information

stored therein. The emergence of a cyber threat, that is finding the source of actualization of certain events in the threat, is characterized by such an element as vulnerability. By integrating a variety of approaches, as well as suggestions for solving this issue, we believe that the following kinds of cyber threats to information security can be identified: disclosure of information resources; violation of their integrity; failure of the equipment itself [2].

All developers of attack detection systems and organizations that use CADS should understand and study their classification in order to choose the best solutions for information security systems. In the study of various aspects of taxonomy and the application of various options, we can achieve a higher level of security of information systems.

The systems for detecting abnormal behavior are based on the fact that CADS has some features that characterize the correct or permissible behavior of the object of observation.

As the world experience has showed, the most effective methodological approach for constructing of innovative intellectual cyber attack monitoring systems is the way to create a hierarchical multilevel structure of cyber attack detection at the beginning of their implementation [3]. Furthermore, a hierarchical approach allows to solve difficult problems of the information protection process managing from cyberattacks in the distributed information systems (IS) as sequence of local tasks, coordinated with each other.

Let's consider one of the effective methods of detecting intrusions and cyber attacks, which is based on the signature approach. Signatory methods allow you to describe a cyber attack with a set of rules or using a formal model, which can be used as a character string, semantic expression in a special language, etc. The essence of this method is to use a specialized database of templates (signatures) of cyber attacks to find actions which fall under the definition of “cyberattack”[4].

The signature method can protect from a viral or hacker cyber attack when its signature is already known (for example, the unchanged fragment of the body of the virus) and it is included in the database of CADS. If the network is experiencing the first attack from the outside, the first infection is still unknown, and the database simply lacks the signature for its search - the signature method CADS will not be able to signal the danger because it considers the attacking activity to be legitimate.

Most of the existing software products which claim to use the signature method, in fact, realize the most primitive way of signature recognition. In such systems, the signature method is implemented as an algorithm that examines only the dynamics of cyberattack development [5]. And it is based on a state machine to assess the scenario of the developing attack. According to the plan, this approach should allow tracking the dynamics of the development of cyber attacks in accordance with the actions of the intruder, while as the module for data collection even the systems for detecting cyber attacks can be used.

Conclusions

Thus, the effectiveness of the signature CADS is determined by three main factors: the efficiency of refinement of the signature base, its completeness from the

point of view of the determination of the signature of the cyber attack, as well as the presence of intelligent algorithms for reducing the attacking party's actions to some basic steps, within which there is a comparison with the signatures.

In order to implement the chosen method of determination and identification of CADS, models of the signature and statistical analyzers of network traffic are offered,and the fuzzy intellectual system is used to determine the sources of cyber-media and the choice of solutions for their elimination [6].

Most of the existing software products which claim to use the signature method, in fact, realize the most primitive way of signature recognition. In such systems, the signature method is implemented as an algorithm that examines only the dynamics of cyberattack development. And it is based on a state machine to assess the scenario of the developing attack. According to the plan, this approach should allow tracking the dynamics of the development of cyber attacks in accordance with the actions of the intruder, while as the module for data collection even the systems for detecting cyber attacks can be used.

In order to implement the chosen method of determination and identification of CADS, models of the signature and statistical analyzers of network traffic are offered,and the fuzzy intellectual system is used to determine the sources of cyber-media and the choice of solutions for their elimination.

Reference:

1. Toliupa S., Parkhomenko I Signature and statistical analyzers in the cyber attack detection system. Scientific and Practical Cyber Security Journal (SPCSJ) № 3 (02). с. 47-53

2. Amer, S.H., Hamilton, J.A., “Intrusion Detection Systems, (IDS) Taxonomy – A Short Review,” DOD Software Tech News, vol. 13, no. 2, June 2010, DOD Data & Analysis Center for Software, Air Force Research Laboratory, Rome, N.Y., pp. 23 – 30.

3. Toliupa S., Nakonechnyi V., Uspenskyi O. Signature and statistical analyzers in the cyber attack detection system. Information technology and security. Ukrainian research papers collection Volume 7, Issue 1 (12). с. 69-79.

4. 17. IDS / IPS. Netgate Documentation: [website]. Washington: Rubicon Communications LLC, 2017. [Electronic resource]. Online: https://www.netgate. com/docs/pfsense/ids-ips/.

5. Ghahramani, Z. An Introduction to hidden Markov models and Bayesian networks / Z. Ghahramani // International Journal of Pattern Recognition and Artificial Intelligence — 2001. — Vol. 15. — P. 9–42.

6. Barbara D. Detecting novel network intrusions using Bayes estimators / D. Barbara, J. Couto, S. Jajodia, N. Wu. // In: Proc. of the 1st SIAM International Conference on Data Mining. — 2001.

[1] **Serhii Toliupa**
PhD in Technical Sciences, Professor at the Department of Cybersecurity
[2] **Volodymyr Nakonechnyi**
PhD in Technical Sciences, Professor at the Department of Cybersecurity
[3] **Maxym Kotov**
3rd year of bachelor's degree student
[4] **Valeriia Solodovnyk**
3rd year of bachelor's degree student
[1-4] *Taras Shevchenko National University of Kyiv*

SIGNALS ENCRYPTION IN WIRELESS DATA INPUT DEVICES

Cryptography is the science that creates strategies for utilizing complex mathematical changes to transmit data through conveyance channels in a frame that no one but authorized individuals can get. Encryption process is a key object in the field of cryptographic research, it is the method of changing the frame of data which is transmitted through open transmission channels [1-5].

The cipher quality of the encryption is measured by the time that it takes to decrypt the content with brute force which is checking of all conceivable key combinations.

There are two types of symmetric encryption [4, 5]: block encryption and stream encryption. An example of symmetric block encryption algorithm is the AES competition finalist, the American encryption standard - Rijndael [6, 7]. In AES 128 form of the calculation, the cipher key comprises of 128 bits isolated by 16 bytes that are composed to the InputKey matrix. The InputKey comprises 4 columns. Utilizing those columns an arrangement of 44 words (w0 - w43) where each word comprises of 32 bits is shaped. Thus, these words become the round keys.

The AES scheme is shown in Figure 1 [6]:

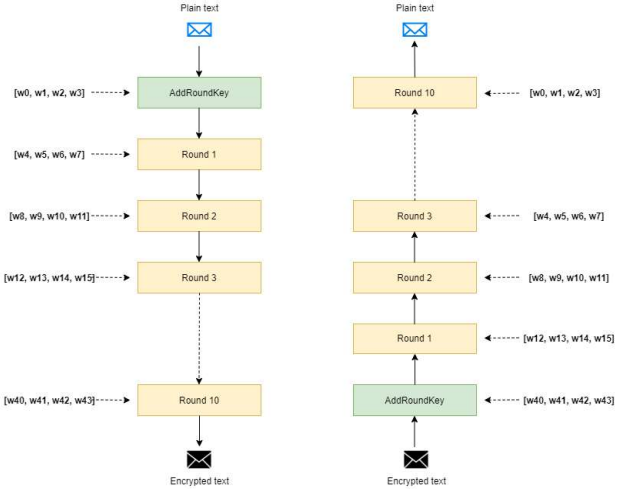

Figure 1 - AES scheme

There are numerous wireless input devices in access at nowadays. As the devices of wireless information input, we consider - the gadget of remote transmission of the entered information to the computer (wireless data input device WDID). Examples of such gadgets are remote consoles or mice, remote headsets, touch input gadgets, etc.

WDID gadgets transmit data through radio waves with a frequency from 27 MHz to 2.4 GHz. WDIDs perform their work with transmitter and a collector. An example of a records acquisition technique is Mousejack.

The essence of this attack is in following stages: a wireless keyboard and a wireless mouse use USB-dongle for signal transmission. a few models of keyboards encrypt those signals, but the majority of mouses do not. In case of keyboards, this works as follows: only USB dongle has the encryption key, which makes it the only item that has the potential to decrypt a signal, an attacker, despite the fact that it intercepts that signal, will no longer be able to decrypt it. But as it was stated most of mouses are not encrypting their broadcast and so the attacker is able to receive unencrypted packets.

The following happens during the first three steps: the user determines the shift (x , y) of the mouse location coordinates, and the transmitter in the mouse transmits the radio signals without encryption to the USB-dongle. At the same time, hackers intercept unencrypted signals using their own personalized USB dongle.

The attacker sends a sequence of requests to connect to the USB dongle during 4-6 stages, then the USB dongle receives the sequence of these requests and connects to the computer of the user.

The attacker sends a series of characters to the user's machine during steps 7-9. If all the stages shown above have succeeded, the hacker has full access to the computer of the victim. This vulnerability can be used on all operating systems since this vulnerability does not apply to operating system vulnerabilities.

References:

1. Dan Boneh, Victor Shoup, A Graduate Course in Applied Cryptography, version 0.4, Stanford University, September 2017.
2. Bruce Schneier. Applied cryptography. Protocols, algorithms, source texts in the C language.
3. S. Toliupa, L Slipachuk,. V. Nakonechnyi. The Process of the Critical In-frastructure Cyber Security Management using the Integrated System of the National Cyber Security Sector Management in Ukraine, 3rd International Conference on Advanced Information and Communications Technologies, AICT 2019 – Proceedings, 2019.
4. William Stallings, Cryptography and Network Security Principles and Practices, Fourth Edition, Prentice Hall, November 16, 2005/
5. Christof Paar, Jan Pelzl, Understanding Cryptography, Springer-Verlag Berlin Heidelberg, 2010
6. Joan Daemen, Vincent Rijmen, AES Proposal: Rijndael, October 1999.
7. Nigel Smart, Cryptography: An Introduction, 3rd edition, Mcgraw-Hill College, December 30, 2004.

[1] **Sviatoslav Tukalo**
Bachelor, MA Department of Telecommunications
[2] **Orest Kostiv**
Senior Lecturer Department of Telecommunications
[3] **Olga Shpur**
PhD, Assistant of the Department of Telecommunications
[4] **Bohdan Buhyl**
PhD, Assistant of the Department of Telecommunications
[1-4] Lviv Polytechnic National University

METHODS DEVELOPMENT TO PROTECT IOT FROM BOTNETS

Modern technologies are evolving very fast. Almost every day we can hear news about new developments in a particular field. One of the fastest growing technologies is the Internet of Things. IoT is integrated into almost all spheres of our life. Also it make demands from developers for new requirements of security standards for this technology. Its implementation directly depends on the security of personal data of users.

Telecommunication company Cisco in its annual report [1] notes that almost 90% of all IoT devices are vulnerable to cyberattacks. The number of such gadgets in the world has become larger than population of the Earth, and the approach to cybersecurity almost equal zero, it becomes obvious that cyber attackers are starting to create IoT-based botnets (Fig. 1).

Fig. 1- The trend of increasing C&C servers

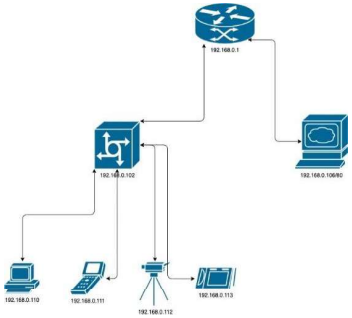

Fig. 2 - Scheme of the experiment

A botnet it`s a typical computer network, that includes in its hierarchy such elements like: botmaster, C&C server and bots. Main principles of botnets are: infection, connection, control and pervasion. Investigation shows, that there are three common and typical Botnet attacks – DDoS, Spam and Brute - force.

For this research botnet with Client-Server architecture was developed. The scheme of the experiment is presented on the Figure 2. The operating system of the attacking host was Kali Linux 64 bit 2020.2 Operating system of the victim's host - Debian 10 64 bit.

For this attack was used usual ICMP network protocol from command ping. The default command view looks like “ping [-AaDdfnoQqRrv] [-c count] [-G sweepmaxsize] destination IP”. As this command is standard for all operating systems, it was initially decided that it was necessary to investigated the possible generated traffic when executing this command. It was explored that the network card processed traffic of 1,31 kbit/s. By using the same configuration for the command ping, but sent at the same time from 4 bots, traffic amounted to 5,25 kbit/s. As a result of the system work, it is established that under normal conditions ping request cannot injure the system, because 100 bots will generate 0.13 Mbit/s traffic, but when changing the packet size and packet sending speed, 100 bots will generate traffic exceeding 77 Mbit/s.

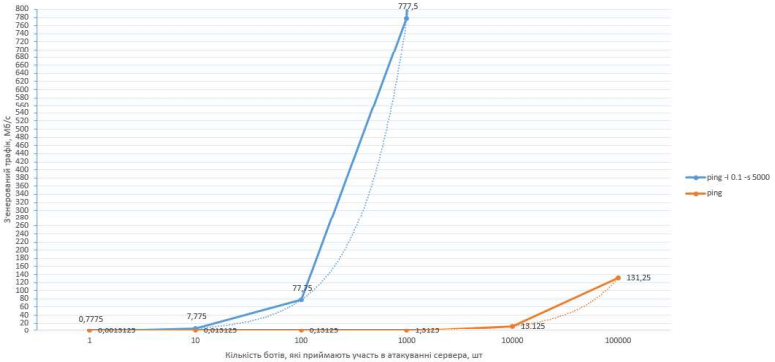

Fig. 3 - Comparative characterization of the generated traffic size using ping commands with parameters and without

By investigating the structure of TCP packet requests, it became clear that the SYN flood can disable the system very quickly, by creating half-open connections. Therefore, a botnet of 100 bots will incapacitate the home WEB server in ~ 3 seconds.

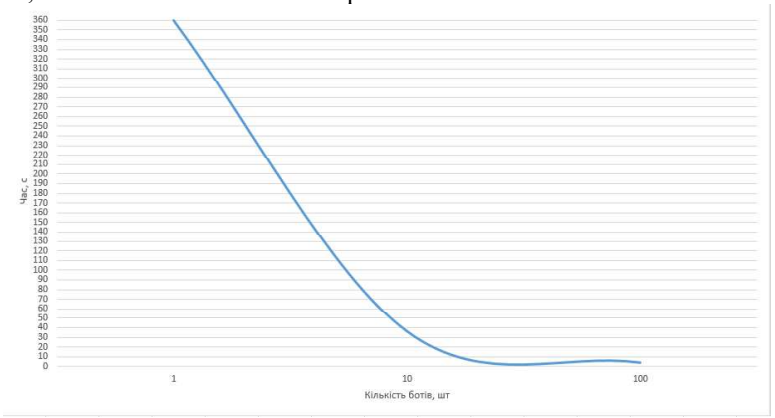

Fig. 4 - Time required for system failure

The basis of the ICMP – flood attack is the system's automatic responses. This means that the first step is ping request disabling.

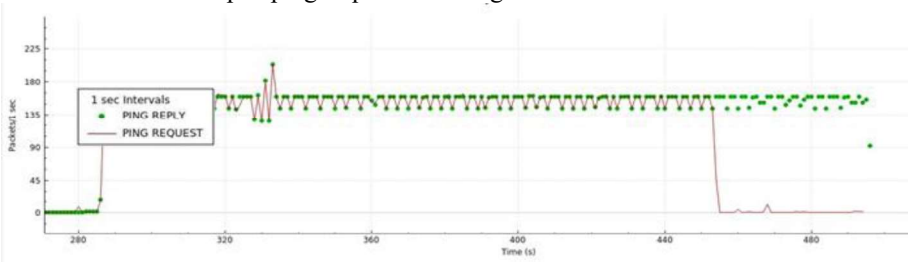

Fig. 5 - The time required to disable the system

Thus, the algorithm for combating the ICMP flood will take the form refer to Figure 6.

Fig. 6 - ICMP flood reflection algorithm

SYN – flood attack can destroy system in a short time. The graph shows that attackwith generated traffic 18 kB / s will destroy system in 90 seconds. The danger is that the network card will stop opening new connections and stop directing traffic to existing ones.

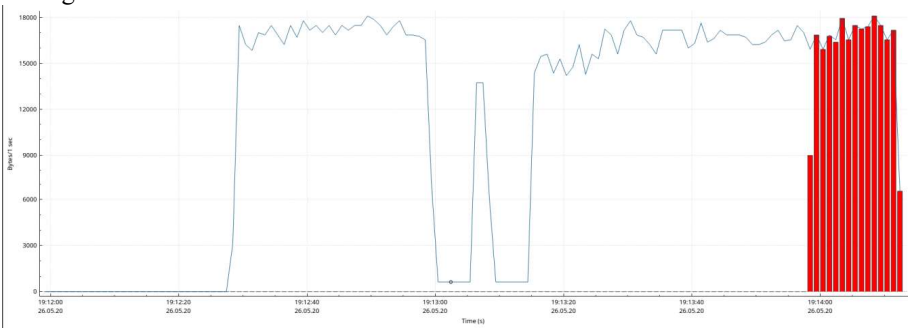

Fig. 7 - SYN - flood traffic generated during the experiment

By changing the system settings (reducing the retention time of half-open connections, syncookies collection and increasing the queue of half-open connections), it was significantly possible to reduce the power of the attack and make it almost safe. Figure 8 shows that the traffic has decreased to 8 kB / s, and there are no more failures to create a new connection.

Thus, the proposed algorithm to struggle this type of attack, turn to Figure 9

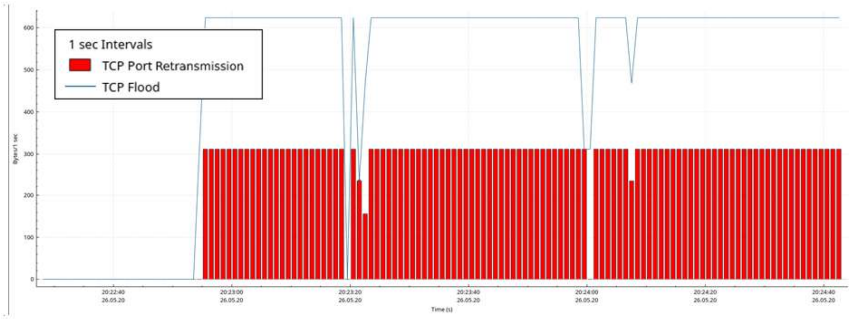

Fig. 8 - SYN - flood traffic after applying the reflection algorithm

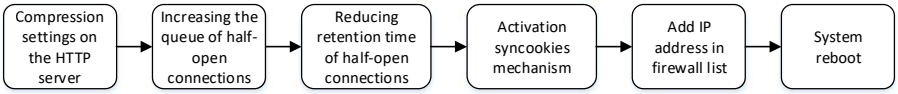

Fig. 9 SYN flood reflection algorithm

Conclusion. In this research presents methods for ICMP and TCP flood attacks repelling. This investigation establishes the time required for the botnet to disable the system and the size of generated traffic. This made it possible to understand the real threat that the bot network is now creating. This research improved algorithm reflection SYN flood by reducing the retention time of half-open connections, syncookies collection and increasing the queue of half-open connections. And also the process of data compression on the web server played an important role in this algorithm improving.

References:

1. Cisco Inc. (2018). Cisco Cyber Security Report 2018 [Electronic resource]. Access mode: https://www.cisco.com/c/uk_ua/products/security/security-reports.html

2. Botnets and their types [Electronic resource] // EC-Council – Resource access mode: https://blog.eccouncil.org/botnets-and-their-types/.

3. Into the Battlefield: A Security Guide to IoT Botnets // Trend Micro. – 2019. – Resource access mode: https:// www.trendmicro.com/vinfo/ph/security/ news/internet-of-things/into-the-battlefield-a- security-guide-to-iot-botnets

4. Water Torture: A Slow Drip DNS DDoS Attack // Secure64 Software Corporation. – 2014. – Resource access mode: https:// secure64.com/2014/ 02/25/water-torture-slow-drip-dns-ddos-attack/.

5. Scanlon M. Peer-to-Peer Botnet Investigation: A Review / M. Scanlon, T. Kechadi // Future Information Technology, Application, and Service, LNEE 179, pp. 231–238, DOI: 10.1007/978-94-007-5063-0_33

6. Almomani A. A proposed framework for Botnet Spam-email Filtering using Neucube // A.Almomani, M. Alauthman, O. Almomani, O. Albalas. // Proceeding of The International Arab Conference on Information Technology (Yassmine Hammamet, Tunisia, December 22-24, 2017). – 2017.

DATA ANALYTICS

[1] **Veronika Bokan**
Master of the Department of Management Technology
[2] **Viktoriia Tsykun**
Master of the Department of Management Technology
[3] **Andrii Khlevnyi**
PhD, Associate Professor of the Department of Technologies Management
[1-3] *Taras Shevchenko National University of Kyiv*

INFORMATION ANALYSIS OF METHODS FOR FORECASTING THE POPULATION OF UKRAINE

Abstract. The urgency of population forecasting is given in the work, the methods of such forecasting are analyzed. The parameters on the basis of which it is expedient to forecast the demographic situation in the near future are identified: the establishment of births, deaths and migration. The expediency of application of methods of neural networks is established, parameters are allocated, prospects of further researches are established.

Keywords: population, forecasting, neural network.

Demographic forecasting is the basis for scientific planning of indicators of socio-economic development of the country, giant companies and local companies. Accordingly, consumers of such forecasts form the transformation of geopolitical processes in terms of creating plans for the production of goods and services, planning in education and medicine and other important areas of life, infrastructure development and housing, and etc. According to the results [1], it is established that in Ukraine there is a rapid decline in population. This depends on: declining population, high mortality, working age, poor health of adults and children, reduced life expectancy and gradual aging, migration.

From the analyzed approaches [2] of demographic forecasting, namely the separation of forecasting by: time, detailing information about the population of Ukraine, regional coverage, purpose, object, it is established that special attention is needed by time forecasts (short-term, medium-term and long-term). Short-term forecasts are considered to cover the period before reaching the age of. The short-term forecast should correspond to high level accuracy and detail. Forecasts, called medium-term, covering a period of twenty to thirty years. Such forecasts are not always the same as in the short-term, but it is possible to show a deeper picture for demographers, noting the prospect of growth or decline of the population and its individual parts with an accepted level of reliability. Long-term forecasts cover fifty or even a hundred years. It is clear that the results of this type of forecast can be operated as scientific predictions. It is established that in terms of economic prospects for the near future it is advisable to pay attention to the development of models and methods of short-term forecasts. The expediency of such a forecast is to plan and achieve economic and cultural goals of companies, the region and the country as a whole. The main indicators on the basis of which it is expedient to forecast the demographic situation in a certain

area for the near future: the establishment of births, deaths and migration.

Short-term demographic forecasting for certain parameters can be implemented by the following methods: regression, cohort-component, analytical, extrapolation, Markov chains, neural networks.

Methods of regression analysis for short-term demographic forecasting. The essence of this approach is to establish qualitative relationships between the factors that shape their intensity. The assessment is based on historical changes in indicators that are considered and act as factor factors and affect the forecast of demographic phenomena. The result of such modeling is the construction of multidimensional regression models, which are the results of the analysis of correlation and regression sets. As an independent variable in such modeling, it is appropriate to use not time, but a numerically defined material characteristic, which is a factor [3]. Therefore, it is appropriate to use this method to determine the forecast by regional coverage.

Cohort-component method. The essence of the method is to form cohorts and determine their changes according to certain parameters, such as age, sex, life expectancy, and so on. This method is most often used in population forecasting. The advantage is that such a number can be considered from the standpoint of the established cohorts. The input data of such modeling are data on the initial number of cohorts. It is appropriate to implement on the basis of the equation of demographic balance.

Analytical method. The essence of the method is to select a function based on historical data that will most accurately describe it in the future. This method should be used for short-term forecasting.

Extrapolation method. The essence of the method is that the calculations are based on exponential and linear functions. Input data contain data on changes in population, both average and absolute, for a particular period or data on changes in the average annual population growth rate. If we assume that the factors or groups of factors that can affect the process of forecasting using the extrapolation method, is unchanged, in order to allow connecting the population to any required time [4]. However, it is advisable to use for short-term forecasting. The method will not change in certain groups of the population. Therefore, it cannot be used to predict, for example, age, working groups.

Markov chains. convenient to use when there is a need to solve the problems of population transition from one cohort to another.

The disadvantages of these methods can be minimized using the neural network approach. The expediency of their application is that in the system of indicators of demographic status linear methods do not cover all the patterns. The task of the neural network is to learn to solve the problem on the basis of a training sample. Such a neural network is able to determine the relationships between the data that enters the input and requires the output signal [5]. The trained neural network is able to summarize the acquired skills and give a forecast for new values of the input signal.

Based on the analysis, it is established that for short-term forecasting of the population of both the region and the country as a whole, it is advisable to use such promising methods as neural networks. Data for training is proposed to be used from

http://www.ukrstat.gov.ua/ from 1989 to 2020. To implement the method, the following parameters are proposed according to which forecasting by neural network methods will be carried out:

- to predict the birth rate: permanent population, migratory population growth, average wages, number of unemployed, number of children who applied to preschool educational institutions, number of doctors, infant mortality rate, indicators of the number of family institutions (marriages and divorces) , state social assistance for children, the number of families of reproductive age who received housing from the state who purchased housing.

- to predict mortality: permanent population, migration growth, life expectancy, number of pensioners, average wages, number of registered crimes, morbidity, number of road accidents, number of suicides, number of deaths in hospitals.

It is established that for the implementation of forecasting it is advisable to choose the Python programming language for simplicity in syntax, broad support of the programming community and a huge amount of available documentation.

The prospect of further research is to build a neural network architecture, its training, testing. This model will be the basis of information technology aimed at combining the analysis and forecast of the population, its labor potential with the socio-economic processes that take place. It will make it easier for stakeholders to draw conclusions that will be aimed at changing the trend of population decline, increasing the socio-cultural level in general.

References:

1. Distribution of permanent population by sex, age groups and type of locality. URL:http://database.ukrcensus.gov.ua/Mult/Dialog/varval.asp?ma=000_0204&path=../Database/Population/02/02/&lang=1&multilang=uk

2. Власенко Н.С., Макарова О.В., Пирожков С.І., та інші. Комплексний демографічний прогноз України на період до 2050 р. / за ред. членкореспондент НАНУ, д.е.н., проф. Е.М. Лібанової. К.: Український центр соціальних реформ, 2006. 138 с.

3. A. Sen, M. Srivastava, Regression Analysis. Theory, Methods, and Applications, Springer-Verlag, Berlin, 2011 (4th printing).

4. George Lindfield, John Penny. Numerical Methods (Fourth Edition), 2019.

5. N.D. Lewis. Neural Networks for Time Series Forecasting, 2017.

[1] **Yuliia Bura**
Master's degree student Department of technology management
[2] **Iulia Khlevna**
Doctor of Engineering Science, Associate professor Department of technology management
[1,2] *Taras Shevchenko National University of Kyiv*

HOUSE PRICE MODELING BY MACHINE LEARNING

Abstract. The objective of this paper is to present the relevance of house price prediction. During the research, we've established that machine learning methods give a fairly accurate result (with a small RMSE error). The aggregate model gives the smallest prediction error.

Keywords: Machine Learning, House Price.

The real estate sector, both in Ukraine and in the world, is one of the sectors with the largest capitalization. Here often appear new profitable projects. These projects play an important role in the formation of capital flows and economic stability. Providing objective information to those who make decisions about certain transactions in the real estate market is an urgent application task. Its solution can be interpreted through the creating of information technology, which is based on the analysis, modeling, and prediction of house prices in the real estate market. The scientific task is the implementation of a rational method of prediction in such technology. This is the task of this article too.

The real estate market is quite sensitive to external events and unpredictable. So, the process of the house price prediction is really complex, and classical methods do not give the desired result in this case [1]. It's expedient to use machine learning methods for such situations of fuzzy variables prediction [2, 3].

To choose a rational method for house price prediction we propose to analyze the following models: *LASSO Regression, Elastic Net Regression, Ridge Regression, Gradient Boosting Regression, XGBoost.*

The study was conducted on a dataset of 1300 houses [4]. The “sales price” was set as the target variable.

We identified 79 independent variables to model the target variable. All data are equally divided into training and test parts.

The result of logarithmic normalization of the target variable for its qualitative modeling is further presented in Fig. 1.

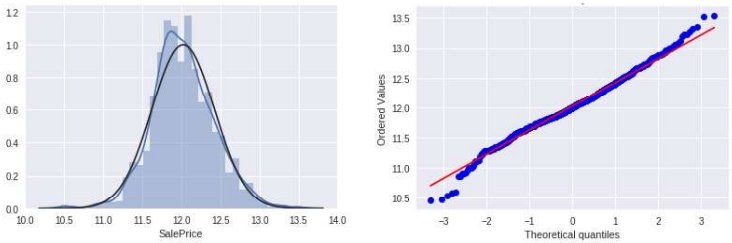

Figure 1. Comparison of distribution graphs of the logarithmically normalized "Sales Price" series with the normal distribution.

The selected models were evaluated for the quality of the model prediction – RMSE error and standard deviation (Table 1).

Table 1

Models prediction results

Model's score	RMSE	Standard deviation
Lasso	0.1115	0.0074
ElasticNet	0.1116	0.0074
Ridge	0.1153	0.0075
Gradient Boosting	0.1177	0.0080
Xgboost	0.1161	0.0079
Averaged base models	*0.1091*	*0.0075*

The gradient boosting model was found to be the least accurate. So, we tried to build an aggregate real estate price forecasting model without it. Obtained result: RMSE = 0.1091, and the standard deviation is 0.0075. This approach improves prediction accuracy by 5%.

Conclusions. We established that better to use a combination of models to obtain the most accurate house price prediction in the real estate market. We determined that such an approach will be the basis for the creating of information technology, and this will be a prospect for further research.

References:

1. YanY. Hui B. (2007). Method for Housing Price Forecasting based on TEII. Methodology Systems Engineering - Theory & Practice, Volume 27, Issue 7, 1-9.

2. Park B., Baeb J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Expert Systems with Applications, Volume 42, Issue 6, 2928-2934

3. Segnon, M., Gupta, R., Lesame, K. et al. (2020). High-Frequency Volatility Forecasting of US Housing Markets. J Real Estate Finan Econ. https://doi.org/10.1007/s11146-020-09745-w

4. Dean C. (2011). Journal of Statistics Education, Volume 19, Number 3

5. Bailey MJ, Muth RF, Nourse HO. 1963. A regression method for real estate price index construction. Journal of the American Statistical Association 58: 933–942.

[1] **Oleksandr Burmistenko**
Doctor of Technical Sciences, Full Professor
[2] **Tetyana Bila**
Candidate of Technical Sciences, Associate Professor
[3] **Volodymyr Statsenko**
Candidate of Technical Sciences, Associate Professor
[4] **Dmytro Statsenko**
Candidate of Technical Sciences, Associate Professor
[1-4] *Kyiv National University of Technologies and Design*

INFORMATION ANALYSIS OF THE BULK MATERIALS CONTINUOUS DOSING PROCESS

Abstract. This work presents the principles of bulk material flows information processing at the plate feeders outlet. The experimental stand structure and the results of data processing are shown.

Keywords: Data processing, data analysis, dosing, bulk materials, plate feeder

Introduction. Bulk materials continuous dosing processes are used in the manufacture of products from polymer materials. The basis of such products is a solid polymer granules mixture with various additives. Mixing complexes are used to obtain them, which include equipment for storing bulk materials (hoppers), dosing devices (feeders) and a mixer [1]. The quality of the mixture is largely determined by the compliance of its percentage composition with the given recipe. Therefore, the accuracy of dosing devices is an important parameter that significantly affects the quality of products.

The movement of bulk materials is a complex discrete process during which flow ruptures, formation of arches, formation of lumps and other phenomena can occur that reduce the accuracy of dosing. The problem gets worse with the continuous operation equipment [2]. Such equipment makes high demands on the accuracy dosing, because bulk material moves in a continuous flow and any deviations lead to the appearance of local changes in the percentage composition of the mixture, which are almost impossible to compensate. The output flow pulsations magnitude allows you to indirectly determine the possible equipment deviations from the specified mode. This information can be used to control both the feeder operating modes and the mixing complex as a whole. These problems determine the relevance of the analysis of the flow bulk materials parameters at the continuous feeders outlet.

The research results. This work discusses plate feeders, which are widely used in industry, because they provide minimal mechanical impact on bulk material and the ability to accurately control its performance. The measurement of the bulk materials flow parameters was carried out using an experimental stand (Figure 1), which included a hopper (H), a plate feeder (F), a flow shaper (S), a strain gauge mass sensor (GS), an analog-to-digital converter (ADC), a microcontroller (MC), a personal computer (PC) and a mixer (M).

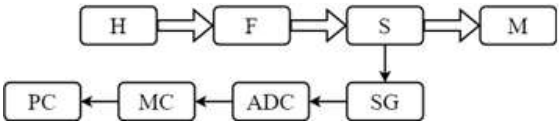

Figure 1 – Stand structure for research of the plate feeders work

A plate feeder transported bulk material to a flow former with a weight sensor installed under the surface. The mass of material determined the sensor signal. Using an ADC, the sensor signals were converted into digital form and read out by the microcontroller at specified time intervals. Then they were transferred to a personal computer for further analysis. An example of the obtained data is shown by dots in Figure 2, a.

Data analysis included two stages:

1. Linear regression equation coefficients calculation passing through the obtained points. The coefficients were determined by using the least squares method. The obtained values are shown in Figure 2, a. Ideally, in steady-state operation, the straight line slope tangent should be equal to zero, and the free term of the equation should equal the specified performance.

2. Spectral analysis of signals using fast Fourier transform (Figure 2, b). The data obtained make it possible to numerically estimate the magnitude of the pulsations and determine possible problems associated with the operating modes of the equipment. For example, the appearance of pulsations with a frequency equal to the plate feeder rotation frequency indicates its incorrect installation.

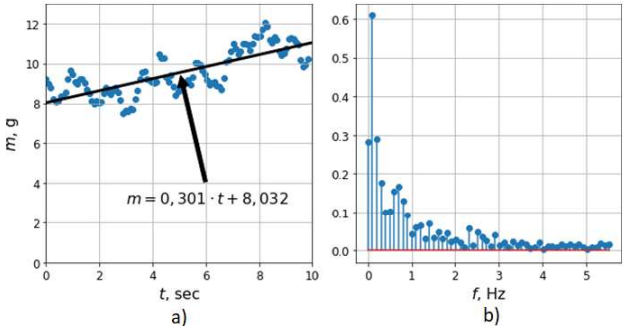

Figure 2 – Sensor Signal Analysis Results

The obtained results can be used as a data source for the dosing equipment control system.

References:

1 V. Statsenko, O. Burmistenkov, T. Bila, D. Statsenko. Determining the motion character of loose materials in the system of continuous action «hopper – reciprocating plate feeder». Eastern-European Journal of Enterprise Technologies. Volume 2/1 (98), 2019, P.21-28. doi: 10.15587/1729-4061.2019.163545.

2 R. Weinekötter, L. Reh. Continuous mixing of fine particles. Part. Syst. December 1994, Volume 12, P.46–53.

[1] **Serge Dolgikh**
M.Sc. Theoretical Physics, Telecommunications Engineering, Senior Project Engineer
[2] **Oksana Mulesa**
PhD, Associate Professor
[1]*Solana Networks, Canada*
[1]*National Aviation University, Ukraine*
[2]*Uzhghorod National University, Ukraine*

COVID-19 EPIDEMIOLOGICAL FACTOR ANALYSIS: IDENTIFYING PRINCIPAL FACTORS WITH MACHINE LEARNING

Abstract. Based on a set of Covid-19 statistical data of national and subnational jurisdictions at the time point of approximately two months after the local onset of the pandemics, an analysis of the factors with strong influence on the reported local outcomes was performed with several different statistical methods. The consistent conclusion of the analysis confirmed epidemiological policy and management as the dominant factors in the outcome. The methods in the study can be used to evaluate principal factors at future time points to reach a confident conclusion.

Keywords: epidemiology, machine learning, regression, factor analysis

Problem Statement

The task of modeling and forecasting time-series processes of different nature is essential and arises in different fields such as planning [1], the study of the dynamics of climate change [2] and importantly in the current situation, health science and epidemiology.

Known models and forecasting methods are based on using integrated information about the background of the predicted processes [1, 3]. Among the tasks of forecasting an important place is occupied by the methods of factor estimation and time-series analysis that includes a variety of methods and approaches including fuzzy sets [4], expert models and methods [5], genetic and neural network methods [6, 7] and others.

A common challenge in the analysis of statistical data related to a developing situation, such as in this work, the developing epidemiological scenario related to a dangerous infection with potentially high impact on health and safety of population, economy and the society as a whole is evaluation of methods and models with the objective of identifying the approaches that could be most effective in describing the process that is being studied.

To avoid or reduce the possible ambiguity related to the selection of the method of analysis of statistical data, in this work we used several common methods of statistical analysis specifically, evaluation and ranking of factor influence with an expectation that if consistency between the results of different methods can be achieved, it would enhance the confidence in the result that can be essential for the development of reliable and effective policies based on the conclusions of factor analysis.

Methods

With variety of statistical methods and techniques used to evaluate the correlation hypothesis as discussed above, we set out to provide an analysis of principal factors influencing the development of the epidemics in the national and subnational jurisdictions based on the available data for the first group of countries that were exposed to Covid-19 pandemics in late January – beginning of February, 2020. This objective is approached by applying several commonly used methods of factor analysis and ranking, looking for consistency of results between different methods. A consistency between the results of different methods would improve confidence in the findings, providing a grounded and reliable statement of their influence on the outcome.

The analysis of scientific publications showed that the following factors have a strong influence on the development of the epidemic including but not limited to the following: the time of the local development of the epidemics; traditions, social and lifestyle factors; demographics including gender and age; the level of the economic and social development; quality standard and epidemiological efficiency of the public healthcare system, and not in the least, the quality of public health policy making and execution. Based on the identified selection criteria and publicly available epidemiological information from a number of trusted sources as indicated below, the dataset of 18 cases of national and subnational public health jurisdictions was constructed. The data included one provincial jurisdiction in Canada (Ontario), one state (California) and one municipal jurisdiction in the USA (New York City); given the high geographical variation of the impacts, data with more detailed geographical breakdown is expected in the future studies. The time point at which the data was collected was *TZ + 3 months*, i.e. approximately two months of the local development of the epidemics in the selected group of jurisdictions.

To evaluate the consistency of results produced by different methods of statistical analysis, several methods were used to evaluate the influence of the selected factors on the overall impact of the developing epidemics:

1. Calculation of correlation between the resulting effect and a specific factor;
2. Linear regression analysis by single factor and a combination of factors;
3. Evaluation of factor importance with Random Forest regression;
4. Evaluation of factor influence or rank with a feature ranking method.

Results

The methods applied to the dataset of early epidemiological data of selected jurisdictions demonstrated consistent results with good agreement between methods. The findings confirmed the importance of clear, timely and evidence-based epidemiological policy as the factor with the highest influence on the development of the epidemiological scenario. This finding is consistently produced by all methods of analysis used in the study.

The analysis offered additional arguments in support of the hypothesis of some form of general population-wide protection effect against Covid-19 as an effect of previous universal immunization program with Bacillus Calmette–Guérin vaccine

(BCG), that has been reported in a number of earlier results adding arguments to the rationale for further studies of the possible correlation and the mechanisms of such general protection with potential benefits that may extend beyond Covid-19 pandemics. Additionally, it established potential significance of secondary factors such as smoking prevalence as in the epidemiological impact, consistently confirmed by several independent methods used in the analysis.

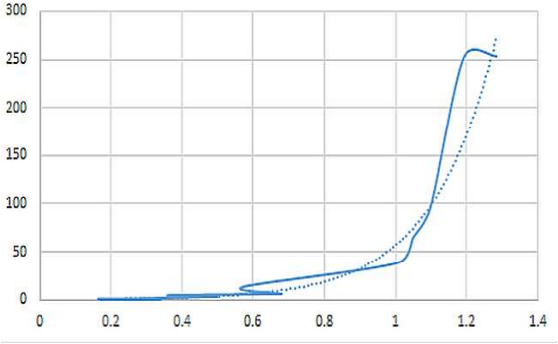

Figure 2 – Epidemiological outcome vs. combined dominant factors

Overall, the authors expect that the methods of factor influence analysis demonstrated in this work can be applied repeatedly over a time series of epidemiological data, allowing to reach confident conclusions by establishing and analyzing the trend over an extended period of time.

References:

1. V. N. Kuharev, V. N. Sally, A. M. Erpert, Economic-mathematical methods and models in the planning and management, Kiev, Vishcha School, 1991.

2. A. S. Kozadaev, A. A. Arzamasians, Prediction of time series with the apparatus of artificial neural networks. The short-term forecast of air temperature, Bulletin of the University of Tambov, Series: Natural and Technical Sciences, №3, is 11, 2006, pp. 299-304.

3. V. Ye. Snytyuk, Forecasting. Models. Methods. Algorithms: Tutorial, Kyiv, 2008.

4. O. Mulesa, F. Geche, V. Voloshchuk, V. Buchok and A. Batyuk, Information technology for time series forecasting with considering fuzzy expert evaluations, 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, 2017, pp. 105-108.

5. A. S. Mendel, Method counterparts in predicting short time series: expert-statistical approach, Machine Telemechanics, № 4, 2004, pp. 143-152.

[1] **Mykhailo Dvoretskyi**
A senior lecturer at the Department of Software Engineering
[2] **Svitlana Dvoretska**
A senior lecturer at the Department of Software Engineering
[3] **Hlib Horban**
PhD., an associate professor at the Department of Software Engineering
[4] **Yuriy Nezdoliy**
A senior lecturer at the Department of Software Engineering
[1,2,3,4] *Petro Mohyla Black Sea National University*

USING THE ANALYTIC HIERARCHY PROCESS FOR OPTIMIZATION THE DATABASE STRUCTURE OF A DISTRIBUTED CORPORATE INFORMATION SYSTEM NODE

Abstract. One of the research aims is to determine and build a mathematical model of the optimality criteria for the structure of a remote node of the distributed corporate information system database. The statistics of user SQL-queries activity is taken into account and presented in the form of a multidimensional database. Criteria of the model effectiveness are formulated and the problem of multicriteria optimization is solved. Choosing the best alternative makes possible to determine the optimal level of the data representation marker.

Keywords: corporate information system, distributed database, SQL-query, multicriteria problem, analytic hierarchy process.

In information systems development, there is a trend of transition from local to distributed databases (DDB). Within one company there is a need to automate different types of accounting. The attempt to automate all types of accounting leads to so-called "universal" corporate information systems. This approach has many disadvantages [1], which can be eliminated by using separate specialized solutions. But this path leads to use of several databases (and perhaps DBMS) that require their synchronization [2-3].

A key factor influencing the reliability and accessibility of such databases is the so-called localization of links. If the database is distributed so that the data hosted in a node is called exclusively by its user, it indicates a high level of link localization [4].

A combined data distribution strategy is the best in terms of combining the benefits of strategies with and without duplication. But when using it, in addition to the task of synchronizing duplicate information, the task of designing the structure of the database is actual, depending which node data belonging to [4-5]. The performance of the system will directly depend on the decision on the need for partial or complete duplication of data.

The purpose of the research is to create a mathematical optimization model and subsequent choosing the best alternative to the marker of data representation of the remote node of distributed CIS. The next obtained multicriteria problem need to be

solved to determine the optimal level of data representation marker. The solution of the problem is also complicated by the fact that the solution space is defined on a set of real numbers, and therefore the set of solutions contains a large number of alternatives. The research is related only to the relational databases.

To avoid the need for further replication some data that required on the DDB node can only be presented on the central node of the database and participate the query through the use of distributed queries. Due to the fact that to represent the data on the remote node it is necessary to use elements of both vertical and horizontal data fragmentation (both projection and selecting), the node relation is a subset or attributes and tuples of the base relation R.

The model of presenting user queries should support the possibility of their further classification according to belonging to a particular workplace, location, user role and other criteria that can be added to the model. That is, the user query is defined as

Q = <DateTime, WorkplaceType, Location, UserRole, Application, R, A, tup>

For the dimensions elements the term of data representation marker is proposed. It reflects the level of data representation necessity at the node of distributed CIS.

So, we have a model where each dimension attribute has a value, a marker and a weight A_{dim} ={Val, Mrk, vol}, where Mrk = {"obligatorily", "necessary", " neutral", "not required" , "forbidden"}, and vol – weight (ignored for the values of the marker "obligatorily" and "forbidden"). By converting a non-numeric linguistic variable of markers into a numeric value ("obligatorily" – "2", "necessary" – "1", "neutral" – "0", "not required" – "-1", "forbidden" – "-2"), the aggregation function was defined:

$$Aggregate_{i=1}^{n} Mrk_i = \begin{cases} 2, \ if \ \exists \ Mrk_i = 2 \\ -2, if \ \exists \ Mrk_i = -2 \ \wedge \nexists \ Mrk_i = 2 \\ \sum_{i=1}^{n} (Mrk_i * \frac{Vol_i}{\sum_{i=1}^{n} Vol_i}) \end{cases}$$

When deciding on the data representation on a remote node, we consolidate the rows of the fact table by the tuple <R, A, tup> and calculate the value of the marker for each of its elements by formula (5). And based on following the decision about data representation is made:

$$Repr\ (Node,R,A,tup) = (Aggregate(R, A, tup)_{i=1}^{n} Mrk_i > koef_{repr}^{node})$$

When deciding on the data representation on a remote node, we consolidate the rows of the fact table by the tuple <R, A, tup> and calculate the value of the marker for each of its elements by previous formula. And based on following the decision about data representation is made:

$$Repr\ (Node,R,A,tup) = (Aggregate(R, A, tup)_{i=1}^{n} Mrk_i > koef_{repr}^{node})$$

where $koef_{repr}^{node}$ – the threshold coefficient of data representation in a certain node Node, that is defined at the range of [-1, 1].

When planning the structure of the database of the remote node of distributed CIS, several factors will be involved - availability and speed of data obtaining,

independence from the central DB node, the DB size, the level of data reliability, the need for further synchronization [6-7]. Criterion of independence from the central database node, and, accordingly, the availability and access speed directly depend on the representation of user SQL-query data on the node of distributed CIS. The criterion of the local database size affects both the performance of queries to the local database and the power of computing resources required to perform database and CIS administration operations. The criteria of the need for data synchronization is the ratio of the number of modified data queries to the total number of queries. In the research the mathematical models of mentioned criteria were obtained.

A multicriteria problem, that was obtained, need be solved to determine the optimal level of data representation marker. The analytic hierarchy process (AHP), which is a general methodology for solving a wide class of decision-making problems, allows to combine a relatively simple mathematical apparatus with knowledge and experience of the decision maker.

When compiling the hierarchy, following relationship between the levels elements was used: goal - stakeholders - criteria - alternatives. The value of the data representation marker (alternative) is a real number in the interval [-1, 1]. It leads to potential large number of alternatives at the 4th level of the hierarchy and therefore the matrices of pairwise comparisons by criteria can become very big. This complicates estimation process for the decision makers. It is proposed to simplify the task by reducing the number of alternatives to 5: "low" (L) – "-1", "lower them medium" (LM) – "-0.5", "medium" (M) – "0", "higher then medium" (HM) – "0.5", and "high" (H) – "1". The level of "decision makers" is represented by the elements "Owner", "Database Administrator", "Database Developer" and "CIS Operator". The obtained hierarchical model is shown in Fig. 1.

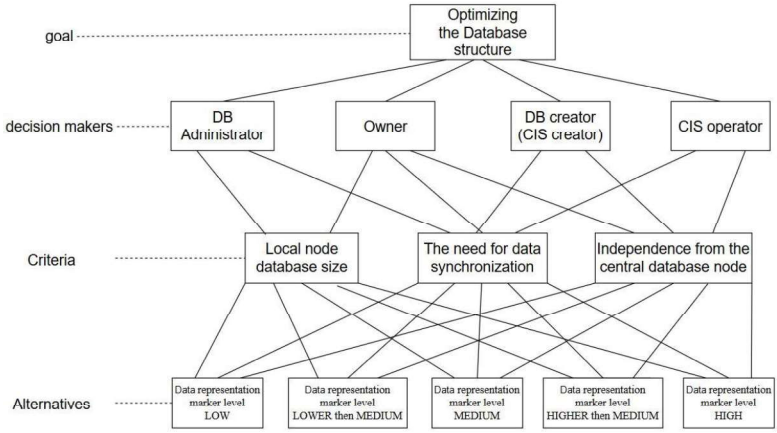

Figure 3: Hierarchical model of the distributed CIS node structure optimization problem

Solving a multi-criteria problem and finding the optimal level of data representation at a remote node increases the level of data availability and efficiency of distributed CIS. Efficiency is defined as the ratio of result and resources, so taking into account the vector of relative weight of the optimality criteria of the model, we calculate the efficiency as

$$Eff = \frac{F_{availab} \times W_1^{criteria}}{F_{size} \times W_0^{criteria} + F_{synchro} \times W_2^{criteria}}$$

Thus, the results of the research allow to increase the efficiency of using the distributed CIS node of the subject area by 25% compared to the presentation of only critical data, and by 11% compared to the presentation of all necessary data of the central database, respectively. The research can be followed by presenting the obtained vector of global priorities in the form of fuzzy sets of one variable. Dephasing the obtained results can make numerical value of the optimal level of data representation at the DCIS node more accurate.

References

1. M. Dvoretskyi, S. Borovlova, Web-application of warehouse accounting in non-automated points of sale, Science works "Petro Mohyla Black Sea National University", Rel. 308. T. 320, Series: Computer technologies, 2018, pp. 45–52 (in Ukrainian).

2. M. Tamer Özsu, Patrick Valduriez. Principles of Distributed Database Systems 3rd ed. Springler, 2011.

3. Automatic synchronization of distributed databases in split mode, [Online]. Availably: http://stimul.kiev.ua/materialy.htm?a=avtomaticheskaya_sinkhronizatsiya_raspredelennykh_baz_dannykh_v_razdelennom_rezh (in Russian)

4. H. Garcia-Molina, J. D. Ullman, and J. Widom, Database Systems: The Complete Book 2nd Edition, Pearson, 2008.

5. Dusan Petkovich. Microsoft SQL Server 2019: A Beginner's Guide, Seventh Edition 7th Edition, Kindle Edition, Mc-Graw-Hill Education, 2020.

6. M. Dvoretskyi, S. Dvoretska, Y. Nezdoliy, S. Borovlova, Data Utility Assessment while Optimizing the Structure and Minimizing the Volume of a Distributed Database Node, in: Proceedings of the 1st International Workshop on Information-Communication Technologies & Embedded Systems (ICTES 2019), Mykolaiv, 2019. pp. 128–137

7. M. Fisun, M. Dvoretskyi, A. Shved and Y. Davydenko, Query parsing in order to optimize distributed DB structure, in: Proceedings of the 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, 2017, pp. 172–178. doi: 10.1109/IDAACS.2017.8095071.

[1] **Ievgen Fedorchenko**
Senior Lecturer
[2] **Andrii Oliinyk**
PhD, Associate Professor
[3] **Alexander Stepanenko**
PhD, Associate Professor
[4] **Anastasiia Kharchenko**
Student
[5] **Marharyta Saman**
Student
[1-5]Department of Software Tools, National University "Zaporizhzhia Polytechnic"

RESEARCH AND DEVELOPMENT OF A GENETIC ALGORITHM FOR DIAGNOSING THE STRENGTH OF THE BLADE STRUCTURE IN GAS TURBINE ENGINES

Abstract. The problem of developing a method of technical diagnostics on the basis of data for the study of the structural strength of the blades is considered. It has been determined that evolutionary methods, including genetic algorithms, are effective methods of computational intelligence that can be used to build diagnostic models. The proposed methods and tools can be used to predict the state of critical load points in the diagnosis of gas turbine blades of aircraft engines during operation.

Keywords: evolutionary method, genetic algorithm, optimization, technical diagnostics, software, forecasting

Introduction

The relevance of the research topic lies in the presence of the need for accurate and fast methods of searching for critical points of functions of different complexity, especially multimodal and multifunctional. Even in such cases, the application of standard search methods complicates the search, makes it more costly and increases the time of information processing. Such a problem is especially acute in areas where diagnostic processes are critical to time and high accuracy, such as aircraft construction [1].

Analysis of literature data and problem statement

Despite the constant development of methods and algorithms for solving technical diagnostics, this task is relevant. This is due to the fact that it is impossible to create a universal method, because the task of technical diagnosis is divided into subtasks. Such tasks include preventive technical diagnostics, technical diagnostics of failures, diagnostics of maintenance of the current state of the object, etc. These subtasks can also be divided by the nature of the diagnosis - it can be a diagnosis by one or more parameters, and so on. The described reasons are the basis for the creation of new methods of technical diagnosis [1].

Development of an algorithm for technical diagnosis

Algorithm of the modernized GA with leader method based on GA with decreasing population size. It was decided to develop a modified method "GA with leader" based on GA with population reduction, which proposes to change the population size depending on the number of most adapted individuals in the population, thus reducing the amount of computation to obtain the optimal solution.

In the proposed method for generating a new set of solutions from the input data, the number of individuals corresponding to the size of the population, the most adapted to the analysis, is selected. Then, when selecting the parent pair from the current generation, a certain number of pairs is selected by the probable rank. In each iteration (in each generation) this number is different. Each pair is selected taking into account the values of the objective functions of the most adapted and least adapted individual in the population:

$$e^{-\frac{f_j - f^t_{worst}}{f^t_{best} - f^t_{worst}}} < rand(e^{-1};1), j = \overline{1;N} \quad (1)$$

where, N is the size of the current population; f_i - the value of the fitness function of the j-th individual; f^t_{best} - the best value of the fitness function of the current population t; f^t_{worst} - the worst value of the fitness function of the current population t; rand $(e^{-1};1)$ - random number from e^{-1} to 1 [2].

During crossing, a descendant G_i is created, which is located at some distance from the ancestor with the best values of the fitness function G_1 in the direction from the ancestor with the worst value of the fitness function G_2. Determining the value of the i-th gene of the gi chromosome-offspring is determined by the formula:

$$g_i = k\left(g_{1i} - g_{2i}\right) + g_{1i} \quad (2)$$

where k∈ [0; 1] - the actual coefficient specified by the user at the stage of initialization of genetic search.

At the stage of mutation, starting from the first chromosome, the whole population is reviewed, and for each chromosome H_j, drop numbers x_i from the interval [0; 1) are assigned. If this number is less than the probability of mutation, then the current chromosome H_i is mutated. In the selected chromosome there is a mutation of genes by some value:

$$g^*_{ij} = g_{ij} + \Delta g_{ij}, \quad (3)$$

where i is the gene number in the chromosome; j - chromosome number; g_{ij} - gene for mutation; g_{ij}* - gene after mutation [3].

The values of the i-th gene gij of the chromosome G_j after mutation can be calculated by formula:

$$g^*_{ij} = \begin{cases} g_{ij} + \Delta\left(p, \max_i - g_{ij}\right), 1 \le i \le w \\ g_{ij} + \Delta\left(p, g_{ij} - \min_i\right), w < i \le K \end{cases}, \quad (4)$$

where

$$\Delta(p,y) = y\left(1 - c^{(1-p/P)^v}\right), \quad (5)$$

where c is a randomly generated number in the interval [0; 1]; p is the number of the current iteration; P is the maximum number of iterations; v is a parameter that

determines the degree of homogeneity (uniformity); min_i and max_i - the minimum and maximum value of the i-th parameter in the solution with the help of the genetic method of the problem; w is a number equal to | K / 2 |; K is the number of genes in the chromosome [4].

The new generation is formed from the existing set of solutions obtained as a result of the application of crossing, mutation and inversion operators. The probability of an individual to be selected for a new generation is calculated by formula (6) [5]:

$$P\left(X^i\right)=\frac{-fitness\left(X^i\right)+D}{N\cdot D-\sum_{j=1}^{N} fitness\left(X^j\right)} \quad (6)$$

After that, the criteria for stopping the evolutionary search are checked (achievement of an acceptable value of the objective function, absence of significant improvements of the values of the objective function during a certain number of iterations, exceeding the maximum possible search time, etc.). In case of non-satisfaction of the stopping criteria, the stages of crossing, mutation and inversion are repeated.

The results of the algorithm

Table 2. Influence of population size on the accuracy and speed of algorithm operation during processing of unimodal functions (number of generations - 50, total discharge - total – 256, of them into a fractional part - 16)

Population size	5	10	20	40	80	160	320	640
Canonical								
Minimum	4	3,9867	4	4	3,7968	3	3	3
time, s	0,8548	0,8392	0,8299	0,8361	0,8392	0,8361	0,8486	0,8642
Genitor								
Minimum	4	4	4	4	3	3	3	3
time, s	1,0077	0,9360	1,1481	1,0826	1,3260	1,5412	2,1122	3,1512
CHC								
Minimum	4	4	4	4	3	3	3	3
time, s	0,9453	0,9703	1,0358	1,0670	1,2417	1,4976	2,0716	3,1512
Island model								
Minimum	-	1.631e +137	2.070e +120	1.198e +51	3	3	3,2000	3
time, c	-	1,7752	1,7971	1,8751	1,9812	2,1590	2,5833	3,4320
Bidirectional GA (DAGA2)								
Minimum	-	5,288e +122	6,549e +60	3,095e +53	1,475e +19	3	3	3
time, s	-	2,7268	2,6239	3,0544	3,7253	5,3071	8,6518	15,703
GA with a decrease in population size								
Minimum	-	3	3	3	3,2000	3	3	3,2000
time, s	-	1,2261	1,3010	1,4664	1,8782	2,7175	4,3118	7,7064

From the results obtained in Table 1, it was concluded that the accuracy of all methods manifests itself at a population size of 80 or more individuals, and becomes stable in the range from 160 The only exception is the method with a decrease in population size, which shows a stable result for all values of the population size.

Conclusions

It has been determined that technical diagnostics is a field of knowledge, that it consists of theory, methods and means for identifying the state of objects. It is noted that, as a rule, the state of an object is determined based on the available observations of it (measured values of input parameters) and a mathematical model that describes the relationship between the input and output parameters of the objects under study, and which is built on the basis of the training sample data. It has been determined that evolutionary methods, including genetic algorithms, are effective methods of computational intelligence that can be used to build diagnostic models.

As a result of the research, the genetic algorithm was refined and adjusted to reduce the proportion by adjusting its parameters to increase the speed of evolutionary optimization. Recommendations are given for tuning the initial parameters of the evolutionary search when using the proposed modification. The parameters of the method, in particular, the size of the population, the number of generations, the bit width and the type of crossover are selected in such a way as to minimize the operating time of the module being developed and to obtain an accuracy within acceptable limits.

References:

1. R. Thompson, Automotive Technology: A Systems Approach, 6th ed., Cengage Learning, Inc, Clifton Park, 2014.

2. M. Busch, Mike Bush on engines: what every aircraft owner need to know about design, operation, condition monitoring, maintenance and troubleshooting of piston aircraft engines, 1st ed. CreateSpace Independent Publishing Platform, 2018.

3. S. Russel, P. Norving, Artificial Intelligence: A Modern Approach, CreateSpace Independent Publishing Platform, 2016.

4. F. Shen, A fast multi-tasking solution: NMF-theoretic co-clustering for gear fault diagnosis under variable working conditions, Chinese journal of mechanical engineering 33, (2020), pp. 16. DOI: 10.1186/s10033-020-00437-3.

5. W. Zamboni, G. Petrone, G. Spaqnuolo, D. Beretta. An evolutionary computation approach for the online/on-board identification of PEM fuel cell impedance parameters with a diagnosis perspective, Energies 12, (2019), pp. 4374. DOI: 10.3390/en12224374.

[1] **Hlib Horban**
PhD., an associate professor at the Department of Software Engineering
[2] **Ihor Kandyba**
A lecturer at the Department of Software Engineering
[3] **Mykhailo Dvoretskyi**
A senior lecturer at the Department of Software Engineering
[4] **Anzhela Boiko**
Ph.D., an associate professor at the Department of Computer Engineering
[1-4] Petro Mohyla Black Sea National University

PRINCIPLES OF SEARCHING FOR A VARIETY OF TYPES OF ASSOCIATIVE RULES IN OLAP-CUBES

Abstract. The classification of association dependencies which can take place among multidimensional data is presented in the article. The representation of templates of inter-dimensional association rules is considered. Generation methods of inter-dimensional and intra-dimensional association rules are presented. Formulas for calculating objective and subjective characteristics of significance of these association rules types are presented.

Keywords: OLAP, Data Mining, multidimensional data, association rule, support, confidence, lift, leverage, template, dimension, measure, attribute, combination, set, difference.

Technologies of on-line Analytical process (OLAP) [1, 2] and data processing [3, 4] are typically employed in trendy data analysis systems and in decision support systems, that alter additional or less effective knowledge analysis. OLAP technology permits conducting user-defined operation like consolidation, detalization, data slice, cube rotation et al. At the identical time data processing investigates some cumulated hidden knowledge that was unknown before that and will be enough helpful within the data analytics process, upon that knowledge is taken from data sheets pre-spawned likewise by means that of database management systems (DBMS). One of the foremost common tasks of Data Mining is association, that represents detection of regularities between connected objects, an example of which can be the rule that event Y follows event X [5]. X is named a condition or an antecedent, and Y is named a consequent. Rules of that sort are known as association rules.

Data Mining strategies and algorithms [6], together with association rule mining likewise, are chiefly supported processing bestowed in tabular type, wherever sets of analyzed knowledge are settled either in one column or in one line, so that they add one dimension. But such knowledge regularities could happen even in three-dimensional data [7]. If to think about a three-dimensional cube rather than relational table data, then an item set for association rule mining may be bestowed as a collection of attribute values for every dimension, likewise as sets of values within the plurality of

dimensions.

The main elements of OLAP cubes are dimensions and measures. Dimension could be a values sequence some of the parameters to be analyzed. Samples of dimensions is time, geographic location, etc. Typically, dimensions contain extra data that permits users to investigate actual knowledge. Values that are obtained at the intersection of cube dimensions and represent quantifying facts are referred to as measures.

Mathematically the hypercube is suitable to represent by following sets:

D – a set of hypercube dimensions for a specific subject area:

$$D = \{D_1, D_2, ..., D_i, ..., D_n\},$$

where D_i – i^{th}-dimension, n – the quantity of dimensions;

A – a set of attributes (values of elements) of hypercube dimensions:

$$A = A_1 \cup A_2 \cup ... \cup A_i \cup ... \cup A_n,$$

where A_i – a set of attributes of dimension D_i, that successively are often diagrammatic as:

$$A_i = \{A_i^1, A_i^2, ..., A_i^k, ..., A_i^m\},$$

where – k-attribute of i^{th}-dimension, m – the quantity of attributes in i^{th}-dimension;

M – a set of values of hypercube measures:

$$M = \{M^1_{I_1,I_2,...I_i,...,I_n}, ..., M^l_{I_1,I_2,...I_i,...,I_n}, ..., M^z_{I_1,I_2,...I_i,...,I_n}\},$$

where I_i – attribute index of i^{th}-dimension, n – the quantity of dimensions, $M^l_{I_1,I_2,...I_i,...,I_n}$ – l-measure for the cube cell with $I_1, I_2, ..., I_i, ..., I_n$ index, z – the quantity of hypercube measures.

If to contemplate OLAP cube rather than relational data, then associate item set of association rules will represent a collection of values (attributes) of every dimension. Association rules that arise in multidimensional data will be classified by the subsequent sorts:

1. Inter-dimensional association rules – rules between attributes of different dimensions:

$$(A_I^x \in D_I) \wedge ... \wedge (A_J^y \in D_J) \rightarrow A_K^z \in D_K;$$

2. Intra-dimensional association rules:

$$(A_I^x \in D_I) \wedge ... \wedge (A_I^y \in D_I) \rightarrow (A_I^z \in D_I) \wedge ... \wedge (A_I^v \in D_I);$$

3. Hybrid association rules – dependencies between dimensions, but some operands can be attributes of the same dimension:

$$(A_I^x \in D_I) \wedge ... \wedge (A_J^y \in D_J) \rightarrow (A_J^v \in D_J) \wedge ... \wedge (A_K^z \in D_K).$$

Hybrid association rules may be known as repetition association rules in distinction to different rules thought of, that essentially represent association rules while not repetitions.

As modern databases can be very large (up to gigabytes and terabytes), you need efficient algorithms to find reflection rules that can be scaled up and that will allow you to find a solution within a reasonable time.

One such algorithm is Apriori, first proposed by Sricant and Agraval [8]. Originally it was developed for relational databases and allowed the generation of frequent data sets from transaction tables.

Frequent subject set in multidimensional data means a set of attribute values for the relevant measurements, the value for which is below the threshold for the minimum support value, which is set by the end user based on his own experience.

This results in frequent subject sets from data first with one dimension, then with two, etc. Finally, frequent subject sets can be found with n dimensions, where n is the total number of measurements in a cube.

In general, let the set of all frequent sets of topics in the OLAP cube be a set of S:

$$S = \{S_1, S_2, ..., S_i, ..., S_n\},$$

where i is the number of elements in a subject set, S_i is a lot of frequent subject sets with the number of elements and, n is the total number of elements in a cube.

In turn, sets of S_1, ..., S_n contain different subject sets for each of the measurements or sets of measurements if the number of elements in the set is greater than one.

In other words:

$$S_1 = \{s_1, s_2, ..., s_n\},$$

where s_1 is a set of frequent single element subject sets in the first dimension of the cube, s_2 in the second dimension and s_n in the n dimension.

In turn, many two-element subject sets can be presented as follows:

$$S_2 = \{s_{12}, s_{13}, ..., s_{mn}\},$$

where s_{12} is a set of frequent subject sets for the first and second dimensions, s_{13} for the first and third dimensions, $m \neq n$.

Let k be the number of elements in the subject set. So, in general:

$$S_k = \bigcup_{i=1}^{C_n^k} \{\underbrace{s_{i_1, i_2, ... i_k}}_{k}\}.$$

It is clear that when creating frequent OLAP cube subject sets, they will not include all the elements included in the corresponding cube measurements. To include an element or a collection of them in such sets, you must first calculate the support for that collection.

It is proposed to create a frequent subject set in the form of a list, where the first element is a sublist containing the sequence numbers of cube measurements according to which the set is generated [24]. in a single element set, such a list contains only one element. This sublist in the first element in the further generation of associative rules

based on subject sets is necessary to identify the measurements for which all sets have been created.

Among multidimensional data similar to tabular one, it is possible to find certain association dependencies represented in the form of rules that can be classified as inter-dimensional, within one dimension and hybrid. The approach to construction of templates of inter-dimensional association rules is proposed by generating all possible combinations of dimensions in OLAP-cube, which allows obtaining possible association rules, as well as the approach to construction of association rules within one dimension by generating all possible combinations of values of a certain dimension, among which search for dependencies is carried out. Appropriate methods have been developed for generating inter-dimensional association rules and association rules within one dimension. In the future, it is planned to study methods of hybrid association rule mining among multidimensional data.

References:

1. E. Thomsen, [OLAP Solutions: Building Multidimensional Information Systems], John Wiley & Sons, New York, 2002, pp. 50–688.

2. R. Wrembel, C. Koncilia, [Data Warehouses and OLAP: Concepts, Architectures and Solutions], Idea Group Inc., 2007, pp. 12–237.

3. D. Hand, H. Mannila, P. Smyth, [Principles of Data Mining], Massachussets Institute of Technology, Cambridge, 2001, pp. 2–546.

4. N. Ye. (Ed.), [The Handbook Of Data Mining], Lawrence Erlbaum Associates Publishers, 2003, pp. 254–690.

5. C. Zhang, and S. Zhang, [Association Rule Mining: Models and Algorithms], Springer-Verlag, Berlin, 2002, pp. 12–238.

6. Kovalenko, Y. Davydenko and A. Shved, Formation of Consistent Groups of Expert Evidences Based on Dissimilarity Measures in Evidence Theory, in: Proceedings of the 14th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2019, pp. 113–116. doi: 10.1109/STC-CSIT.2019.8929858

7. H. Zhu, [Online analytical mining of association rules. Master's thesis], Simon Faster University, Burnaby, 1998, pp. 5–51.

8. Symeonidis and P. Mitkas, Agent intelligence through Data Mining, Springer Science+Business Media, Heidelberg, 2005, pp. 3–200.

[1]**Andrii Khlevnyi**
PhD, Associate professor Department of technology management
[2]**Bohdan Koval**
Master's degree student Department of technology management
[3]**Svetlana Shabatskaya**
Associate professor Department of Medical and Biological Physics and Informatics
[1,2] *Taras Shevchenko National University of Kyiv*
[3]*National Medical University*

DEVELOPMENT OF A FRAUD DETECTION SYSTEM IN PAYMENT SERVICES USING CRISP-DM METHODOLOGY

Abstract. The paper presents the development of a system for detecting fraud in payment services using CRISP-DM methodology. It appoints that the methodology partially satisfies the requirements of financial institutions and proposes its adaptation.

Keywords: methodology, data analytics, payment services.

The demand for various payment services is constantly increasing nowadays. It is coupled with the development of technology and increasing need for such systems. High transaction speed, transparency, full control over payments, and ease of use are essential for clients. However, in addition to the benefits of modern payment services, there is a dynamic problem associated with maintaining the integrity of transactions. Therefore, one of the urgent issues for banks is to solve the problem of detecting and preventing illegal actions with their financial resources. The best way to deal with fraud is to prevent it. The warning is possible due to systematic data processing, with fairly well defined stages. To address this issue, it is appropriate to apply a methodological approach. Given the basics of data analysis [1,2] it has been determined, that the use of CRISP-DM is appropriate for the development of a system that will detect fraud at the transaction stage. Let's take a look at the basic principles and concepts underlying the methodology of developing a system for detecting fraud in payment services.

According to CRISP-DM [2] the lifecycle of a fraud detection system project consists of 6 stages. It has been confirmed that any methodology does not work without adjusting it to the relevant business needs [3]. Therefore, the sequence of steps is not considered from the point of their strict compliance and depends on the requirements of the applied field - financial system. According to the results of the study, all development stages of a system for detecting fraud in payment services have been taken. It has been discovered that at 3 - 5 stages data analysts play the key role, and at stages 1, 2, and 6 the key is the project team. Let's look through these stages and fill them with the main tasks from the point of developing a system for detecting fraud in payment services.

1. Business understanding. At this stage, we need to explore the business processes of the financial institution that owns the payment services. Identify the main types of fraud the institution deals with, and its competitors. We need to set a business goal - to save money and increase customer loyalty by minimizing fraudulent

transactions. Investigate competitors and identify the risks that occur in the development of such systems, prepare the project plan to develop a system for fraud detection in payment services and its expected results, establish responsible persons for each stage of the methodology.

2. Data understanding. To implement this stage, we need to determine data sources for each type of fraud, configure basic parameters of the data collection, set attributes, perform data cleaning and pre-processing, identify valuable subsets to form hypotheses of fraud in the system, so we can further analyze hidden patterns.

3. Data preparation. Data analyst is responsible to pre-process the data before modeling at this stage.

4. Modeling. Here we choose the methods using which the transactions' analysis and fraud detection will be resolved, and implementing the model. Also, we determine the methods of testing, training and evaluation of models.

5. Evaluation. The list of approved models and the stakeholders' feedback on the results is determined at this stage. Key persons prepare alternative solutions and strategies, that may be used to improve the models, the following steps are appointed.

6. Deployment. Develop a specific methodology for implementing a fraud detection system in payment services of financial institutions.

References:

1. Foster Provost and Tom Fawcett.Big Data.Mar 2015.51-59.http://doi.org/10.1089/big.2013.1508

2. Shafique U. Qaiser H. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 12 No. 1 Nov. 2014, pp. 217-222 © 2014 Innovative Space of Scientific Research Journals http://www.ijisr.issr-journals.org/

3. Khlevna J.L., Khlevyi A.O. Osnovy` formuvannya meta-metodologiyi upravlinnya proektamy`. XIV- Mezhdunarodnaya prakty`cheskaya konferency`ya "Upravleny`e proektamy`: sostoyany`e y` perspekty`vy", My`kolayiv, 11 - 15 September 2018p.

4. Fraud Detection Techniques: Data and Technique Oriented Perspective / S. Sorournejad, Z. Zojaji, R.E. Atani, Amir Hassan Monadjemi / Cornel University Library, 2016. Mode of access: https://arxiv.org/ftp/arxiv/papers/1611/1611.06439.pdf.

[1] **Nikolay Kiktev**
[2] **Taras Lendiel**
[3] **Volodymyr Osypenko**
[1,2] *National University of Life and Environmental Sciences of Ukraine*
[1] *Taras Shevchenko National University of Kyiv*
[3] *Kyiv National University of Technologies and Design*

APPLICATION OF THE INTERNET OF THINGS TECHNOLOGY IN THE AUTOMATION OF THE PRODUCTION OF COMPOUND FEED AND PREMIXES

1. Introduction

An increase the efficiency of using the technological opportunities of agricultural units is achieved through extensive automation of production processes and the development of a computer control system. In recent years, there is actively developing a new direction - cloud technologies and the Internet of Things (IoT), which can also be used in the automation of agricultural production, including in the production of poultry feed and premixes. The IoT is the most advanced tool in industrial automation.

2. Problem Statement and Literature Servey

Given the above, to increase the efficiency of feed production and product quality it is necessary: to improve the control system of technological processes in the production of poultry feed by developing a two-tier computer-integrated control system; to apply modeling of individual technological processes; at the second level of the integrated control system to apply software to calculate the best feed recipe for nutritional qualities; using of modern cloud technologies and the Internet of Things for remote monitoring of the process via the Internet. Purpose of the research. This work solves the problem of developing a hardware and software complex for managing the production of poultry feed using cloud technologies and the Internet of things. With the advent of production automation, this area also applied to poultry feed production. Thus, in [3] researchers investigate how to obtain the optimal composition of feed. It is usually performed using the simplex method [4]. This technology will be included in AquaSim, a set of custom IoT productivity tools for Skretting [2].

3. Formulation and solution of the problem of optimizing the feeding ration of animals. Control system operation algorithm

The objective function of the feeding optimization problem can be written as follows:

$$Z = \sum_{j=1}^{m} C_j X_j \to \min,$$

where C_j - cost or purchase price of the j-th type of feed; X_j - the required amount of the j-th type of feed in the daily diet, under restrictions (conditions) - nutrients in the diet contain at least the required amount:

$$\sum_{j=1}^{m} A_{ij} X_j \geq B_i,$$

where A_{ij} - the content of the i-th nutrient per unit of the j-th type of feed; B_i - daily requirement of the animal in the i-th nutrient.

The operation of the feed production control system is as follows. The grain comes from the loading hoppers through the pipeline to the ripper passing the flow sensor, which removes the flow rate. The signal from the flow sensor is fed to the flow meter which control signal through the actuator closes or opens the valves of the grain loading hoppers. After the baking powder, the raw material enters the hammer crusher. It is proposed to create a control system for the technological process of feed production using the technology of the Internet of Things, namely the principle of remote control and monitoring of technological processes. In project we use an Arduino MEGA2560 + WiFi R3 controller from RobotDyn [1]. In addition to Wi-Fi, the microcontroller is able to run programs from external flash memory with SPI interface. Schematically, the control system is shown in Fig. 1. The system works as follows: after power supply, the entire system is initialized; the current value of time t is compared with the critical value of time *tk* (time for which the whole technological process); measurement of technological parameters; check of wireless communication; under the condition of wireless connection, the measured value is displayed; waiting for the command; go to step 2; in the absence of wireless communication, the system goes into automatic mode; after the transfer of control action to the actuators, the data is sent to the personal computer of the general control system; go to step 2; when the critical time value is reached, the program is terminated. The control system is programmed in the "Arduino IDE" environment.

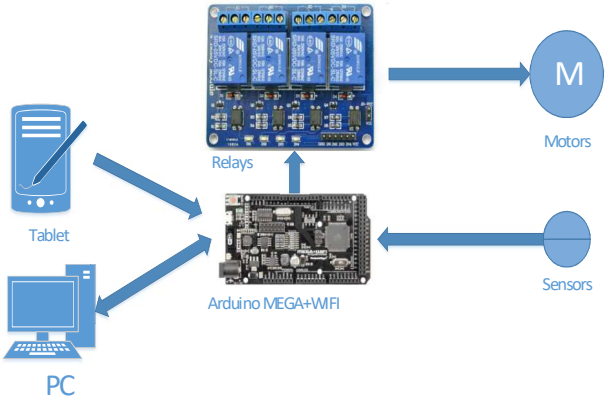

Figure 4: Symbolic circuits of the control system

4. Processing Experimental Results and discussion

Conducted experimental researches have shown that there is developed the control system of technological process of production of compound feed with use of technology of the Internet of things. There is implemented an algorithm of control of technological process with use of technology of the Internet of things and is considered at the same time the objective function of a problem of optimization of feeding. In

particular, there is created a prototype of the installation with the proposed module for monitoring and maintaining technological parameters. The layout included an Arduino MEGA2560 + WiFi R3 controller from RobotDyn, level sensors, actuators and a PC.

5. References:

1. V. Lysenko, T. Lendiel, and D. Komarchuk. "Phytomonitoring in a greenhouse based on arduino hardware." 2018 International Scientific-Practical Conference (PIC S&T) (2018): 365-368.

2. K.R. Raghunandan, L.J. Quadras, S. Gurunandan, S.S. Karthik. "Usage of Internet of Things in Agriculture Automation." International Journal of Computer Trends and Technology (IJCTT), Vol. 58, Issue 1, April (2018): 35-39.

3. A.O. Kashkariov. "Pro efektyvnist` skladannya receptiv kombikormu." [On the effectiveness of compound feed recipes]. Konferenciya molodyh vchenyh TDAU, (2010): 1-3. (In Ukranian).

4. N.A. Kiktev. "Infomacionnye tekhnologii v reshenii zadach upravleniya proizvodstvom kombikormov." [Information technologies in solving problems of compound feed production management]. Innovacii v sel'skom hozyajstve, № 3 (13), Moscou (2015): 53-57. (In Russian).

Natalia Kondruk
Candidate of Technical Sciences, Associate Professor of the Department of Cybernetics and Applied Mathematics
Uzhhorod National University

SEGMENTATION OF DATA SETS BY DIFFERENT TYPES OF CLUSTERS

Clustering is a powerful tool in the field of Data mining, when there is no a priori information about the relationships between data. Currently, many cluster analysis algorithms are successfully used in various application areas, where there is a need to divide similar in certain features objects into subsets [1-3]. A crisp split into clusters is possible only with very different features of clustering objects. Therefore, fuzzy methods are increasingly used to solve real problems, in which the division of objects is carried out to determine the degree of belonging of objects to clusters.

All existing methods can be classified according to the similarity measures they use [3, 4]. On the other hand, it determines the different geometric shape of the formed clusters and allows obtaining qualitatively different applied interpretations of the obtained homogeneous segments of data sets. Therefore, it is the specifics of applied problems that make it impossible to automatically transfer methods to another application area without the risk of deliberately obtaining a bad solution. Therefore, it is advisable to develop an information system that would have a fairly wide range of tools for grouping objects by different similarity measures. This makes it possible to effectively solve a lot of applied problems in different subject areas. The main works in which the technology is presented, which allows to solve this problem are presented in [4-7].

The focus of the system is a single-level clustering method based on fuzzy binary relations described in [5]. The flexibility of this algorithm allows you to form different geometric shapes of clusters of datasets by simply changing the appearance of the degree of similarity of objects. In this case, the similarity of the objects O_i and O_j by some criterion is characterized by a fuzzy binary relation R on the set of vector features with the membership function μ_R. The closer the value μ_R is to 1, the more similar the objects will be to this criterion. Thus, in [4-7], three types of similarity measures of objects are proposed: length-based, angular and distance.

To form elliptically similar clusters, it is expedient to use the "distance" similarity measure, which is described by a fuzzy binary relation R^V [4]. The fuzzy binary relation R^K [7] characterizes the angle of deviation between the feature vectors. Its use makes it possible to carry out clustering with conical clusters. The length-based similarity measure R^D allows splitting the feature vectors of objects into clusters by concentric spheres [5].

Conical clustering can be effectively used to solve multi-criteria linear programming problems with a large-scale criterion space [7], which arise, in particular, in mathematical modeling of balanced nutrition problems. One of the steps in solving such problems is to cluster their criteria space. In this case, the relationships between the criteria are determined by their angular similarity R^K. Clustering by elliptical

clusters is most common in many application problems, as the similarity of objects is based on a "distance" similarity measure. Also in [6] two synthetic sets of two-dimensional data of Gaussian type are generated and efficiency of application of a clustering method based on fuzzy binary relations at various indices of an estimation of quality of partition is investigated. Clustering by concentric clusters (clusters in the form of concentric spheres) [5] made it possible to group objects by length-based similarity of their feature vectors and to obtain a qualitatively new applied meaningful interpretation of the formed homogeneous groups in practice. In addition, this approach allows for both crisp and fuzzy data clustering.

In perspective researches the combined index of an estimation of clustering quality which is adapted to use of various similarities measures of a fuzzy binary relations method will be created; development of a software system that will ensure the segmentation of data sets into different geometric shapes clusters without prior determination of the clustering threshold.

References:

1. T. Sajana, C. S. Rani, K. V Narayana, A survey on clustering techniques for big data mining. Indian journal of Science and Technology, 9(3), 2016, pp. 1-12. doi: 10.17485 / ijst / 2016 / v9i3 / 75971

2. A. Amelio, A. Tagarelli, Data mining: clustering. Encyclopedia of Bioinformatics and Computational Biology, 2018, pp. 437-48. doi: 10.1016 / B978-0-12-809633-8.20489-5

3. K. Chitra, D. Maheswari, A comparative study of various clustering algorithms in data mining. International Journal of Computer Science and Mobile Computing, 6(8), 2017, pp. 109-115.

4. N. Kondruk, Clustering method based on fuzzy binary relation, Eastern-European Journal of Enterprise Technologies, 2017, pp. 10–16. doi:10.15587/1729–4061.2017.94961

5. N. Kondruk, Use of length-based similarity measure in clustering problems, Radio Electronics, Computer Science, Control, 2018, pp. 98–105. doi:10.15588/1607-3274-2018-3-11.

6. N. E. Kondruk, A comparative study of cluster validity indices, Radio Electronics. Computer Science. Control, 4, 2019, pp. 59 – 67. doi: 10.15588/1607-3274-2019-4-6.

7. N. E. Kondruk, M. M. Malyar, Structuring of the ctriterional space by an angle similarity measure, Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics, 2020, pp. 85 – 91. doi: 10.24144/2616-7700.2020.1(36).85-91

[1] **Bohdan Koval**
Master of the Department of Management Technology
[2] **Iulia Khlevna**
Doctor of Technical Sciences, Associate Professor of the Department of Technologies Management
[1,2] *Taras Shevchenko National University of Kyiv*

FRAUD DETECTION TECHNOLOGY IN PAYMENT SYSTEMS

Abstract. The paper outlines the relevance of fraud prediction from the standpoint of the integrity of the study. It proposes the solution - the development of technology to detect fraud in payment systems and gives the definition of such technology. It has been established that in terms of technology it is important to develop an effective and optimized model for the classification of fraud in payment systems from the standpoint of all stages of the study.

Keywords: data science, machine learning, deep learning, data visualization, binary classification.

The fast expansion of the practice of financial institutions of user autonomy is a requirement of today. Financial institutions and customers face new challenges related to fraudulent malicious activities. The consequences of which are violations of the integrity and truthfulness of transactions, financial losses, reduced customer loyalty and their loss. To prevent this, we need to transform approaches and means to monitor, detect and control illegal actions. Of course, the best way to combat fraud is to prevent it. Such a signal is possible through the development of a system based on the prediction of fraudulent actions at the level of suspicious transactions and forecasting the probability of its occurrence.

According to the analysis [1, 2] it has been pointed, that the main essence of the presented works is the models of classification of transactions for fraud by different methods. The use of combined methods is presented in the works [3, 4]. It has been revealed that the research of the model of user behavior with the subsequent assignment of the transaction to fraudulent or non-fraudulent is the most relevant. Various methods and algorithms can be used to solve this problem. However, there is no powerful algorithm in the literature on credit card fraud that would be the standard for all financial institutions. [5]. Therefore, the study of fraud detection models, changes in their parameters, combination of algorithms to maintain each other's advantages and cover their weaknesses in detecting fraud with financial payment systems from the standpoint of systematization, namely in the form of consistent actions, is the technology of scientific and practical interest.

Technology of fraud detection in payment systems (TFDPC) has been proposed. TFDPC is a set of systematized ways to provide forecasting, detection and control of fraudulent transactions in financial systems. At the basis of such methods are models, methods and algorithms of machine learning.

In terms of the construction of TFDPC, it has been proposed to develop an

effective and optimized model for the classification of fraud in payment systems from the standpoint of all stages of the study. To implement the full cycle of TFDPC we chose the Python3 programming language for its simple syntax, broad support of the programming community and a huge amount of available documentation.

It has been established that to implement the solution it is advisable to choose the following libraries: pandas, numpy, matplotlib, scikit-learn, K-nearest neighbors, random forests, xgboost and others.

A dataset has been formed, which is based on a set of transaction data of an anonymous payment system, namely data of banking operations performed by individuals alone and consists of 6,362,620 records. 10 transaction attributes are selected.

Research analysis and data preparation consisted of: determining which type of transactions is most often fraudulent (and accordingly adapting the data frame), substantiation of anomalies in the funds transfer, analysis of anomalous transactions by initiator, quantitative detection of certain anomalies in transactions.

We apply visualization of the difference between fraudulent transactions and regular transactions and determine the correlations of attributes in regular and fraudulent transactions (Fig. 1).

Figure 1. Heats maps of attributes correlation in regular and fraudulent transactions

The metrics used to evaluate the model are based on the area under the response accuracy curve (AUPRC), rather than the usual area under the recipient performance curve (AUROC).

Models were created and analyzed, focusing on the detection of anomalies and supervised training: logistic regression, K-nearest neighbors, support vectors machine (SVM), the Bayesian classifier. The best result is achieved by applying an algorithm based on ensembles of decision trees that works effectively on unbalanced data. Among

these algorithms (based on decision tree ensembles) there are 2 most effective - Random Forest and XGBoost, and the last, gradient boosting algorithm, still shows the best result. In addition, XGBoost allows you to weigh the positive class (fraud) more efficiently than the negative class (no fraud) - which allows you to more efficiently process unbalanced data.

The constructed algorithm has AURPC score of 0.9986, which indicates a very high efficiency of the classifier. The accuracy of the model implemented using the technology of extreme gradient boosting is 99.97%.

The obtained results are not just high, but the technology can be recommended for use in business and banking, because out of 554,082 test transactions, only 3 transactions that were classified as genuine (non-fraudulent) turned out to be fraudulent, 166 actually genuine transactions were identified as fraudulent. Accordingly, the construction of TFDPC embodies: research analysis, data visualization with subsequent adaptation of the data set, technology creation using existing classification algorithms, visualization of the obtained model and results.

References:

1. Fraud Detection Techniques: Data and Technique Oriented Perspective / S. Sorournejad, Z. Zojaji, R.E. Atani, Amir Hassan Monadjemi / Cornel University Library, 2016. Mode of access: https://arxiv.org/ftp/arxiv/papers/1611/1611.06439.pdf .

2. Lebichot, B., Le Borgne, Y.-A.: Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection. In: Oneto, L., Navarin, N., Sperduti, A., Anguita, D. (eds.) Recent Advances in Big Data and Deep Learning, pp. 78–88. Springer, New York (2019)

3. Kuznyecova N.V. Analiz ta prohnozuvannya ryzykiv shaxrajstva z kredytnymy kartkamy. Informatics and Mathematical Methods in Simulation Vol. 8 (2018), No. 1, pp. 16-25

4. Kuznietsova, N.V. Scoring Technology for Risk Assessment of Fraud in Banking / Selected Papers of the XVI International Scientific and Practical Conference "Information Technologies and Security" (ITS 2016). — 2016. — Pp. 54-61 .

5. MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, "Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria", International Journal of Computer Applications (0975 – 8887) Volume 52– No.3, 2012.

6. Bailey MJ, Muth RF, Nourse HO. 1963. A regression method for real estate price index construction. Journal of the American Statistical Association 58: 933–942.

[1] **Yaroslav Linder**
Ph.D. in Physics and Mathematics, associate professor
[2] **Maksym Veres**
Ph.D. in Physics and Mathematics, associate professor
[3] **Kateryna Kuzminova**
[1,2,3] *Taras Shevchenko National University of Kyiv*

MODELING AND PREDICTION OF COVID-19 USING HYBRID DYNAMIC MODEL BASED ON SEIRD WITH ARIMA CORRECTIONS

Abstract. The stages of the proposed method are building a SEIRD compartment model with vital dynamics, estimating its parameters, calculating and predicting the difference between the SEIRD model solution and the observed data using the ARIMA model, and adjusting model prediction using this newly obtained data on the residuals. The proposed method was tested on the data on the epidemic`s dynamic in Ukraine. The validation results indicate the method`s aptitude to real-world usage.

Keywords: COVID-19, SEIRD, ARIMA, Hybrid Dynamic Model.

Introduction

As the coronavirus pandemic continues to rattle the world, humanity craves for means to alleviate the situation if not overcome the crisis entirely. Quality estimations and predictions of future dynamics of the disease spread will ensure better prevention and thorough preparation for exacerbations of the problem (such as expected rises in infection cases after the holidays or lockdown lifts). Rational use of resources may help avoid future boiling points for the healthcare and other systems critical to the delivery of the COVID-19 response.

The susceptible exposed infectious recovered model (SEIR), which is based on differential equations, is one of the most widely adopted methods for modeling the epidemic of the COVID-19 outbreak [1]. The SEIR model replicates the “time-history” of any epidemic or pandemic outbreak, and it presents the model of dynamic interaction between people with four different health conditions or phases of the pandemic, namely the susceptible (S), exposed (E), infective (I), and recovered (R). SEIRD model, as a generalization of the SEIR model, has an additional variable – Deceased individuals. A “Formal Characterization and Model Comparison Validation” based on the SEIRD model, which uses the data from Korea and Spain, is proposed by Casas et al. [3]. The proposed model showed the predicted parameterization with empirical evidence and a decision support system (DSS) is implemented to study the nature of the pandemic in Catalonia [3]. A data-driven model to predict the spread of Covid-19 for an upcoming week using the SEIRD model is studied and tested for datasets obtained from Italy, India, and Russia [2].

The hybrid dynamic model framework

Upon investigation, we introduce a novice model based on an enhanced SEIRD model and ARIMA model. As shown in Figure 1, the stages of the proposed method are building a SEIRD compartment model with vital dynamics, estimating its parameters, calculating and predicting the difference between the SEIRD model solution and the observed data using the ARIMA model, and finally adjusting model prediction using this newly obtained data on the residuals.

The model consists of such stages:

1. At the first one, we estimate SEIRD model parameters using historical data, trying to lessen the difference between the model`s output and observed data. This model is responsible for long-term prediction (i.e., 60 days or 100 days).
2. Calculate residuals between observed infected, recovered, and deceased percentage of the population and corresponding solutions of the SEIRD model.
3. Build three ARIMA models on the time-series of each of these residuals. Prediction of these ARIMA models will compensate residuals between the SEIRD model and historical data in order to make predictions mode accurate.
4. Validate the prediction of the obtained model using the data on the number of infected, recovered, and deceased individuals as of the most recent days, data on which was not included while working with the model on previous stages.

The compartments of the model are as follows:

- $S(t)$: Susceptible individuals - stock of healthy people who may be infected; population inflow due to births is taken into account.
- $E(t)$: Exposed individuals - virus carriers in the latent stage, during which they
- are not virus spreaders. Usually corresponds to an asymptomatic phase of the disease.
- $I(t)$: Infectious individuals - virus carriers able to spread the disease to individuals in contact with them.
- $R(t)$: Recovered individuals - stock of healthy people who are immune to COVID-19.
- $D(t)$: Deceased individuals - population loss due to the disease, natural deaths included.

The model itself is comprised of a system of differential equations:

$$\begin{cases} \frac{dS}{dt} = \Lambda N - \mu S - \frac{\beta SI}{N} \\ \frac{dE}{dt} = \frac{\beta SI}{N} - (\mu + \sigma)E \\ \frac{dI}{dt} = \sigma E - (g + \mu)I \\ \frac{dR}{dt} = g\left(1 - \mu_{COVID}(t)\right)I - \mu I \\ \frac{dD}{dt} = g\,\mu_{COVID}(t)I \end{cases} \quad (1)$$

with constraints at time t=0 S=S_0, E= E_0,I= I_0,R = R_0, D=D_0 and parameters Λ – population`s birth rate; $\mu$ – population`s mortality rate; β – rate of virus transmission,

which is the probability of transmitting disease between a susceptible and an infectious individual; σ – rate of latent individuals becoming infectious (average duration of incubation is $1/\sigma$); g – recovery rate, which can be initially estimated as $g = 1/D$, where D is the average duration of infection; $\mu_{COVID}(t)$ – death rate due to COVID-19, which is estimated by an inverse exponential formula $\mu_{COVID}(t) = \alpha e^{-\xi t}$. The population size $N(t) = S(t) + E(t) + I(t) + R(t)$ is not fixed due to its global birth and mortality rates taken into account at any given time t.

Results

In this section, we will provide results of hybrid model approbation on data from the Ukrainian finance analytics website [4].

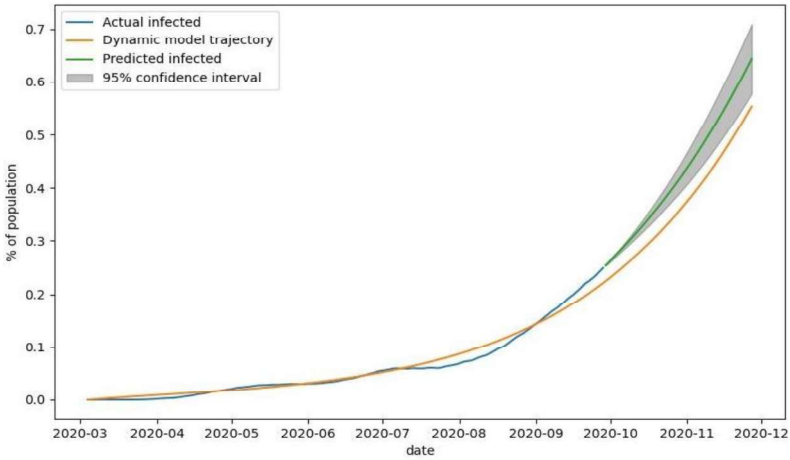

Figure 1 – The observed number of infected individuals (blue), number of infected individuals modeled with SEIRD model (yellow), and predicted number of infected individuals (green) by SEIRD model and corrected by ARIMA residual prediction with 95% confidence interval (grey)

The analysis of modeling and prediction of the number of infected individuals (Figure 5) shows that the number of observed cases of the disease grew steadily during the first half of the outbreak (mid-July) and is very accurately modeled with our method. The deviation of the predicted number of infected individuals from the observed data in the second half of July and August is most likely caused by the insufficient number of tests for COVID-19 performed during this period. The inconsistency in testing and changing levels of quarantine severity explain further deviations of observed data from the output of the SEIRD model. The prediction, corrected by ARIMA residual estimation, steadily increases, with optimistic and pessimistic scenarios (lower and upper bounds of the grey area, respectively) deviating by less than 0.1%.

Table 1

Quality measures of the fitted model for validations set

	MAE	MSE	MSLE	Normalized MAE	Normalized MSE	Max. deviation
Infected	$1.13 \cdot 10^{-4}$	$2.51 \cdot 10^{-8}$	$2.50 \cdot 10^{-8}$	$3.59 \cdot 10^{-2}$	$2.62 \cdot 10^{-3}$	8.6%
Recovered	$2.76 \cdot 10^{-4}$	$9.25 \cdot 10^{-8}$	$9.21 \cdot 10^{-8}$	$1.1 \cdot 10^{-1}$	$1.66 \cdot 10^{-2}$	15.4%
Deceased	$9.28 \cdot 10^{-6}$	$1.24 \cdot 10^{-10}$	$1.24 \cdot 10^{-10}$	$8.41 \cdot 10^{-2}$	$1.11 \cdot 10^{-2}$	15.5%

As shown in Table 1, all measures of the prediction quality for the infected, recovered, and deceased fractions of the population are very low.

Discussion

The proposed hybrid model consists of a dynamic SEIRD model with vital dynamics and decaying COVID mortality rate and three ARIMA models that cancel out dynamic model residuals and enhance prediction quality. The model was tested on Ukrainian COVID statistic data. Obtained validation results allow us to draw conclusions that the proposed hybrid model has good prediction ability and decent performance. Obtained long-term predictions reflect the general dynamic of the outbreak and are especially useful for the healthcare system workers and government officials. Obtained short-term predictions allow us not only to forecast the future number of infected, recovered, and deceased patients but only estimate forecast error under adverse or optimistic circumstances. The proposed method can be used as an effective tool for prediction and analysis of the dynamics of the COVID-19 pandemic.

References:

1. Hethcote, H.W., 1989. Three basic epidemiological models. In Applied mathematical ecology (pp. 119-144). Springer, Berlin, Heidelberg.

2. Rapolu, T., Nutakki, B., Rani, T.S., and Bhavani, S.D., 2020. A Time-Dependent SEIRD Model for Forecasting the COVID-19 Transmission Dynamics. medRxiv.

3. Fonseca i Casas, P., García Carrasco, V. and Subirana, J., 2020. SEIRD COVID-19 Formal Characterization and Model Comparison Validation. Applied Sciences, 10(15), p.5162.

4. Ukrainian finance analytics website, 2013. URL: https://index.minfin.com.ua/ua/reference/coronavirus/ukraine/.

[1] **Volodymyr Mikhieiev**
Student of the Department of Management Technology
[2] **Olga Mezentseva**
Candidate of Economic Sciences, Associated prof. of the Department of Technology Management
[1,2] *Taras Shevchenko National University of Kyiv*

ANALYSIS AND FORECASTING OF ENVIRONMENTAL POLLUTION BY CARBON DIOXIDE

Abstract. The main idea of this work is to analyze carbon dioxide pollution and apply several methods of mathematical modeling for this. In this case - the Gaussian model of the distribution of pollutants in the atmosphere and the matrix method of the effect of carbon dioxide on the state of the atmosphere.

Keywords: Gaussian model, forecasting, air pollution

The solution to the problem of managing the impact of industrial facilities on atmospheric pollution by emissions in the theoretical and methodological aspect requires the development of technologies for assessing the state of the environment and the dynamics of changes in the ecological and economic situation, as well as the formulation of problems related to environmental and mathematical modeling and the use of forecasting models. Modeling and forecasting, as the main tools of the system for managing the impact of industrial facilities on atmospheric pollution, allow not only to develop scenarios and options for ecological and economic development, but also to determine the maximum level of air pollution through the optimum of the objective function.

In the last decade, the expansion of the framework of the environment protection management system due to the inclusion of an environmental safety block in it required the solution of many methodological, regulatory, legal, and information problems. One of which consists in the methodological support of diagnostics of environmental safety and assessment of its state in order to rank disadvantaged areas and determine the priorities for their development[1].

One of the most effective methods is based on the Gaussian scattering model. When pollutants escape into the atmosphere, a cloud is formed, which is carried away along with the surrounding atmospheric air in the direction of the wind. In the process of movement, turbulent mixing with the surrounding atmospheric air occurs, which leads to the expansion of the cloud in space and a change in the concentration of pollutants in it. As a result, a spatial distribution of the concentration of pollutants is formed, which in the most general case is described by the normal (Gaussian) law[2].

In the case of using the matrix method, the values of the air pollution indicator, obtained in the form of relative values, characterize the growth rates of emissions (pollution) for individual industries and the average growth rate for industry. Forecasting is performed based on the dependence $Y = f(X, t)$ and a specific analytical expression defined for each industry. In this case, the calculation of the forecast values

of the growth rates of the mass of emissions by each industry is carried out by a simple substitution of time periods for the subsequent forecast interval[3].

Conclusions. Especially important and urgent are the tasks of monitoring the concentration of pollutants and forecasting the state of the air basin in industrial regions. Solving these problems will allow us to rationally approach the issue of locating new enterprises in the region, make administrative decisions in the field of environmental safety, and develop effective measures to reduce the level of air pollution. Predicting the level of air pollution will warn the population about possible dangers and strengthen environmental control by society. Using monitoring systems, organizations will be able to adjust the work schedule of employees in the open air, and government agencies will warn about the dangers of various events on days when the norms are exceeded, and industrial enterprises can regulate the level of emissions of harmful substances into the atmosphere.

References:

1. Vyvarets A.D., Belik I.S., Stepanova N.V., Leontyev Y.V., Nikulina N.L., Atmospheric pollution assessment industrial emissions, 2017. URL: https://elar.urfu.ru/bitstream/10995/40861/1/ozapv_2006.pdf.

2. Antonova M.A., Vorobev A.V., Vorobev A.V., Dutova E.M., Pokrovskiy V.D., Modeling the distribution of pollutants in the athmosphere, 2019. URL: http://earchive.tpu.ru/bitstream/11683/55288/1/bulletin_tpu-2019-v330-i6-17.pdf.

3. Matveev Y.N., Maslennikov B.I., Karelskaya K.A., Stukalova N.A., Mathematical modeling of processes the spread of a pollutant in soils and the atmosphere, 2016. URL: https://naukovedenie.ru/PDF/65TVN516.pdf.

Julia Minaeva
Ph.D., Associate Professor, Associate Professor at the Department of Intellectual Technologies.
Taras Shevchenko National University of Kyiv

PROCESSING OF MULTIDIMENSIONAL AND MULTI-ASPECT DATA IN CONDITIONS OF UNCERTAINTY

Currently, the fuzzy set theory is a powerful mathematical apparatus that has a wide range of applications. However, some difficulties and inaccuracies in the fuzzy set theory (FST) application were found, in particular, in conditions of insufficient information, when the membership function (MF) assignment is impossible or associated with difficulties [1].

Fuzzy variables (FV), fuzzy numbers (FN) are the main objects of FST, their use reflects the uncertainty [2]. The limits of fuzziness are determined by the expert, using the available a priori knowledge of the system. Fuzzy numbers are in many aspects similar to probability distributions, and in [3] is brought an example of constructing a fuzzy number— an analog of a normal probability distribution — by using the convolution theorem.

Fuzzy set theory declared the universality of its models and the generality of their application, although several processes are not amenable to meaningful (formal) representation in the form of fuzzy sets (FS) due to their complexity lack of study. For example, the process of using FST in the study of biomedical processes is associated with some difficulties, the main of which is the omission or distortion of data, which does not allow the expert to assign a membership function. A new class of tasks has been identified that have a high semantic readiness for the use of the FST apparatus, but the incompleteness of the data prevents this [4].

FST was created as a means of solving problems (primarily - management) under conditions of uncertainty in the 2D data space. Further expansion of FST to 3D space is caused by new types of tasks that are difficult to transform into 2D space without losing the task representation adequacy. This emphasizes the relevance of finding new methods, models, and tools of solving problems under conditions of uncertainty in a multidimensional space, taking into account the multifaceted nature of the data.

Fuzzy sets (FS), and not only FN, are a product of the human mental activity, the universal set must be calculated based on the initial data set, in particular, the definition of the universal set (US) interval [min/max] requires data, but the vast majority of works almost "leave behind" this fact. This is especially noticeable in the case of the influence of BIG DATA where, on the one hand, the use of FST can give a certain effect, on the other hand, the fundamental difficulties in determining the US limit the possibilities of using an effective mathematical apparatus.

Under uncertainty, a multidimensional (multi-aspect) object can be represented by a subset of ordered sequences (multi FS analog) – a subset of OS or a subset of ordered pairs – SOP, (analog of FS) [1], which are equivalent in terms of proximity of F-norms. In our case, the application of a subset of ordered pairs and subset of ordered

sequences for the analysis and modeling of uncertainty is based on the condition of invariance of norms).

A multi-aspect object is considered as:

$$\left\{ \begin{array}{c} x_1^{(1)} \quad x_1^{(2)} \cdots x_1^{(n)} \\ \vdots \\ \underbrace{x_m^{(1)} \quad x_m^{(2)} \cdots x_m^{(n)}}_{Aspects} \end{array} \right\} \quad (1)$$

Fuzzification (MF calculation) is offered in 2 cases:

- the expert can assign heuristic FNs for each aspect;
- the choice or assigning of MF is limited.

A 3D initial data set, as a rule, is presented in the form of slices or fibers, which significantly complicates the process of forming FS, requires, in turn, new assumptions, in particular, the decision that can be obtained in one direction not always can be taken as a general, also, it is difficult to take into account distorted or omitted data, pre-processing is required, especially in terms of restoring some subset of the original data. In 3D space, there is, on the one hand, the need to take into account the uniqueness of the problem, on the other hand - the ability to bring the initial problem to a level where it is possible to apply standard methods for one- or two-dimensional spaces. Let's note, that 2 type-FS are objects of 3D space, fuzzy sets of such type are extensions of 1 type - FS, which belong to objects from 2D space

In conditions of uncertainty, the object under study has a number of hidden properties that can be detected by structuring the object (data set) in the form of 2D or 3D tensors, performing the next step of tensor decomposition, and obtaining a fuzzy-like subset of ordered pairs similar to FS. In addition, a powerful blurring tool that allows obtaining SOP is using of special matrices. This gives the ability for blurring not only a subset of values, such as a universal set but also a more complex object, such as a block matrix.

References:

1. Minaeva, Julia & Filimonova, Oksana. Alternative subsets of ordered pairs and their application in decision-making problems under conditions of uncertainty. Management of Development of Complex Systems, 41, 68 – 82 (2020). dx.doi.org\10.32347/2412-9933.2020.41.68-82.

2. Hanss. Applied Fuzzy Arithmetic. An Introduction with Engineering Applications. Springer-Verlag Berlin Heidelberg 2005- 260 p.

3. Kofman A. Introduction to the theory of fuzzy sets: Per. with French. M.: Radio and communication. 1982.

4. E. Acara, D. M. Dunlavy, T. G. Kolda, M. Morup. Scalable Tensor Factorizations for Incomplete Data. arXiv:1005.2197v1 [math.NA] 12 May 2010.- 34p

[1] **Anastasiia Mudra**
Student of the Department of Management Technology
[2] **Olga Mezentseva**
Candidate of Economic Sciences, Associated prof. of the Department of Technology Management
[1,2] *Taras Shevchenko National University of Kyiv*

EXAMINATION OF THE DEPENDENCE BETWEEN CRIMINAL'S APPEARANCE AND HIS OFFENSE USING MACHINE LEARNING

Abstract. The main idea of this paper is to present the relevance of researching of dependency between tendency to some actions, that in our case is type of offense, and appearance, in particular, face. Also describes the solution method – using convolutional neural network.

Keywords: convolutional neural network, Criminal's facial features, Offense

Not so far after the discovering of photography, some scientists began to notice similarities in the photographs of criminals taken after their arrest. If you believe their words, criminals are united by common facial features.

Modern scientists have already tried to prove this theory using the capabilities of artificial intelligence. Such experiments were held in America and China, but the results were not enough. This situation only proves the relevance of the topic and the fact that there is still a lot of work to be done in this direction.

A new view on the problem, that is describing in this article is trying to connect criminal's appearance with a type of his crime. Why exactly appearance? A growing number of studies have linked facial images to personality. It has been established that humans are able to perceive certain personality traits from each other's faces with some degree of accuracy. Studies focusing on the objective characteristics of human faces have found some associations between facial morphology and personality features. For instance, facial symmetry predicts extraversion. And, actually, we try to establish the connection between personality and the offence's type, because there is a wide range of different types of crime, that requires different traits from assailant and have different impacts on victims [1].

The databases of criminals are in every country, but they are confident. That is why the first information for model are put from Interpol from "Red Notices" category. Red Notices are issued for fugitives wanted either for prosecution or to serve a sentence. A Red Notice is a request to law enforcement worldwide to locate and provisionally arrest a person pending extradition, surrender, or similar legal action. Naturally, the data set needs more additions [2]. It is necessary to repeat the experiment with more people of different ages, gender, ethnic groups and with more information about them, like motives.

The most logical solution for solving this problem is to use the capabilities of a neural network. There is a need to process a lot of information, that, first of all, contains photos of criminals. It is the hardest part for neural network. When processing images,

there is a need to scan photos from different angles. That is why we decided to use convolutional neural network, which successful work with images is confirmed. Its architecture includes 2 main paradigms: local perception and shared weights.

Further, it technology can be used in systems of face recognition. Coupled with an automated biometric software application, this system is capable of identifying or verifying a person by comparing and analyzing patterns, shapes and proportions of their facial features and contours. Unlike a person [3], the computer vision algorithm has no subjective "baggage", emotions, prejudices regarding experience, race, religion, political beliefs, experience. It doesn't get tired, it doesn't need sleep or food. Thus, sometimes, it can help different specialists to improve their work.

Conclusions. Identifying patterns between a person and a crime can provide an opportunity to make a set of rules by which people and their tendencies can be classified and described. The results from the study will not try to call someone a criminal, rather it is an attempt to find out if there is an objective dependence of appearance and inclination to certain behavior. If the correlation will be confirmed, then further these data can be used in order to prevent the development of undesirable behavior. For example, it can be a classification of young people and further conversations with a psychotherapist, who, as a more educated specialist, can help solve some internal conflicts that a person does not admit, ignores or when a person does not have someone, who can help in overcoming this issues.

References:

1. Xiaolin Wu, Xi Zhang, Automated Inference on Criminality using Face Images, 2017. URL: https://arxiv.org/pdf/1611.04135.pdf.

2. Osin E., Novokshonov A., Shutilov K., Davydov D., Kachur A., Assessing the Big Five personality traits using real-life static facial images, 2020. URL: https://www.researchgate.net/publication/341568689_Assessing_the_Big_Five_personality_traits_using_real-life_static_facial_images.

3. Nurul Azma Abdullah, Md. Jamri Saidi, Nurul Hidayah Ab Rahman, Chuah Chai Wen, Isredza Rahmi A. Hamid, Face recognition for criminal identification: An implementation of principal component analysis for face recognition, 2017. URL: https://aip.scitation.org/doi/pdf/10.1063/1.5005335#:~:text=Face%20Recognition%20for%20Criminal%20Identification%20is%20a%20face%20recognition%20system,be%20removed%20from%20the%20image.

[1] **Dmytro Orlovskyi**
PhD, Associate Professor
[2] **Andrii Kopp**
Senior Lecturer
[1,2] *National Technical University "Kharkiv Polytechnic Institute"*

A BUSINESS INTELLIGENCE DASHBOARD DESIGN APPROACH TO IMPROVE DATA ANALYTICS AND DECISION MAKING

Abstract. This paper considers a problem of dashboard design, since usage of inappropriate visuals may mislead users and shift their focus to wrong things. Bar charts, line charts, and pie charts are considered as the most common visualization graphs. Proposed approach includes dataset preparation and analysis phases. While dataset preparation phase is mostly focused on star schema transformation into flat structures, dataset analysis phase proposes recommendations on which visualizations may be placed on a designed dashboard.

Keywords: Data Analytics, Business Intelligence, Dashboard, Star Schema.

In today's competitive market situation, it is extremely important for small business and large corporations to have permanent access to analytical reports regarding their business activities. Such access may be granted by modern data analytics and data visualization techniques covered in this section. However, it is a challenging problem to design information technologies in this field.

The dashboard design problem includes selection of visualizations, such as graphs and charts, which should be placed in a limited space. If choose inappropriate visualization charts that do not fit nature of data presented in datasets prepared for visualization, developed Business Intelligence (BI) dashboard applications may mislead business users and shift their focus and attention to unimportant or wrong things. Thus, dashboard design problem is extremely relevant nowadays when big data volumes processing and analysis is vital for making business decisions. Since the problem is not trivial and complex enough, it is required to propose a BI dashboard design approach.

Proposed approach is based on the relational algebra methods used to process Data Mart (DM) and Data Warehouse (DW) data structures in order to prepare datasets for analysis and suggest appropriate visualizations that should be placed on dashboards. The dashboard design process, which is underlying for a proposed approach, includes steps related to dataset preparation and dataset analysis that lead to recommendations on dashboard visualizations.

Proposed dashboard design process generalizes all the tasks required to prepare DM or DW measures and dimensions for visualization. At first, it is necessary to transform a star schema structure into a flat dataset. Since the star schema data warehouse model is one of the simplest but, on the other hand, is the most widely used data structure in BI domain, in this paper will be considered exactly this kind of storages [1]. This may be done using the SQL language join operators [2]. Then, it is required to prepare subsets of the general flat dataset. Prepared subsets will be used as

data sources for future visualizations (graphs and charts) that should be then placed on a dashboard. As well as the generic dataset, these subsets may be also prepared using the SQL language and its powerful selection and projection capabilities, and analytical functions [2]. After that, thresholds for pie charts and bar charts should be selected. When all the previous steps are completed, recommendations regarding the visualization charts and graphs, which should be used to display on a dashboard prepared data subsets, may be obtained.

With respect to the generated recommendations, the content of a designed dashboard may be created by data analysts or other stakeholders. The dashboard design process diagram is shown in Figure 1.

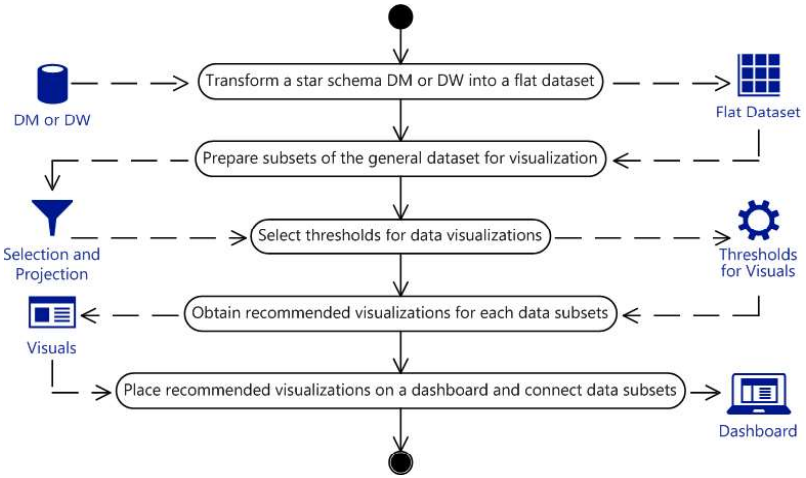

Figure 1 – Dashboard design process diagram

Described process allows users to choose appropriate visualization charts and graphs. Outlined approach is based on data mart or data warehouse transformation from the star schema, extremely popular and quite simple data structure, into the flat dataset that should be used to produce data subsets, which then may be used as data sources for visualizations, which are usually bar charts for comparisons, line charts for trends over time, pie charts for parts of a whole, and cards for scalar values [3]. Future research includes software implementation of the proposed approach, as well as the research of placement of visualizations on a dashboard's space.

References:

1. P. Bhatia, Data Mining and Data Warehousing: Principles and Practical Techniques, Cambridge University Press, 2019.
2. R. Ghlala, Analytic SQL in SQL Server 2014/2016, John Wiley & Sons, 2019.
3. R. Telg, T. A. Irani, Agricultural Communications in Action: A Hands-On Approach, Cengage Learning, 2011.

[1] **Vladyslava Rudenko**
Student of the Department of Management Technology
[2] **Olga Mezentseva**
Candidate of Economic Sciences, Associated prof. of the Department of Technology Management
[1,2] *Taras Shevchenko National University of Kyiv*

INFLUENCE ANALYSIS OF DIFFERENT MANAGEMENT METHODOLOGIES ON THE RESULT OF BIG DATA PROJECTS

Abstract. The paper considers various methodologies of project management and their interaction on BIG DATA projects. In essence, big data project management relies on general project management methodologies, but not all of them can be successfully applied to the full. Because this is a relatively new industry, big data may require something new or at least a combination of standard approaches.

Keywords: analysis of project management methodologies, BIG DATA, Data Science

Since Data Science, and Big Data in particular, is still evolving and lacks direct form, there can be no single answer to the question of which methodology works best in such projects.

In general, the following methodologies can be distinguished:

• CRISP-DM as a traditional approach to project management in the field of Data Science.

• Waterfall as a traditional approach.

• Scrum as an Agile approach.

• Kanban as an Agile approach.

This is not a complete list of methodologies used in big data projects. However, tools based on the above approaches will help identify key points where each approach can be useful to develop a model that will work for a specific Big Data project.

Cross Industry Standard Process for Data Mining (CRISP-DM) is a standard that describes general processes and approaches to data analytics used in industrial data-mining projects, regardless of the specific task and industry [1]. The CRISP-DM standard includes six iterative phases in data processing project management: understanding a business problem, understanding and retrieving data from different sources, preparing data, modeling data when building and evaluating a model, and actually evaluating that includes visualization and communication, and deployment and maintenance with final reports and project overview [2]. This methodology has a flexible cyclical nature and a focused approach, but it does not work for teams and does not cover communication issues at all. CRISP-DM can be called a sequence of works required to perform in Data Science projects [3]. This list of jobs can be kept in mind when managing Big Data projects, but it is too general, so other tools should be considered.

An approach such as Waterfall gives a clear and consistent picture of all the tasks that have been identified since the beginning of the project. The project or its individual phases are divided into smaller parts that depend on each other. In working with data, it is advisable to try the project in action as early as possible in order to quickly check and test hypotheses. If we stick to the waterfall style, we first need to make a complete and polished model, and then apply it in a work. Changes in this approach are not expected, however, in Big Data projects they may occur. This methodology does not work for the data processing. However, it may be useful to plan the use of some methodology tools, such as a Gantt chart. Thus, the main disadvantage of Waterfall is its lack of coverage of change management, which is not compatible with big data processing projects.

The main problem of Data Science projects in general is the misunderstanding between them and business goals [4]. Big Data projects add specific issues such as huge amounts of data and continuous changes. Agile methodologies can handle these problems. Scrum is one of the world's most common Agile approaches, in which large projects are divided into smaller phases, called sprints, and last from 1-2 weeks to 1-3 months. Each sprint has a fixed time frame and should achieve the results that have been set at the meetings. Scrum is largely focused on customer feedback. It is adaptive and flexible, with a high degree of autonomy, which in terms of Data Science allows you to optimize predictability. However, unlike Scrum in Software Engineering, where there is a constant increment to demonstrate results, Big Data has no feedback material. Sprint requirements are changing. In working with data, there is an iterative between the phases of data preparation and modeling, where everything can change. For example, a new hypothesis appeared, data were prepared, the hypothesis was tested, and everything planned for the end of the sprint is no longer of value. Everything needs to be reworked because the concept has changed. If you take into account this feature and slightly change the classic approach of Scrum, it can work well on Big Data projects. The continuous flow of a huge amount of information from the Internet, corporate systems or devices falls under the definition of Big Data at high speeds of download or accumulation [5]. Response speed and high frequency of result presentation is one of the 12 basic principles of Agile Manifesto.

Consider another Agile approach - Kanban. This methodology uses the board as a project and the cards as a task. The traditional kanban board includes three columns – To Do, In progress and Done [6]. For data processing projects, you can add additional columns: data preparation, development, coding, testing, and so on. It is necessary to improve the standard approach until it is suitable for working directly with the Big Data project. To use kanban in data analysis, some of the separate phases can also be combined - for example, modeling and data preparation can become one phase, because working with data and experiments occur at the same time. The approach places more emphasis on work in progress without reference to dates and roles, the work in it is highly visualized [7]. Disadvantages of the methodology - the lack of emphasis on dates and deadlines - can, on the contrary, be called advantages for big data processing projects.

It is difficult to find one methodology that would work well during the project life cycle in the Big Data field. Big data processing projects can combine several tools

of different methodologies in their management, and this is completely natural. It is worth exploring which tools influence the success of Big Data projects and synthesize them into a single approach that can have a positive impact on project results.

References:

1. Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

2. V. Gogunskii, O. Kolesnikov, G. Oborska, , ... S. Harelik, , D. Lukianov. Representation of project systems using the Markov chain, in: Eastern-European Journal of Enterprise Technologies, volume: 2(3-86), 2017, pp. 60-65. doi:10.15587/1729-4061.2017.97883

3. Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009); A Data Mining & Knowledge Discovery Process Model. In Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438-453, February 2009, I-Tech, Vienna, Austria.

4. V. Morozov, O. Mezentseva, M. Proskurin. Application of game theory for decisions making on the development of it products, in: Advances in Intelligent Systems and Computing, volume: 1, 2021. doi: 10.1007/978-3-030-54215-3_24

5. K. Senthuran, R. Sithamparanathan, S. Evans. Markov Decision Process-Based Opportunistic Spectral Access, in: IEEE Wireless Communications Letters, volume: 5 , issue: 5, 2019, pp. 544-547. doi: 10.1109/LWC.2016.2600576

6. Georg N. Krieg. Kanban-Controlled Manufacturing Systems. — Germany: Springer-Verlag Berlin Heidelberg, 2005. — 236 с. — ISBN 3-540-22999-X.

7. D. Anderson: Agile Management for Software Engineering - Applying the Theory of Constraints for Business Results. Prentice Hall, 2004, ISBN 0-13-142460-2

[1] **Tetiana Shelest**
Assistant Professor, Technology Management Department, Faculty of information Technology
[2] **Veronika Yeremieieva**
Master of degree, specialty Project Management, Faculty of information Technology
[1,2] *Taras Shevchenko National University of Kyiv*

ANALYSIS OF THE POSSIBILITY OF USING VR TECHNOLOGIES IN ENVIRONMENTAL AWARENESS PROJECTS

Reorganizing the projects connected with eco awareness is important in Ukraine because the waste management system is imperfect and has potential for development.

The main goal of this thesis is to collect data that will confirm the reasons that make the system imperfect, to analyze the ways of solving this problem and practices that will allow Ukrainians preserve the environment.

According to USA Today research, Ukraine ranks 9th out of 10 countries in the world that produce the largest amount of waste in relation to the population. According to the statistics of 2019, Ukraine produces 10.6 metric tons of waste per capita per year, and the total amount is approximately 474 106 065 metric tons. From this waste, only 3.2% are recycling [1].

Despite the fact that Ukraine produces a lot of waste and do recycle only small part of it, we do also import sorted waste from Europe. In 2018 we have imported around 100 thousand tons of plastic for 40 million dollars, glass waste for 11 million dollars and paper for 80 millions [2].

As a conclusion, Ukraine imports sorted waste from Europe because can not provide good quality of sorting in our country. The key problem in this case is that people do not know how to do it in a right way.

Analyzing the data of 2018 year from questionary “Environmentally conscious citizen - the key to successful implementation of the Association Agreement" was informative for our study. Let’s pay attention on two indicators: ways to solve environmental problems and practices that preserve the environment [3].

Table 1 – Ways to solve eco problems [3]

Ways to solve problems	Positive answers, %
Increase information on environmental issues	32.4%
Ensuring better application of current environmental legislation	27.3%
High panalties for violations of environmental legislation	44.7%
Implementation or dissemination of training programs	26.8%

Basing on this statistics (tab. 1), we can ensure that Ukrainians do need more information about ways of sorting. Moreover, introducing trainings, workshops and lectures to increase eco awareness can build new eco habits.

Here (tab. 2) we can see that more that 60% think that sorting waste is the key to build ecofriendly environment and around 40% are willing not to buy plastic for one use.

Table 2 – Practices that preserve the environment [3]

Practices that preserve the environment	**Positive answers, %**
Choosing more eco friendly transport	34.7%
Avoid buying disposable plastic products	39.3%
Sort the most part of waste	60.2%
Reduction of energy consumption	31.3%

This analysis showed that people are ready to change their environment, learn the sorting rules and rebuild their habits. But in Ukraine we do not have any learning content about eco awareness that would be interesting for kids as well as for adults.

Educational systems are now being radically restructured and raising the awareness of citizens can be regulated through multi-touch interactive VR technologies.

It is important that we can improve the experience of participants through manipulating and controlling VR environment. Through thematic educational modules it is possible to improve the experience of learners and increase their sensitivity to current environmental challenges [4].

We can install stationary interactive racks near supermarkets and sorting tanks. Each rack will have several VR modules, between which the user can switch and find out the most interesting information. Modules can include: an overview of things that can be made from recycled materials, a module with information about waste cycle in Ukraine, sorting rules and tips on how to minimize the waste amount.

With the help of interactive communication, the information is perceived much better. Therefore, the use of VR environment in various projects to increase environmental awareness, will introduce a completely new and experimental learning. Taken together, this will give the people an impetus to increase their eco-consciousness and learn the rules of waste sorting.

References:

1. H. Byrnes, T. Frohich, USA Today: Canada produces the most waste in the world. The US rank. URL: https://www.usatoday.com/story/money/2019/07/12/canada-united-states-worlds-biggest-producers-of-waste/39534923/

2. Will the planet turn into Plastic Porridge? URL: https://re-solutions.com.ua/ru/rus-plastyk-nastupaet-zachem-ukrayna-pokupaet-chuzhye-othody/

3. Environmental protection and citizens of Ukraine. Research of practices , values and judgements. May 2018, URL: https://www.rac.org.ua/uploads/content/481/files/envportraitpollreport2018.pdf

4. T. Mikropoulos, Virtual realities in environmental education., URL: https://www.researchgate.net/publication/226299007_Virtual_realities_in_environmental_education_The_project_LAKE

[1] **Serhiy Shtovba**
Doctor of science, professor
[2] **Mykola Petrychko**
Master, PhD student
[1] *Vasyl' Stus Donetsk National University*
[2] *Vinnytsia National Technical University*

AN INFORMETRIC ASSESSMENT OF VARIOUS RESEARCH FIELDS INTERACTIONS ON BASE OF CATEGORIZED PAPERS IN DIMENSIONS

Abstract. We calculated the level of interactions between all the pairs of the research groups and between all the pairs of the research divisions for 4 five-year periods. Paired interaction for research divisions shows that every consecutive five-year period has decreased irregularity of its distribution of interaction. Cu-mulative interaction shows that all research divisions tend to have their stickiness decreased over time. Paired interaction for research groups showed the same re-sults but difference between two consecutive periods is greater by a larger factor.

Keywords: informetrics, research group, research division, interactions, interdisciplinary, Dimensions, ANZSRC, Jaccard index, stickiness index, distribution, categorization.

Interdisciplinarity is a fashion direction in modern education and science. Interdisciplinarity is impossible without interactions between research from various fields. For quantitative assessment of level of interdisciplinarity, of level different research fields interactions many approaches are proposed in recent years. Currently the most widely used approaches to quantitative measuring of interdisciplinarity use bibliometric. They take into account co-authorships, collaborations, references, citations and co-citations. There are a few researches related to measuring interdisciplinarity using bibliometric [1-5]. In this research we used the same approach to assessing interdisciplinarity as in [5] - Jaccard index. We measured interdisciplinarity for Australian and New Zealand Standard Research Classification (ANZSRC) system. Our assessing is based on Dimensions' categorized papers where classes are research divisions and groups from ANZSRC. There are 22 research divisions with 157 groups. Now, Dimensions indexes over 110M research papers. Each paper is assigned to one or several research divisions and groups. Such categorization is carried out by Dimension itself by special software based on machine learning guided by topic experts. The title and abstract of the paper are source data for this categorization.

We analyzed the interaction of research divisions and research groups for 4 five-year periods. We analyzed paired and cumulative interaction for both research divisions and research groups. Only research groups of different research divisions were considered. Paired interaction allowed finding pairs that interact the most and analyze the behavior of different interactions over time.

Using paired interaction, we analyzed leaders and the most changeable research

divisions and groups in terms of interaction. Paired interaction for research divisions shows that every consecutive five-year period has decreased irregularity of its distribution of interaction. This also confirms the Gini index and the number of pairs with non-zero interaction. Cumulative interaction shows that all research divisions tend to have their stickiness decreased over time. The same effect shows Gini index for the stickiness index that decreases from 0.197 in the first period to 0.188 in the last period.

Paired interaction for research groups showed the same results but difference between two consecutive periods is greater by a larger factor. Cumulative interaction for research groups showed that with time research groups tend to have a more equally distributed stickiness index. It is also observed from the number of research groups that increased their stickiness index from zero to greater than zero and Gini index.

On base of interacting assessments, we identified five research groups that have been assigned suspiciously to divisions in ANZSRC. They are as follows: 0105 Mathematical Physics, 0602 Ecology, 0909 Geomatic Engineering, 1502 Banking, Finance and Investment and 2001 Communication and Media Studies. All the mentioned research groups have more similar research groups in concurrent divisions with statistically significant level. Those five research groups also have strong semantic ties with concurrent divisions, this can be an argument for adapting of ANZSRC.

References:

1. Wagner, S., Roessner, D., Kamau, B., Klein, J., Boyack, K., Keyton, J., Rafols, I., Borner, K.: Approaches to Understanding and Measuring Interdisciplinary Scientific Research (IDR): A Review of the Literature. Journal of Informetrics. 5, 14-26 (2011). doi: 10.1016/j.joi.2010.06.004.

2. Porter, A.L., Cohen, A.S., Roessner, J.D., Perreault M.: Measuring researcher interdisciplinarity. Scientometrics. 72, 117–147 (2007). doi: 10.1007/s11192-007-1700-5.

3. Noorden, V.R.: Interdisciplinary research by the numbers. Nature. 525, 306-307 (2015). doi: 10.1038/525306a.

4. Karlovcec, M., Mladenic, D.: Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. Scientometrics. 102, 433-454 (2015). doi: 10.1007/s11192-014-1355-y.

5. Shtovba S., Petrychko M. Jaccard Index-Based Assessing the Similarity of Research Fields in Dimensions // CEUR Workshop Proceedings, Vol. 2533 “Proc. of the First International Workshop on Digital Content & Smart Multimedia”. – 2019. – P. 117-128.

[1] Taborovskyi Andrii
Master's degree student Department of technology management
[2] Kolesnikova Kateryna
Doctor of technical sciences, Professor of the Department of Technologies Management
[3] Khlevnyi Andrii
PhD in Engineering Science, Assistant Professor of Department of Technologies Management.
[1-3] Taras Shevchenko National University of Kyiv

THE IMPACT OF AUTOMATED AND ROBOTIC WAREHOUSES ON THE SCOPE OF SUPPLY CHAIN PROCESS

Abstract. The essence and purpose of systems of automation and robotics of warehouses are considered, their advantages are defined, the offered software-analytical decision is described in the form of expert system.

Keywords: automation, robotics, supply chain, warehousing, warehouse capacity.

Warehousing is an important and integral part of every business. Its task is to save stocks of raw materials and finished products. It plays an important role in the movement of tangible assets, raw materials, materials, fuel, tools, equipment, spare parts, clothing, and other products.

Warehouse - a room or a set of rooms intended for storage of materials. Supply Chain Management combines the management of supply and demand within and between companies. A recent network of companies that cooperate with the aim of offering products and services is called an expanded enterprise. [2]

Automation and robotization of warehousing allows to introduce the technical, logistical and analytical measures that increase the level of productivity of facilities in the warehouse and reduce the impact of the human factor on the activities of the warehouse. The number of resources spent on the technical implementation of automated and robotic components of the warehouse is significantly less than the number of resources required for the maintaining of non-automated and non-robotic warehouse.

An urgent task is the choice of bots-carriers for their use in the department of planning the placement of boxes with goods for the automated warehouse. A solution in the form of an expert system is proposed. Its task is to recommend to the user on decision-making when choosing a manufacturer of carrier bots for automated warehouse.

During the development of the expert system, knowledge was extracted by analyzing the characteristics of carrier bots by the method of main components, cluster analysis and building a decision tree.

According to the results of the principal components method, three main components Comp1, Comp2 and Comp3 have been identified.

The first main component is determined by more than 66.3% of the following indicators: power, charge duration, charging speed.

The second main component is determined by more than 78.6% of the following indicators: the number of charging cycles, charging duration, price. The third main component is determined by more than 79% of the following indicators: power, charge duration, price.

The results of component and cluster analyzes were analyzed when constructing decision trees using the See5 / C 5.0 system. The result of the See5 system was expressed in the form of decision trees and 42 if-then rules that provide answers to the manufacturer's belonging to the class according to its characteristics.

The results obtained during the construction of decision trees completely coincided with the results of component and cluster analyzes.

The created expert system can become a prototype for a real decision-making system for an automated warehouse.

References:

1. Belinskyi P. I. Management of manufacturing and operations : P.I. Belinskyi, Y. Fedcovich national universist of Chernivtsi. Kyiv, 2005 – 380 p.

2. Donald Bowersox Integrated Supply Chain : Bowersox D, Olymp-business, 2017 – 500 p.

3. Dibska V.V. Warehouse management in Supply Chain: V.V. Dibska, Alfa-press, 2014 – 466 p.

[1]**Oksana Tereshchenkova**
Ph.D., associate professor
[2]**Kostyantyn Kondrashov**
graduate student
[1,2]*Kherson State Maritime Academy*

INFORMATIONAL EXPERT SYSTEM FOR MINIMIZING THE TIME FOR SEARCHING OF FAILURES OF SHIP ELECTRICAL EQUIPMENT

Abstract. This article is devoted to obtaining and comprehensively studying methods and models for fault trees and decision trees construction for identifying a defect in specific diagnostic objects and their structural units, as well as methods used to defects finding. The basic subjective and objective conditions that affect the time spent by maintenance personnel on the readapting to work of the failed ship system are systematized. The article substantiates the need of the transition from existing paper documentation to electronic maintenance documentation using the expert system.

Keywords: object of diagnostics, structural units, complex technical system, decision maker, expert system, alarm monitoring system, decision support system.

Each ship system can be represented as a complex of structural units interconnected and interacting with each other. In turn, each structural unit can be decomposed into many simple elements interconnected and interacting with each other.

Accordingly, the more structural units create a system, the more complex the system becomes and the more difficult it is to identify a malfunction in this system.

Thus, conventionally, all ship systems can be divided into 5 levels, according to the criterion of complexity when troubleshooting in this system.

Analysis of the failure diagnostic tools used by the operator in real navigation conditions to find and eliminate the causes of malfunction of shipboard automated systems and mechanisms is an actual problem. Quick search and elimination of a defect affect the level of ship's safety.

The troubleshooting process is the most difficult at electrical equipment repairing, as modern automated systems are a complex interconnected network of electrical and electronic circuits. The task of faulty element findingis finding of the sequence of checks when a minimum of time is spent on defect searching [1-3].

To be able to show the possible amount of time spent searching and repairing, an experiment was conducted on one of the container ships of the shipping company Mediterranean Shipping Company (MSC) m/v MSC "Brunella" [4, 5].

We used the archive logbook of the Kongsberg K-Chief 600 alarm monitoring system, the container ship MSC "Brunella", it was built in 2015. The total number of parameters controlled by the AMS is 3410 units [6].

Data of ship failures recorded in the ship's logbook for six months were conditionally divided by the level of complexity of the systems in which they occurred and are summarized in the table shown in Tab.1.

The purpose of the experiment was to calculate the average number of possible

causes of these failures, as well as the number of possible ways to eliminate them and the time taken to eliminate them.

Based on the obtained data, we construct a variational series of observations for the number of malfunctions, occurred within six months on the ship, and the number of their possible causes.

Let's find the relative frequency of events for 6 months W_i, for each level of complexity of the systems.

Table 3

The number of malfunctions recorded by the AMS system for six months				
Simple elements	Simple systems	Medium difficulty systems	Complex systems	Very complex systems
Total – 2	**Total – 126**	**Total – 198**	**Total – 168**	**Total – 12**

$$W_i = \frac{N_i}{n} \quad (1)$$

where N_i is the number of failures at a given interval; n is the total number of malfunctions within six months.

Let's substitute the numerical data, we obtain: $W_1 = 0.046$; $W_2 = 0.241$; $W_3 = 0.379$; $W_4 = 0.311$; $W_5 = 0.023$.

Find the numerical parameters: average value and variance.

Sample average:

$$\overline{X_B} = \frac{1}{n}\sum_{i=1}^{m} * N_i * X_i \approx 15 \quad (2)$$

Dispersion of discrete random variance:

$$D_B = \frac{1}{n}\sum_{i=1}^{n} * (X_i - \overline{X_B})^2 = 28{,}88 \quad (3)$$

Then, the standard deviation (standard error):

$$\sigma_в = \sqrt{28{,}88} \approx 5 \quad (4)$$

Thus, the average number of possible causes of an accidental failure recorded by the AMS system $\overline{X_B} = 15$ with the standard deviation of $\sigma_B = 5$.

One-sigma interval (confidence probability is 67%) for the given random variable is from 10 to 20 possible reasons.

This means that very often even experienced electricians will spend quite a lot of time guessing about the causes of the breakdowns and how to fix it.

The increasing of SAS effectiveness can be achieved in two methods.

The first method is the highly qualified personnel training. In order to quickly search and eliminate the OOD defect, the decision-maker must have the extensive

knowledge, experience and a wide range of personal qualities. In addition, he should be able to adapt to objective reasons that make troubleshooting difficult.

The problem is that availability of these qualities in one decision-maker (it is extremely unlikely), the process of defect searching can be taken place rather long. It is due to the information content received by the operator, in each case, is often excessive. The same OOD is represented by different models, and the information content about its elements and connections, as well as various features significantly exceeds the level necessary for defect searching.

So, it is impossible to draw up quickly a clear pattern of action at defect searching. The decision maker is always forced to keep in mind all the methods and algorithms of checks, to understand when to replace one method by another. In the process of searching of the same defect, he should constantly think about what to use at a given time. In this case, the factor of the human psyche works as limitedness to process a large amount of information (from 5 to 9) per unit of time.

As a result, even a competent decision maker falls into the mandatory time frame; it increases the troubleshooting process.

The second method is increasing the reliability of OOD by strengthening of the control over the operability of the main OOD nodes and the connections between them.

The problem here is that the structural, circuit and technological capabilities for improving the reliability of ship systems are limited, and, in practice, exhausted. Moreover, increasing of the OOD reliability due to the structural complication of diagnostic systems involves growing the number of measurements with dimension enhancement of the diagnosed circuit. It requires an increase of the control points in OOD; it inevitably raises a new problem related to the diagnostic systems reliability. In addition, their false positives can trigger a chain of incorrect operator actions leading to an accident or disaster.

As a result, even complex diagnostic systems help to reduce the number of failures of electrical equipment by timely informing the operator about violations in the operation of a particular mechanism, but, unfortunately, it doesn't contribute to the quick searching and elimination of a defect, in the case of ship system failure. And it requires the high qualification of the service personnel and a longer duration of the checks. In the conditions of autonomous navigation and with low qualification of the staff it can lead to undesirable consequences.

All described above clearly demonstrates the urgent need for the implementation of special information expert systems, which allow, even with low qualifications of the service personnel and low efficiency of control of OOD, to quickly search for defects in a failed ship system.

The proposed system will be built on the basis of knowledge, which includes the experience of experts in repair and troubleshooting. The knowledge base is formed on the basis of expert evaluation (experts are electricians with experience of at least 5 years, as well as superintendents of crewing firms with the same experience).

The system uses the approach that implements the task of separating of information stored in a common database and directly in the knowledge base (a set of decision tables).

To implement this approach, linking variables (link tables) are used.

By the use of these communication tables, a variable from the knowledge base is connected with the data stored in a common database of equipment and ready-made troubleshooting algorithms.

The knowledge base includes the knowledge and assessments of experts in failures, as well as databases with structural diagrams, principled schemes of elements and components, as well as troubleshooting algorithms.

Filling occurs from ship's logbooks. The number of malfunctions detected by the AMS system for vessels of the type container ship is recorded. For entry into the database, faults are ranked by their level of complexity. All entries are transmitted to the crewing company by the superintendent. The database is filled on the basis of the data of the logs collected from all ships of the crewing during the entire period of ship running.

The final product is software that provides the operator with complete, but not redundant information on the necessary malfunction, as well as a clear sequence of actions for its quick elimination.

The decision-making operation in the ES of a ship electrical engineer is: the registered error of the AMS is entered into the system window. The user receives all the necessary documentation of the unit that gave the error signal, as well as a set of strategies for troubleshooting. The variability of possible problems increases with the complexity of the mechanism. It becomes necessary to choose the most effective strategy to reduce the time of elimination. There is a table of opinions of experts who had the similar problems.

References:

1. B. Palyukh, T. Kakatunova, O. Baguzova, Intelligent decision support system for managing complex objects using dynamic fuzzy cognitive maps, Software Products and Systems, 2013.

2. C. Moreno, E. Espejo, A performance evaluation of three inference engines as expert systems for failure mode identification in shafts, Engineering Failure Analysis, 2015.

3. E. Liberado, Novel expert system for defining power quality compensators, Expert Systems with Applications, 2015.

4. KONSBERG. Standard K-Chief 600 Alarm and Monitoring System / 354760 / Rev.D March 2013 © Kongsberg Maritime AS

5. KONSBERG. Kongsberg K-Chief 500/600 Marine Automation System Installation Manual / 311956 / F March 2013 © Kongsberg Maritime AS.

6. Y. Krainyk, Y. Davydenko and V. Tomas, Configurable Control Node for Wireless Sensor Network, in: Proceedings of the 3rd International Conference on Advanced Information and Communications Technologies (AICT), Lviv, Ukraine, 2019, pp. 258–262. doi: 10.1109/AIACT.2019.8847732

Anastasiia Vavilenkova
Doctor of technical science, professor, docent
National Aviation University 1

RAGULARITY OF CONTEXT UNITS IDENTIFICATION IN ELECTRONIC TEXT DOCUMENTS

In this article had been analyzed actual software services, that can build relation's tree and make syntactical analysis. Each of them transforms primary text into the data structure with special features. The author proposed to use components of logic and linguistic models for automatic generation of grammar colocations. Also author suggested the rules for context units identification for complex sentences of natural language.

All early efforts to extract knowledge from the textual information by difference scientists leads in grammars by Homskiy and transformation grammars [1–2]. Grammars of regularity do not come up with collocations like analysis ones. However, almost all linguistic theories describe linear sequence of the sentence units by mean of hierarchic structure of gramma components [3]. Traditional ways of analysis by key parameters and standard answers are not possible to analyze natural language text at all and are not helpful for context analysis [4–5].

Today the main aim of natural language researches is automatic creation of context data structures for formalization of logical links by mean of particular algebraic construction. From the other hand, these researches give practical value for automatic analysis and synthesis of natural language texts by computer technologies.

Despite almost a century of research in artificial intelligence, context units identification still can not be realized in correct form for complex sentences.

One of the most essential thing for all systems that woks with natural languages must be using the process of grammatical analysis [6].

There is a full correspondence between grammar structure and logic form of natural language sentence of natural language sentence [7]. Considering grammar organization of the sentences, we have such graduation of sentences' members [8]:

- subject of the sentence – subject x;
- predicate of the sentence – relation p;
- object of the sentence – object y or subject-matter of relation z;
- definition – characteristic of subject g, characteristic of object q or characteristic of subject-matter of relation r;
- circumstance – characteristic of relation h.

A set of words connected between each other by logic links, will be lettered $sp_j, j = \overline{1,m}$, where m – amount of the collocations in the sentence.

According to the Ukrainian and English language rules, collocations can be formed between those members of sentence [9–10]:

- "definition – subject" – $sp_j = g \cup x$;

– "predicate – object" – $sp_j = p \cup y$;
– "definition – object" – $sp_j = q \cup y$;
– "object – object" – $sp_j = y \cup z$;
– "object – object" – $sp_j = r \cup z$;
– "circumstance – predicate" – $sp_j = h \cup p$.

The author formulated special rules for identification context units according to the rules for creating different collocations in flexional natural languages, examples of what were depicted above. It was developed 32 rules with additions for punctuation symbols in complex sentences and for considering homogeneous parts of the sentence. Some of these rules are represented below.

1. If the first world is adjective, numeral, pronoun or participle and the part of speech for second word is noun, their characteristics of case, number and genus are similar, the words are made collocation. For example, "*mathematical modelling*", "*computer modelling*", "*three pets*", "*her name*", "*designed room*".

$$if\left((cm(S_i)=2)\,and\,(cm(S_{i+1})=1)\right)and\left(g(S_i)=g(S_{i+1})\right)$$
$$and\left(n(S_i)=n(S_{i+1})\right)and\left(k2(S_i)=k2(S_{i+1})\right)$$
$$then\left(S_j=S_i\cup S_{i+1}\right).$$

2. If the first world is noun and second word is noun of personal name too, their characteristics of case and number are similar, the words are made collocation. For example, "*Dnipro river*".

$$if\left((cm(S_i)=1)\,and\,(cm(S_{i+1})=1)\right)and\left(g(S_i)=g(S_{i+1})\right)$$
$$and\left(n(S_i)=n(S_{i+1})\right)\ then\left(S_j=S_i\cup S_{i+1}\right).$$

3. If the first world is verb and the second word is noun in genitive case, the words are made collocation. For instance, "*read book*".

$$if\left((cm(S_i)=5)\,and\,(cm(S_{i+1})=1)\right)and$$
$$\left((g(S_{i+1})=2)\vee(g(S_{i+1})=4)\right)\vee\left(g(S_{i+1})\neq 1\right)$$
$$then\left(S_j=S_i\cup S_{i+1}\right).$$

4. If the first world is verb, second word is preposition and third word is noun in subjective case the words first and third are made collocation. For example, "*created for children*".

$$if\left((cm(S_i)=5)\,and\,(cm(S_{i+1})=9)\ and\,(cm(S_{i+2})=1)\right)and$$
$$\left(g(S_{i+2})\neq 1\right)then\ \left(S_j=S_i\cup S_{i+1}\cup S_{i+2}\right).$$

5. If the first world is verb and second word is pronoun not in subjective case, the words are made collocation. For instance, "*integrated scheme*".

$$if\left((cm(S_i)=1)\,and\,(cm(S_{i+1})=1)\right)and\left(g(S_i)=g(S_{i+1})\right)$$
$$and\left(n(S_i)=n(S_{i+1})\right)\ then\left(S_j=S_i\cup S_{i+1}\right).$$

Using developed rules and according finding regularity it had been possible to create a system for context units identification. For the complex Ukrainian language sentence the system creates such context units (Figure 1).

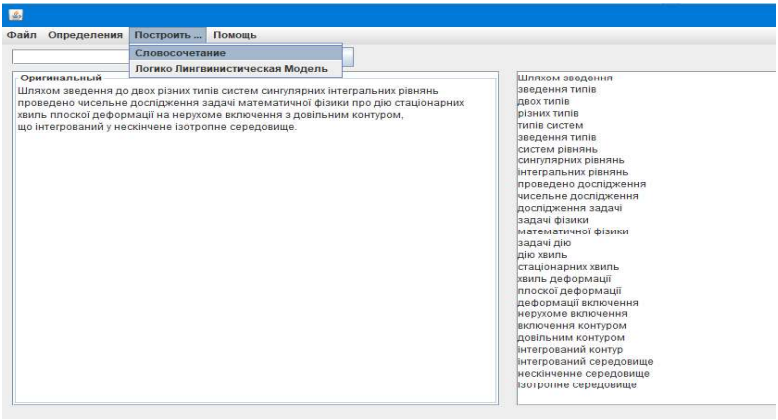

Figure 1 – Results of automatic identification of context units

It had been demonstrated the outcome of using these rules. It is a program, that extract all collocations from different types of natural language sentences.

References:

1. V. Evans, Lexical concepts, cognitive models and meaning-construction. in: Cognitive Linguistics, Edinburg university press Publ. Vol. 17, 2006, pp. 73-107.

2. F. H. George The foundations of cybernetics, Gordon and breach science publishers Ltd., U.K., 1977.

3. W. Che, Y. Zhang, Deep learning in lexical analysis and parsing, Springer Nature Singapure Pte Ltd., ch.4 in: Deep learning in Natural Language Processing, 2018, http:// doi.org/10.1007/978-981-10-5209-5_4.

4. Y. Zhang Discriminative syntax-based word ordering for text generation, in: Computational linguistics, Vol.41, 2015, pp. 503-538.

5. D. Chen, C. D. Manning, A fast and accurate dependency parser using neural networks, in: *p*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp.740–750.

6. A. Bashmakov, I. Bashmakov Intellectual information technologies, M., MGTU Publish, 2005.

7. A. Vavilenkova, Basic principles of the synthesis of logical-linguistic models, in: Cybernetics and systems analysis, Vol. 51(5), 2015, pp. 826-834, http:// doi.org/10.1007/s10559-015-9776-z.

8. V. Shyrokov, System semantics of explanatory dictionaries, in: Cognitive Studies, Vol. 12, 2015, pp. 95-106.

9. A. Vavilenkova, Analysis and synthesis of logic and linguistic models, TOV "SIK GROUP Ukraine", 2017.

10. A. Broder, Identifying and Filtering Near-Duplicate Documents, in: proceedings of the 11th annual symposium, Canada, Montreal: Springer, 2000. pp. 1–10.

[1] **Hlib Yefremov**
Student of the Department of Technology Management
[2] **Kateryna Kolesnikova**
Doctor of Technical Sciences, Professor of the Department of Technologies Management
[1,2] *Taras Shevchenko National University of Kyiv*

OPINION MINING METHODOLOGY IN MARKET RESEARCH

Abstract. The objective of this paper is to present the relevance of using opinion mining methods and tools in market research. During the research, we've established that opinion mining methods proved to be of significant capability to provide insights and conclusions needed for developing market strategy.

Keywords: Opinion Mining, Data Mining, Sentiment Analysis, Marketing.

Marketology is moving at a great pace to becoming a lot more significant and valuable than it was a couple of years ago. With the number of gadgets and digital services common people are using gradually growing, the amount of data trail they leave in cyberspace is growing almost exponentially, left basically as dead weight, unused. Analyzing these data trails is, most certainly, crucial for obtaining and/or maintaining a good grasp of your clients' preferences and opinions. Opinion mining methods is a great way to find out the latter.

Textual information in the world can be broadly classified into two main categories, facts and opinions. Facts are objective statements about entities and events in the world. Opinions are subjective statements that reflect people's sentiments or perceptions about the entities and events. Much of the existing research on text information processing has been (almost exclusively) focused on mining and retrieval of factual information, e.g., information retrieval, Web search, and many other text mining and natural language processing tasks. Little work has been done on the processing of opinions until only recently. Yet, opinions are so important that whenever one needs to make a decision one wants to hear others' opinions. This is not only true for individuals but also true for organizations[1].

One of the main reasons for the lack of study on opinions is that there was little opinionated text before the World Wide Web. Before the Web, when an individual needs to make a decision, he/she typically asks for opinions from friends and families. When an organization needs to find opinions of the general public about its products and services, it conducts surveys and focused groups. With the Web, especially with the explosive growth of the user generated content on the Web, the world has changed. One can post reviews of products at merchant sites and express views on almost anything in

Internet forums, discussion groups, and blogs, which are collectively called the user generated content. Now if one wants to buy a product, it is no longer necessary to ask one's friends and families because there are plentiful of product reviews on the Web which give the opinions of the existing users of the product. For a company, it may no longer need to conduct surveys, to organize focused groups or to employ external consultants in order to find consumer opinions or sentiments about its products and those of its competitors. Finding opinion sources and monitoring them on the Web, however, can still be a formidable task because a large number of diverse sources exist on the Web and each source also contains a huge volume of information. In many cases, opinions are hidden in long forum posts and blogs. It is very difficult for a human reader to find relevant sources, extract pertinent sentences, read them, summarize them and organize them into usable forms. An automated opinion mining and summarization system is thus needed[2]. Opinion mining, also known as sentiment analysis, grows out of this need.

Given a set of evaluative text documents $\boldsymbol{D}$ that contain opinions (or sentiments) about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document $\boldsymbol{d \in D}$ and to determine whether the comments are positive, negative or neutral[3,4].

Opinion mining and summarization process contain three main steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization(Figure 1).

Opinion Retrieval is the process of gathering review text from review websites. Different review websites involve reviews for products, movies, hotels and news. Information retrieval techniques like web crawler can be applied to accumulate the review text data from many sources and store them in database. This step includes retrieval of reviews, micro blogs, and user's comments.

Next basic step in opinion mining is classification of review text. Given a review document $\boldsymbol{D = \{d1…..d1\}}$ and a categories set $\boldsymbol{C = \{positive, negative\},}$ sentiment classification is to classify each di in $\boldsymbol{D}$, with a tag expressed in $\boldsymbol{C}$. The method involves classifying review text into two forms namely positive and negative.

Summarization of opinion is a main part in opinion mining process. Summary of reviews provided should be established on lineaments or subtopics that are mentioned in reviews/blogs/comments etc.

Conclusions. We established that opinion mining is an emerging methodology of data mining applied to summary the knowledge from large volume of data, left by people in the Web, and it's a promising new domain which can greatly increase the quality of the potential market research for your needs. For example, it is critical for a product manufacturer to know how consumers perceive its products and those of its competitors. This information is not only useful for marketing and product benchmarking but also useful for product design and product developments.

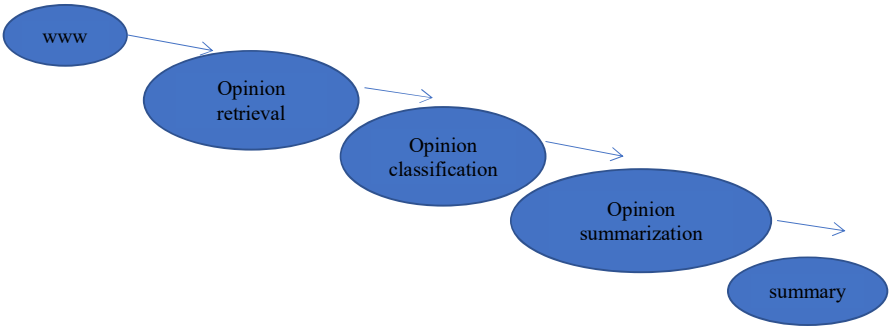

Figure 1. Architecture of Opinion mining.

References:

1. 1. Carenini, G., Ng, R. and Zwart, E. “Extracting Knowledge from Evaluative Text”. Proceedings of the Third International Conference on Knowledge Capture (K-CAP’05), 2005.

2. Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher"A Survey Paper: Areas, Techniques and Challenges of Opinion Mining",International Journal of Computer Science, November 2013

3. McKinsey&Company. “Marketing & Sales Big Data, Analytics, and the Future of Marketing & Sales” (March 2015).

4. Michael Svilar, Arnab Chakraborty and Athina Kanioura “From hype to real help: Finding valuable consumer insight in a stream of data”.

5. Bailey MJ, Muth RF, Nourse HO. 1963. A regression method for real estate price index construction. Journal of

6. the American Statistical Association 58: 933–942.

7. Bailey MJ, Muth RF, Nourse HO. 1963. A regression method for real estate price index construction. Journal of

8. the American Statistical Association 58: 933–942

[1] **Vladyslav Yeshchenkov**
Student of the Department of Management Technology
[2] **Olga Mezentseva**
Candidate of Economic Sciences, Associated prof. of the Department of Technology Management
[1,2] *Taras Shevchenko National University of Kyiv*

IDENTIFICATION OF THE MAIN PROBLEMS OF COLLECTION AND ANALYSIS OF SPEECH DATA USING MACHINE LEARNING

Abstract. The main idea of this paper is to identify and present the main problems that data analysts face when collecting and analyzing speech data. The development of human-machine interfaces has been developing very rapidly nowadays. Invention of voice assistants, smart house technologies is a confirmation of that. Speech technologies are the latest technologies of the 21st century that are used to control computers, cars, and household appliances using voice.

Keywords: machine learning, data analysis, automatic speech recognition.

Despite the large number of researches, modern speech recognition systems remain insufficiently perfect, many problems associated with the process of automatic speech recognition remain unsolved. This indicates the urgency of studying the process of recognizing speech signals and the development of algorithms and methods for implementing this process.

For a machine learning system being able to learn, a huge amount of data is required. This becomes a problem when we speak about audio data. We face into privacy and data security issue. Biometrics like voice, face, fingerprints, and other personal traits are widely used as robust features to identify individuals in authentication systems. It is important to keep the biometric data secure to protect the privacy of users, and we require privacy-preserving machine learning algorithms that can perform the authentication using the secure data. That's why the speech data that collected by personal devices is confidential and barely can be used in data analysis. The actual workaround is using of data generated through paid research and studies, but it is very limited approach.

Speech recognition systems can be sufficiently accurate when trained with enough data having similar characteristics to the test conditions. However, there still remain many circumstances in which recognition accuracy is quite poor. These include moderately to seriously noisy or reverberant noise conditions, and any variability between training and recognition conditions with respect to channel and speaker characteristics (such as style, emotion, topic, accent, and language). While systems are getting better there's still a big difference in their ability to understand different accents of Ukrainian for example. And even a simple cold can be a reason for voice commands not to work as well as usual. In other case, when there is too much background noise speech recognition will be challenging. Making it especially hard to use them effectively in the urban outdoors or large public spaces/offices. For the purpose of

cleaning speech signal from noise is currently successfully applied method of wavelet transformation of an audio signal. Using this method, it seems possible to isolate voice on audio recording, even if present strong background noise [1].

Speech has no natural pauses between the word boundaries, the pauses mainly appear on a syntactic level, such as after a phrase or a sentence. This introduces a difficult problem for speech recognition — how should we translate a waveform into a sequence of words? One way to simplify this process is to give clear pauses between the words. This works for short command-like communication, but as the possible length of utterances increases, clear pauses get cumbersome and inefficient. Additional complexity appears in languages which are phoneme-based, which is typical for Slavic languages. So, the words get extra parts and become different to those from dictionaries. This complicates the process of speech analysis.

Natural language has an inherent ambiguity, i.e. we can not always decide which of a set of words is actually intended. The main ambiguity concept is homophones. The concept “homophones” refers to words that sound the same, but have different orthography or meaning. How can we distinguish between homophones? It’s impossible on the word level in ASR, we need a larger context to decide which is intended [2].

Conclusions. In this paper, we identified the main problems of automatic speech recognition, but not all of them. But one thing we can say certainly, audio data analysis is a challenging process. And as we can see, the most problematic issues are connected to the large search space and the input data variability. Thus, the speaking style, speed and even gender of person can cause errors in speech recognition or learning the system. This is why the search for ideal methods and algorithms for speech processing is still ongoing. And with the growing demand for human-machine interfaces, it becomes even more relevant.

References:

1. M.Z.Hein, The current state of the problem of processing, analysis and synthesis speech signals, 2018. URL: http://www.mathnet.ru/links/6fc916ae022b05118d03dd6b04428d3c/cn182.pdf.

2. Markus Forsberg, Why is speech recognition difficult, 2003. URL: https://www.researchgate.net/publication/228763868_Why_is_speech_recognition_difficult.

[1] **Dmytro Zhovtukhin**
Master student of the Department of Management Technology
[2] **Oleksii Yehorchenkov**
PhD, Associate Professor of the Department of Technology Management
[1,2] *Taras Shevchenko National University of Kyiv*

CLASSIFICATION OF BOTTLES IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

Abstract. The purpose of this paper is to present the relevance of used bottles classification, as well as to developed efficient deep learning models. During the research, we've trained several convolutional neural networks to choose optimal one. As a result, we`ve obtained separate models with height accuracy score and small time of performance on test dataset.

Keywords: Deep Learning, Neural Network, Image Classification.

Annually in Ukraine about 12-15 million tons of household rubbish are generated [1], 40% of which is used for packing. Almost 96% of all trash, including plastic, is sent to landfills, where it poisons the earth for years. For example, a plastic bag decomposes for 500 years, an ordinary bottle of water - a millennium. Ukrainians started sorting garbage only in 2018, but despite this, the problem of recycling is still acute.

To facilitate the process of sorting garbage, we offer to use a system of bottles classification based on the processing of their images by convolutional neural networks. Automating the sorting process will help to speed up the recycling and, in the long run, save money that is now spent on manual labor. The reason of choosing this method is its similarity to human analysis. The machine, as well as person, finds some patterns with differ one type of bottles from others. Same approach very popular in machine learning world to classify objects that can be represented by their images.

For training was chosen 2 popular architectures: ResNet-18 [2] and Mobilenet_v2 [3]. First one famous for its accuracy, second for small computational needs. In ResNet was added one block of Linear, ReLu, Dropout layers, Mobilenet has two additional blocks.

Due to the fact that the available dataset was not found, for some time was collected personal dataset with such classes: plastic bottles with 1206 images, glass bottles with 561 images and 902 images of aluminum cans. Each class was split on 3 sets for train, validation and test. For test - 100 images by class and others images divided into test and validation by ratio 8:2.

The dataset is small for independent training. Two approaches have been used to solve this problem. First, a pre-trained model on PyTorch was used. So, all layers except few last one was frozen to save previous coefficients. Secondly, augmentation was added to the training pipeline. Sometimes one of these changes could occur: horizontal flip, vertical flip, random affine transformation, color changes.

As metric was used accuracy, as loss function – weighted cross entropy, optimizer - Adam. Epoch number – 75, batch size – 25, initial learning rate – 0.001 with decay rate optimization. GPU provided by Google Colaboratory was used for training acceleration.

As the result we have 4 models: ResNet and Mobilenet_v2 without augmentation and with it. Their performance can be seen on Table 1.

Most important features are accuracy, size and speed. Most errors occur plastic bottles class that was false predicted by developed models. Size of ResNet-18 are bigger and speed are slow as was expected. But its accuracies are better. In the other case, Mobilenet_v2 fast and small and not so bad in accuracy score.

It is interesting that the best loss score is 0. It may cause by overfishing because model perfectly learned training set. So, it is better to avoid of using that model.

Table 1

Models performance results

	Test Accuracy	Best Loss	Size, MB	Training time, sec.	Mean speed on CPU/GPU, sec.
ResNet-18	0.9967 (best)	0.000 (best)	42.94259 (worst)	1325.55	0.188/0.0138
ResNet-18 with augmentation	0.9967 (best)	0.005	42.94212	4080.65 (worst)	0.198/0.0139 (worst)
Mobilenet_v2	0.9834	0.004	11.32676 (best)	1269.44 (best)	0.102/0.0157 (best)
Mobilenet_v2 with augmentation	0.98 (worst)	0.017 (worst)	11.32725	3821.54	0.148/0.0157

Conclusions. We established that there isn`t model that can provide both heigh accuracy and quick performance at the same time. If the bottles classification system must run in real time on a device with low computational potential, is better to choose Mobilenet_v2 architecture with no augmentation during training process. For the best accuracy score, ResNet-18 architecture is more wisely to choose. Despite slightly worse results of speed and loss score, we recommend using augmentation in order to avoid overfitting. Also, this work can be the basis for classification system of used bottles and shows a prospect for further research.

References:

1. State Statistics Service of Ukraine: http://www.ukrstat.gov.ua/.
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2015). Deep Residual Learning for Image Recognition.
3. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks.