IDENTIFICATION IN INFORMATIVE SYSTEMS ON THE BASIS OF USERS' BEHAVIOUR

I.V. Ruban
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
ihor.ruban@nure.ua

V.O. Martovytskyi
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
vitalii.martovytskyi@nure.ua

A.A. Kovalenko
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
andriy.kovalenko@nure.ua

N.V. Lukova-Chuiko
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
lukova@ukr.net

***Abstract*— Distributed informative systems which unite the technology of client's server with global web have posed numerous problems. It turned out that standard methods of identification have already become obsolete. Particularly the problem is that the generally accepted division of methods of a control over physical access and a control over access to information are not effective any more. To solve this problem there is a need to apply for the methods of identification which will have possibility to identify users with the help of aggregate of actions implemented by users in the process of work with DIS.**

Keywords*— Distributed informative systems, Neural network, access to authentication, identification

I. INTRODUCTION

Identification and authentication can be considered to be the basic program and technical means of security as the rest services are turned to maintain named subjects. Identification and authentication are the first line of defense of organization's informative zone

According to the report [1] by the institute SANS «IT Security Spending Trends» which indicates the percentage of cost of organizations to ensure different technical means of informative security. Figure 1 presents the diagram of every technological unit cost.

Having analyzed the diagram it is possible to determine two types of technologies of cybersecurity which are presented by the most important such as:

- access to authentication;
- broaden defense from malware;

Distributed informative systems now become more popular than before. It is doubtful that there is possibility to find any app which does not use components from various suppliers. The more difficult the modern apps are the more needs there are to use components distributed by remote machine. Services which are used by web-apps are very often placed not on the territory of the institution's informative center.

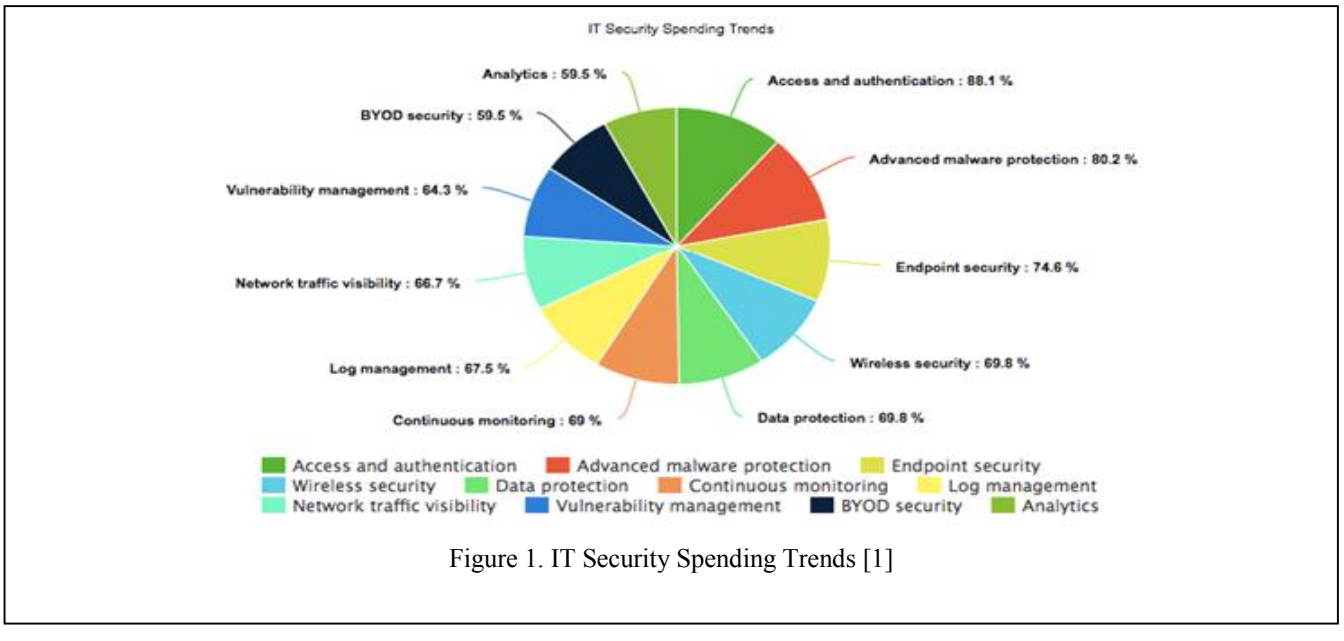

Figure 1. IT Security Spending Trends [1]

978-1-7281-1814-7/19/31.00 © 2019 *IEEE*

Distributed informative systems which unite the technology of client's server with global web have posed numerous problems. It turned out that standard methods of identification have already become obsolete. Particularly the problem is that the generally accepted division of methods of a control over physical access and a control over access to information are not effective any more. To solve this problem there is a need to apply for the methods of identification which will have possibility to identify users with the help of aggregate of actions implemented by users in the process of work with DIS.

II. Analysis of References and Posing of a Goal

The defense of information of computer systems and webs is a complex task the solution of which can be done with the help of introduction of different security systems. One of the main roles in the task solving is played by the element which ensures the control over the access to the resources of computer systems. This element implements its functions using the procedure of identification and authentication of users.

The classic method of users' authentication is the use of unique information – the password which is known by the user and which is presented by him in the process of authentication [2].

Biometric authentication is a prospective direction which allows solving numerous tasks which arise in the process of using traditional procedures of authentication.

Nowadays there are two groups of methods of biometric authentication which are based on statistic methods and on dynamic methods [4]. Having analyzed the state of security of modern methods of authentication [4-5] it is possible to conclude that the method based on the estimation of emotional state and mimics has numerous disadvantages such as:

- Strong dependence on psycho-physical state of an operator. If the person is ill it is possible that he can fail his authentication.
- Strong dependence on the means received from the biometrical characteristics.

Existing program realizations of such systems are characterized by insufficient reliability of authentication. Relevant formation of new methods, algorithms and their program-apparatus realizations increase the effectiveness of systems of identification and authentication.

Thus there is a prospect to create methods and algorithms of identification and authentication based on the IS users' behavior.

III. The Main Material

The most wide-spread today are the methods of users' identification based on using passwords which, unfortunately, may be lost, stolen or compromised in numerous ways. Thus it can be great to implement the idea of a combination of a standard password with the method of users' identification based on their behavior in system. In this case even if intruder gets access to the password the general access to the computer system may be denied due to the behavior identification that is depicted on Fig. 2. In this case the structure of identification due to user's behavior in the system is the following: if the user types incorrect password he is denied in access immediately. If the password is correct in some duration of time T there has place the process of gathering data about his behavior after that this data is compared with the one registered user's example from the data base which has already come through identification. Depending on the required accuracy in the process of data comparing the access can be allowed or denied.

The task of identification may be divided into several stages the main of which are the process of teaching system and, relevantly, the process of recognition. At the beginning the system takes and saves into the data base the information of user's behavior in system; based on the values of these settings there is a template or a profile of user's behavior. Then there occurs the process of comparing the given template with already stored in the data base in system that is actually identification.

Identification happens due to the results of a control over user's behavior in the process of working in informative system. As original data there is used matrix X indicators in the proves of monitoring the user's behavior and vector column Y, which consists of multiplicity {0.1}, where 1 occurs if user's behavior is relevant to the identified user and 0 if not.

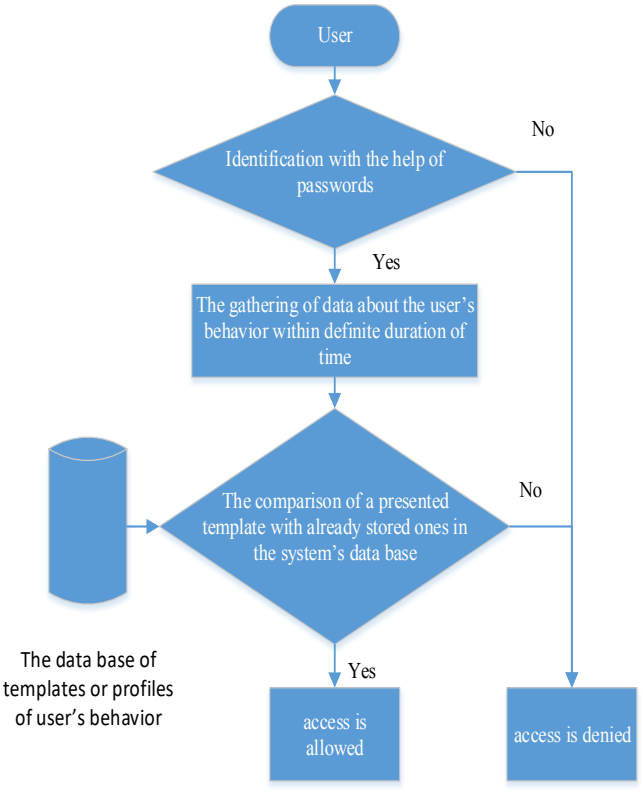

Figure 2. The structure of user's identification depending on his behavior in system

Matrix X consists of vectors which have the following form:

$$x_i = \{S_i^T, ST_i^T\}, i = 1, \dots, l, \quad (1)$$

where S_i^T – is a multiplicity of states of a user within the period of i session, and ST_i^T – is a multiplicity of statistic parameters based on the multiplicity S_i^T, for example the duration of user's presence in any state out of many multiplicities S_i^T, the average time of his presence in any state

S_i^T and others. . The sessions are determined so that they can not be longer than time *T* or has more than *B* states. Thus the session may be considered finished when the user is in *B* states or when session takes more time than *T* time.

The task of template forming of user's behavior will consist of function construction

$$\alpha: X \to Y, \quad (2)$$

Which can identify random object $x_i \in X$. $x_i \in X$. To construct the function of template α it is possible to use linear model with a vector with parameter $w=\{w_0, w_{1,\dots,} w_n\}$, where *n* is a length of vector x_i. Thus the function of template α has the following form:

$$\alpha(x, w) = w_0 + w_1 x_1 + \cdots + w_n x_n, \quad (3)$$

in the process the task of user's identification can be settled to the task of binary classification with multiplicity $Y=\{-1;1\}$. In this case the user's template will have the following form:

$$\alpha(x, w) = sign \sum_j^n w_j x_j, \ (x_0 = 0) \quad (4)$$

Found parameters will be the templates of user's behavior and will ensure an optimal value of a functionality of a quality. In this task the functionality of mistakes will be minimized that is an average quantity of mismatches, where $L(\alpha, x_i)$ – is a function of loses.

The function of mistakes has the following form:

$$Q(\alpha, X) = \frac{1}{l}\sum_i^l L(a, x_i) = \frac{1}{l}\sum_i^l [a(x_i) - y_i] \to min, \quad (5)$$

To estimate the quality of user's classification there are such measures:

recall:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

precision:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

where TP TP is the quantity of correctly classified identified users, FN is the quantity of incorrectly classified intruders, FP is the quantity of incorrectly classified identified users. The measure which takes into account the balance between the completeness and accuracy is called F-measure and is calculated in the following way:

$$F = \frac{2 * Precision * \text{Recall}}{Precision + \text{Recal}} \quad (8)$$

Another measure of the estimation of binary classification quality is a space under ROC-curve is known as AUC-ROC (Area Under ROC Curve), which formally distinguishes the possibility that the object of class 1 gets the estimation higher than the random object of class-1.

IV. THE RESULTS OF EXPERIMENT

As an experiment it was decided to analyze the user's behavior of PC in the process of using Internet browser. For that within a month through proxy server there was an analyses of user's behavior on Internet, moreover, every user had to access using their own account

After that there was formed a sample which initially contained the following indicators:

- site1 – index of the first visited site during session
- time1 – time of visiting the first site during session
- site12 – index of the 12th visited site during session
- time12 – time of visiting the 12th site during session
- target – target variable takes the value from 0 to 4 in accordance with every user's computer.

The session of users is highlighted in the way that they can not be longer than 25 minutes or contain more than 12 sites. Namely, the session is considered to be finished when a user visited 12 sites in a raw or when the session has taken more than 25 minutes. These parameters have been received from average time of every user's visit of Internet and average quantity of visits of sites per one authorization in system.

After that for every user there has been formed learning aimed sample where the variable item gets the value in 1 for the sessions of an authorized user and in 0 for the sessions of other users.

Session is the sequence of sites' indexes and the data in this form is not useful for the linear methods. We convert this table in the way that every possible site is relevant to its definite feature and its value would be relevant to the quantity of visits of the sites during the session. Then there would be created a feature which will be like the numbers in the way YYYYMM (YYYY – year, MM – month) from the date when of session for example 201804 – 2018 year and the 4th month. So we will take into account monthly linear trend within the whole period of presented data. As well it is possible to add the following feature:

- the feature of the quantity of unique sites;
- Time of the beginning of session (from 0 to 23);
- binary feature is 1, if the session starts in the morning and 0, if the session starts later.

As the process of forming the user's profile uses linear methods it is recommended to bring features to one scale actually that is what has been done at this stage.

After this stage of preparation there has been conducted the analyses of the estimation of the quality of different classifiers which is presented in the table 1.

TABLE I. MEASURES OF THE QUALITY OF DIFFERENT METHODS OF CLASSIFICATION

Method of classification	*F-measure*	*AUC*
Logistic regression	0.834	0.785
Naïve Bayesian classifier	0.8	0.609
SVM	0.852	0.711
Neural network	0.827	0.893

Experiment shows that the best has been neural network at F-measure as well as at AUC.

The next stage of the experiment has been in forming a template of behavior for every user separately and forming of

the general template of users which has been identifying every user of the system. Neural network has been in role of a classifier.

The results of the work are presented at the Picture 3.

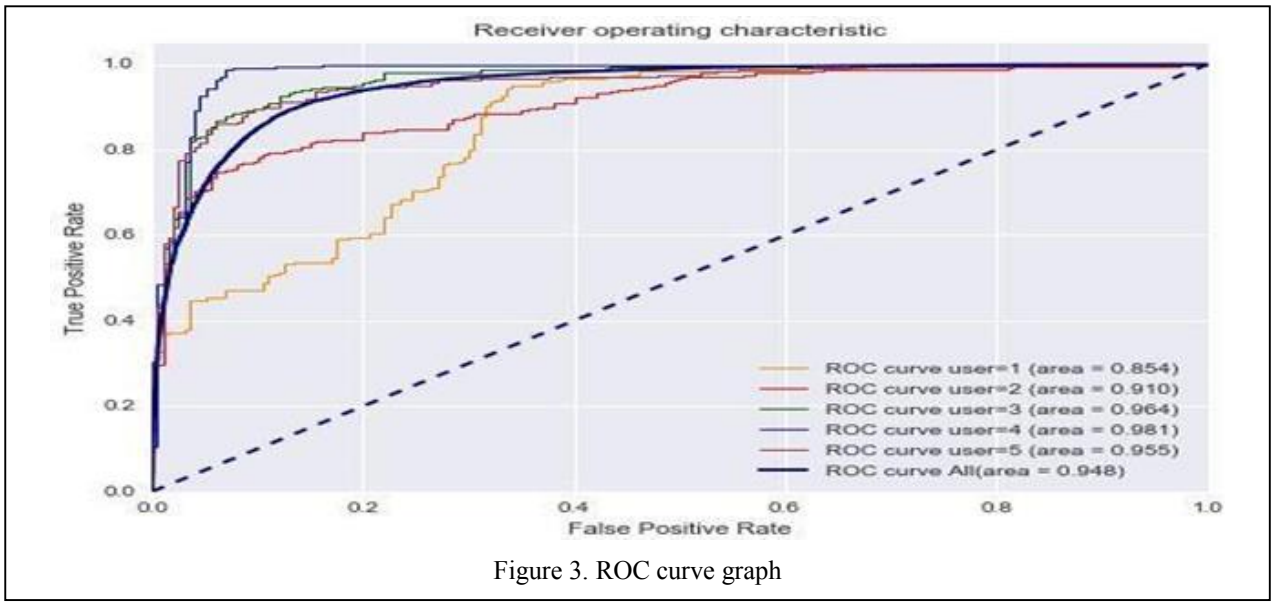

Figure 3. ROC curve graph

Looking at the graphic of ROC-curve it is possible to conclude that multi-dimensional model shows more qualified results. But the usage of the model in a working system of an access control on the basis of user's behavior analyses is ineffective because with the access of a new user arises the need to reteach the whole system in general.

V. CONCLUSION

Summing up the results of the experiment it is possible to distinguish the following advantages and disadvantages of user's identification due to his/her behavior in informative system.

Advantages.

- The easiness of realization and implementation. There is no need in special machine provision, data is taken from the system monitoring of the informative system state that means that the usage does not require the buying of any additional equipment. That is the cheapest way of identification due to biometric characteristics of an access.
- There is no need for user to take any additional steps or have any additional skills, it is impossible to copy a behavior profile of authenticated user.
- The possibility of hidden identification.

Disadvantages.

- At the stage of initial using there is a need in additional time to form a behavior profile.
- The huge dependency on numerous states which may be reflected a definite system's user.

There is also a need to point out that the user's identification using only the analyses of user's behavior is unacceptable in the systems with a high level of security. But in combination with other models and methods of security of informative systems for example with those in mentioned articles [7-10] can appear to be quiet effective.

[11]

REFERENCES

[1] SANS 2016м IT Security Spending Strategies Survey // [Online] https://www.sans.org/webcasts/2016-security-spending-strategies-survey-100997

[2] How to Get the Absolute Most from Your Cybersecurity Budget // [Online] https://www.stickman.com.au/how-to-get-the-absolute-most-from-your-cybersecurity-budget/

[3] Kim H., Lee E. A. Authentication and Authorization for the Internet of Things //IT Professional. – 2017. – Т. 19. – №. 5. – С. 27-33.

[4] Ali M. L. et al. Keystroke biometric systems for user authentication //Journal of Signal Processing Systems. – 2017. – Т. 86. – №. 2-3. – С. 175-190.

[5] Wu F. et al. A lightweight and robust two-factor authentication scheme for personalized healthcare systems using wireless medical sensor networks //Future Generation Computer Systems. – 2018. – Т. 82. – С. 727-737.

[6] Gałka J., Grzywacz M., Samborski R. Playback attack detection for text-dependent speaker verification over telephone channels //Speech Communication. – 2015. – Т. 67. – С. 143-153.

[7] Ruban, I., Martovytskyi, V. and Lukova-Chuiko, N. (2016), "Designing a monitoring model for cluster super–computer", East-ern-European Journal of Enterprise Technologies, Vol. 6, No. 84, pp. 32-37.

[8] Kharchenko V.S., Illiashenko O.A., Kovalenko A.A., Sklyar V.V., Boyarchuk A.V. (2014), "Security informed safety assessment of NPP I&C systems: Gap-Imeca technique", Proceedings of the 2014 22nd International Conference on Nuclear Engineering ICONE22 July 7-11, 2014, Prague, Czech Republic, ICONE22-31175, pp. 1-9.

[9] Ruban I., Martovytskyi 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 //Cybernetics and Systems Analysis. – 2018. – Т. 54. – №. 2. – С. 302-309

[10] G. Kuchuk, A. Kovalenko, I.E. Komari, A. Svyrydov, V. Kharchenko. Improving big data centers energy efficiency: Traffic based model and method. Studies in Systems, Decision and Control, vol 171. Kharchenko, V., Kondratenko, Y., Kacprzyk, J. (Eds.). Springer Nature Switzerland AG, 2019. Pp. 161-183.