behavioral biometrics,
user profile, intruder detection,
user verification

Tomasz WESOLOWSKI[1], Przemyslaw KUDLACIK[1]

USER PROFILING BASED ON MULTIPLE ASPECTS OF ACTIVITY IN A COMPUTER SYSTEM

The paper concerns behavioral biometrics, specifically issues related to the verification of the identity of computer systems users based on user profiling. The profiling method for creating a behavioral profile based on multiple aspects of user activity in a computer system is presented. The work is devoted to the analysis of user activity in environments with a graphical user interface GUI. Mouse activity, keyboard and software usage are taken into consideration. Additionally, an attempt to intrusion detection based on the proposed profiling method and statistical measures is performed. Preliminary studies show that the proposed profiling method could be useful in detecting an intruder masquerading as an authorized user of the computer system. This article presents the preliminary research and conclusions.

1. INTRODUCTION

Companies, offices and institutions often process and share a lot of confidential information. Among them financial documents, ongoing contracts and project documentations together with personal and sensitive data of employees or business partners can be found. This fact brings the need to protect these documents and data sets against unauthorized access, which often can be crucial for the company. It is therefore necessary to take such a security measures for the IT systems that will allow very high level of access control. The goal is to allow the access to the specific resources only to the authorized individuals. Access control is a task carried out at the level of the operating system that either allows or refuses to perform an operation basing the decision on rights assigned by the administrator. The rights are assigned to specific groups or individual users, in line with the institution's security policy.
It is necessary however to confirm the identity of the person working at the computer. The most common method of authentication is "log in" mechanism that requires a user name (login) and password. This type of authentication is a onetime operation usually associated with the moment of starting a work in the computer system. However, according to the open character of many of the computer stations (like those in hospitals dealing with sensitive data) this kind of security measures is not sufficient and to increase the level of security more advanced methods are necessary.
One of the solutions is the use of biometrics, which identifies users by their individual physical characteristics, for example, methods based on facial recognition or fingerprint analysis. However, even these methods of recognizing a particular person are not able to guarantee that someone not authorized will not overtake the access to the system where the authorized user is already logged in. Situation like that can happen when the legitimate (authorized) user leaves for some time the workspace after logging in into the system.

[1]University of Silesia, Institute of Computer Science, ul. Bedzinska 39, 41-200 Sosnowiec, Poland
e-mail: tomasz.wesolowski, przemyslaw.kudlacik@us.edu.pl

Table 1. Prefixes and additional data description for recorded user activity.

Prefix	Event	Additional data
M	mouse move	x,y – coordinates of mouse pointer
L	left mouse button down	x,y – coordinates of a mouse click
l	left mouse button up	x,y – coordinates of a mouse click
R	right mouse button down	x,y – coordinates of a mouse click
r	right mouse button up	x,y – coordinates of a mouse click
S	mouse scroll	x,y – coordinates of mouse pointer
K	key down	encrypted code of the key
k	key up	encrypted code of the key
W	window switched	encrypted window name and optionally application name

To ensure a high level of security of information systems it should be continuously monitored if the working person is the legitimate user. Such a monitoring is performed by the Intrusion Detection Systems (IDS) that constantly monitor all operations performed by users, and then try to verify user's identity. The basis of this authorization method is the analysis of user's activity during the interaction with the computer system. Based on this information user's profiles are created and by means of various methods the reference profiles are compared with the user's activity data. If such a comparison is made on the fly (in real time) we are dealing with the so-called online IDS and if the collected activity information is compared after some time it is an offline IDS. The proposed in this paper user profiling method can be used in both of those IDSs, however due to the necessity to analyze long periods of activity time it is best suitable for offline IDS.
Subsequent sections of this paper describe the format of the data used in the experiments, present state of the art in profiling computer system users, introduce the profiling method that combines the usage of mouse, keyboard and software and finally present the conclusions obtained from the results of experiments.

1.1. Data Set

The data set was created using the dedicated software implemented in the Computer System Division of Computer Science Institute at University of Silesia. The software works in background of an operating system registering user's activity at every moment of his work. Despite this, the application is unnoticeable to the user and does not affect the comfort of work. In order to ensure a sufficiently high level of security of the data processed a number of mechanisms protecting the data was implemented. First of all, the alphanumeric keys are coded together with the names of the windows in which the user is working. Only the codes of functional keys are stored in an unencrypted form. In order to encode the data the MD5 function and unique seeds were used. In some cases the seed of the encryption was shared by different users in order to allow the simulation of an intruder attack for the studies based on keyboard usage. The software records activity saving it in log files compatible with the CSV format. The first line starts with a token RES followed by the resolution of the screen. Starting with the second line the activity data is stored and as it can be observed most of it consists of mouse events. Each line starts with a prefix identifying the type of event. The prefix is always followed by the time stamp and optionally additional data related to the event. The prefixes and corresponding event types together with additional data information are presented in Tab. 1

The number of files and size of data recorded for each user is shown in Tab. 2.

2. State of the Art

In behavioral biometrics the idea of user verification based on users activity in a computer system is very well known. For a long time a human-machine interaction was based on giving text commands that the machine could recognize and process. It was possible to record these commands and analyze

them in order to create a characteristic user profile. Very popular work in this area was performed by Schonlau et al. – SEA. His team collected a set of data consisting of text commands [12] and conducted a series of studies related to statistical analysis of this data and detection of intruders – in particular masqueraders [11]. Our work concerning SEA data set and intrusion detection based on text commands analysis was presented in [8]. Another approach consisted of research that were carried out towards analyzing the direct use of the keyboard. The method proposed in citeidWesolowski: Bleha1990 focuses on how the user enters his personal password, when logging on to the system. The descriptions of parameters used to indicate the quality of the methods are presented in Tab. 3. The FN ratio for the method was 1,75%, and the FP ratio 0,43%. Another method based on the keyboard dynamics was proposed by Monrose [7]. The focus was not on what is entered, but how is it done. Individual rhythm of typing was determined and any deviations from the average values were treated as an alarm. The effectiveness of this solution oscillated between 80-90%. Along with the development of computer systems due to the increasing computing power the way of human-computer communication also has evolved. Nowadays, most of computer systems is running operating systems with GUI. To facilitate the work of computer users, so called Human Interface Devices (HID) were created that allow user to control the graphical environment easily. For this reason, currently attempts of user profiling and intrusion detection using behavioral biometrics are based on data of user activity associated with HID. The most popular HID is a computer mouse. In [2] the dynamics of mouse usage based on the moves characteristics is presented. Efficiency of identity verification reached approx. 95%. Another solution is based on the separation of mouse events (move, click) and their subsequent analysis [9]. The distance, angle and speed of movement of the cursor was calculated and then the average values were determined. Gross deviations from the calculated average values consisted anomalies that were classified as the presence of an intruder. False alarms (FP) accounted for 0.43%, while the undetected intrusions (FN) 1.75%. Other approach was based on calculating the vector of characteristic features [5]. The calculated vector includes information about the average number of the left and right mouse button clicks and a thorough analysis of the trajectory of the cursor. The effectiveness in verifying the identity of the person was around 80%. There are also solutions using both data from the keypad, or mouse. In [1] a detailed analysis of the movements made with a mouse was made, next the average number of clicks in specific fields of the screen was determined, and the number and type of keystrokes was examined. Then an attempted to verify that the user is legitimate was made by comparing the current activities with the typical use of the mouse, and keyboard. When relying on an analysis of mouse activity the effectiveness of the method was 91%, in the case of the keyboard 53%. This solution takes into account the activity associated to the use of both of the mentioned devices, but the information gathered is not treated as a whole, and is divided into two independent approaches. The new approach to user profiling was based on an analysis of how they interact with different types of software. An example would be a method based on the use of e-mail client software [14]. The analysis of the various aspects related to, among others, the type of attachment, size and number of incoming and outgoing messages, as well as a list of recipients was performed. When a significant deviation from the average value of each parameter was detected the software reported an alarm. False alarms ratio was around 30-50% and 5-10% of intrusions were undetected. Another method [4] analyzed mainly the content of individual e-mail messages focusing on the linguistic analysis, and thus related to the individual way of writing sentences, vocabulary used, the length of words, etc. Such a solution achieved efficiency in verifying the identity of approx. 80%. The above methods apply either the keyboard or mouse usage information, rarely both of them. However, a user working in a graphical environment uses both of these devices at the same time. The approach presented in this article is based on both: the data related to the use of the keyboard and mouse combined, and further expanding the data base by information on the usage of computer programs.

3. PROFILING A USER

Basic principle of the method proposed in this paper is to create a characteristic profile of each individual user working in the computer system. The individual profile should reflect as closely as possible the way the user works in the graphical environment. This allows the monitoring system to have knowledge of the typical behavior of individual users. As a result, this will enable the identification of any unusual situations, especially those that might indicate a detection of an intruder. The behavioral profile is created based on the training data provided to the profiling system. In this case, these are the files described in Sec. 1-1. The profile describes the way individuals work and is presented as a vector of features. Different types of events are counted and then the average number of occurrences in the analyzed data by 15-minute work period is calculated. This value is an element of the relevant vector of features. The time interval of 15 minutes was determined experimentally. Each feature is represented as a numerical value that reflects its level of severity. The length of the vector is therefore dependent on the number of extracted behavioral features. Each of them should be chosen to constitute the most unique value that characterizes user actions. For example, a feature indicating the number of used screens would not meet this assumption, as its value would be mainly 1, rarely 2, and occasionally 3. The number of screens is also not at all directly related to the way a user works. And what is most important, the value of this feature does not enable the detection of an intruder. Therefore an important issue is the right choice of the features stored in the profile.

3.1. PROFILE FEATURES

The proposed profile consists of five groups of features related to different types of recorded events. The first part of the profile consists of 13 elements and is related to the organization of work by the user. Any person manages the time in a different way and has a different rhythm. Some users after starting the system work constantly, while others do more or less frequent breaks. The number and length of these interruptions is dependent on the person, and may be one of the individual characteristics. Each event recorded in the log file carries the time stamp of its occurrence. On this basis, it is possible to specify the intervals between occurrences. If the break lasts less than 60 seconds it is not considered to be a break at work. On the basis of the length the interval is assigned to one of 13 groups. The indexes 1 to 9 are respectively a break of 1-2 minutes, 2-3 minutes, ... , up to 9-10 minutes. Than the values of the intervals are increased from one to five minutes, so that the indexes 10 to 13 represent a break of 10-15 minutes, ... , up to 25-30 minutes. Next all the 13 groups are counted and stored in the profile. As a result basing on the first group of features it is possible to specify how many breaks and of what length occurred in the analyzed user activity data.

The second part of the profile consists of 4 values related to a computer mouse usage. Successively the features are associated with: single and double click of the left mouse button, clicks of the right mouse button and the number of mouse wheel rotations (scrolls). All of those values are a numerical representation of how the mouse is used when interacting with a computer. Depending on the experience of the user the individual tasks are variously accomplished. For example, some people often use main menu while selecting the desired option, while others prefer to call a function from the context menu. Both ways have different characteristics in the recorded activity. As a result, an increase in the number of events, involving respectively the left or right mouse button click can be observed.

Another group of extracted features relates to the use of the keyboard. Observing the frequency of pressing specific keys it can be determined how the user performs various actions in a computer system with a graphical interface. This part of the profile vector consists of 36 items, each of which is associated with another key. In the presented method the dynamics are not at all taken into consideration.

The penultimate group in a vector of features constituting the user behavioral profile refers to events related to the system windows and running applications. The group consists of 3 values: the number of different windows opened, how often the user was switching between them and the average time of working with each window.

The last group of elements of the vector is the longest and is based on the previous three. It is related to the usage of computer mouse, a keyboard, and the parameters associated to the windows. The individual values are calculated in the same way, but they represent the characteristics of user work when using one of 10 defined computer programs that are identified by the software for activity registration.

3.2. COMPARING PROFILES

For the IDS to be able to determine whether a working at the moment computer uses is a legitimate user, it must have the information about user's typical behavior. This information specifying the individual characteristics is presented as a behavioral profile described earlier. Comparing the current user's activity with that observed previously and saved as a profile allows to specify the degree of similarity in the manner of interaction with the computer system. As a result, it allows the verification of the person working at the computer. In order to verify the user an understandable for IDS value, saying how much current and the reference profile are similar, must be designated. In the case of this method the mutual correlation of the profiles is calculated. For this purpose, the Pearson linear correlation coefficient r_{xy} was used (Eq. 1).

$$r_{xy} = \frac{\sum_{i=1}^{u}(x_i - x')(y_i - y')}{\sqrt{\sum_{i=1}^{u}(x_i - x')^2}\sqrt{\sum_{i=1}^{u}(y_i - y')^2}}, \tag{1}$$

Overall, the correlation is used to measure the strength of the relationship between the variables X and Y [13]. In this case, both of these variables have the form of a feature vector calculated individually for each user based on the recorded activities - they consist profiles of a user. In (1) x_i is the value of i-th element of the vector X, and x' is the arithmetic mean of all its elements. Respectively the same meaning have y_i and y' referring to the vector Y. While the value of u is the length of the vector, and thus the number of its elements (it is the same for X and Y). The calculated value of r_{xy} must be in the range of <-1,1>. The closer it is to 1, the greater the degree of similarity between the variables. Correlation coefficient of 1 means 100% similarity of the compared vectors, and so that they are identical. The IDS can use the presented method to verify the user's identity by calculating the correlation coefficient of the reference profile (X), with the current profile (Y). The value obtained as a result is the basis to determine whether a user working at the moment is a legitimate user or a masquerader. The next step in the process of authentication, significantly contributing to its effectiveness, is the appropriate designation of the parameter t - the threshold value of the correlation coefficient, which allows to classify a person as the legitimate user or an intruder. If the value of r_{xy}, indicating the degree of similarity of the compared profiles, is lower than the assumed value of the parameter t, then the IDS reports an alarm. This situation occurs when the current profile is not sufficiently similar to the reference profile of selected user. When the value of r_{xy} is equal to or greater than t, then one can conclude that the system is used by the authorized person. It is therefore important to set the value of t optimally, otherwise it will result in an erroneous verification of user's identity. The difficulty in choosing the appropriate threshold is one of the drawbacks of this type of IDS [6] because the number of false alarms and undetected intrusions, resulting in the overall effectiveness of the method, depends on it.

4. EXPERIMENTS

In order to examine the mechanism described in this article the database of different behavioral profiles was needed. The information containing a user activity was collected using the specially designed software described in Sec. 1-1.

The data was collected on 10 different computers of 10 different users. Each user works alone on his/hers computer, therefore, the data is not contaminated with false activity. Users were selected in a way to represent population diversity of: age, gender, profession and experience in computer usage. The

Table 2. Number of user files and their size in the test database.

user	files	size
user 1	26	31.1 MB
user 2	5	22.3 MB
user 3	4	2.6 MB
user 4	17	72.4 MB
user 5	6	12.6 MB
user 6	4	6.4 MB
user 7	19	17.1 MB
user 8	5	16.7 MB
user 9	16	15.5 MB
user 10	21	184 MB
summary:	123	381 MB

Table 3. Description of examined parameters.

parameter	description	calculation
$alarms$	number of raised alarms	-
$false\,alarms$	number of falsely raised alarms	-
$false\,negative$	number of missing alarms	-
$true\,positive$	number of correctly raised alarms	-
FP	percentage of false alarms (false positive)	$FP = (false\,alarms/alarms) \cdot 100$
FN	percentage of false negative	$FN = (false\,negative/alarms) \cdot 100$
TP	percentage of true positive	$TP = (true\,positive/alarms) \cdot 100$
E	efficiency	$E = TP - FN$

analyzing software was installed and was working approximately for one month. The number of files and size of generated trace of activity for each person was obviously different because of individual frequency and intensity of computer usage. The short description of the database volume is shown in Tab. 2.

Each file of the database was created every time the user's operating system was restarted. It is important to mention that the files can contain information gathered for periods longer than one working day because of hibernation mechanism commonly used in contemporary operating systems. This fact is not an issue though, and does not influence the quality of the stored data.

Behavioral profiles were obtained according to the method described in Sec. 3-1 for each of 123 files. The process produced 123 usage profiles and 10 average profiles calculated for each user in the database.

For examination purposes intruder activity was simulated by comparison of the average profile with each of 123 obtained profiles. Therefore, the simulation concerned activity of 10 different users in one workplace for 10 different computers. In that case it is easy to calculate the test should raise an alarm 1107 times, which generally can be obtained by the following equation

$$N_a = \sum_{i=1}^{10} (N_f - N_f^i) \ , \tag{2}$$

where N_a represents the total number of alarms, N_f represents the total number of profiles (files) and N_f^i stands for the number of profiles (files) for i-th user.

The examinations were performed for two methods of calculating the correlation trigger level t. The first method assumed an arbitrarily given value of t for all users (global value). The tests concerned calculations of several parameters for each t level, which were precisely described in Tab. 3.

The obtained results showing the influence of t value on described parameters are presented in Tab. 4.

The second group of tests were performed for individually calculated value of t that was adjusted to each user. The t level was in this case calculated as an average avg of all t values obtained for all profiles of that one user. The obtained results showing the influence of changing t in described situation are presented in Tab. 5.

Table 4. The influence of globally adjusted values of t parameter on number of alarms and efficiency of the method

t	$alarms$	$false\,alarms$	$false\,negative$	$true\,positive$	FP	FN	TP	E
0,60	1032	22	97	1010	2,13%	8,76%	97,87%	89,11%
0,65	1056	24	75	1032	2,27%	6,78%	97,73%	90,95%
0,70	1089	33	51	1056	3,03%	4,61%	96,97%	92,36%
0,75	1106	36	37	1070	3,25%	3,34%	96,75%	93,40%
0,80	1131	48	24	1083	4,24%	2,17%	95,76%	93,59%
0,85	1155	60	12	1095	5,19%	1,08%	94,81%	93,72%
0,90	1173	69	3	1104	5,88%	0,27%	94,12%	93,85%

Table 5. The influence of individually adjusted values of t parameter on number of alarms and efficiency of the method.

t	$alarms$	$false\,alarms$	$false\,negative$	$true\,positive$	FP	FN	TP	E
avg – 0,10	1041	26	93	1015	2,50%	8,40%	97,50%	89,10%
avg – 0,05	1071	36	72	1035	3,36%	6,50%	96,64%	90,13%
avg	1093	46	60	1047	4,21%	5,42%	95,79%	90,37%
avg + 0,05	1122	65	50	1057	5,79%	4,52%	94,21%	89,69%
avg + 0,10	1161	93	39	1068	8,01%	3,52%	91,99%	88,47%

The results for two described approaches are accordingly depicted in Fig. 1 and 2. The visualization, which is almost linear in both cases, shows the area of Equal Error Rates (EER), which can be assumed respectively at the levels around 3.30% for t near 0.75 and 5% for t near $avg + 0.025$ (see Tab. 4 and 5).

Considering the ERR ratio it can be seen that significantly better results were obtained for the method of global t adjustment. The trigger levels computed individually for subsets of data usually give better results as they are suited to the particular needs of the smaller data range. In opinion of the authors the average is just not the approach working in described environment, therefore, a different form of individual adjustment have to be applied in future research.

Nevertheless the overall analysis indicates that the solution can be very accurate. Probably the obtained level of results is caused by the differentiation of individuals creating the database. The real-world scenarios of computer activity logs will hardly ever have such advantage. However, the very important objective of the research have been evaluated: it is possible to distinguish user's activity in a computer system at the relatively high level of accuracy. That was the main reason of designing the logging software allowing to collect information coming from different aspects of user's work in a contemporary operating system.

The last stage of examinations was performed to analyze the general influence of the learning set size. The idea in this case was very simple: use 4 different user work periods to calculate the average profile and compare the results obtained for all 123 files like it is described earlier. The table 6 shows how the amount of data in the learning set was chosen for all users . Four levels gave four different periods of user's activity, which approximately corresponds to $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ and to a whole length of the

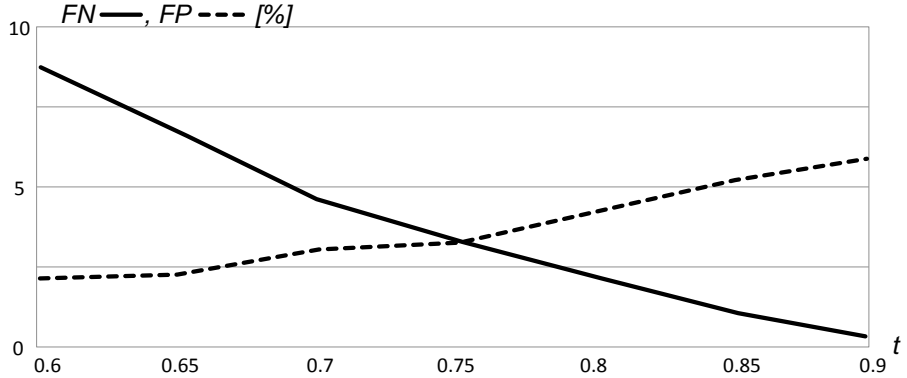

Fig. 1. The influence of globally adjusted values of t on FN and FP ratios.

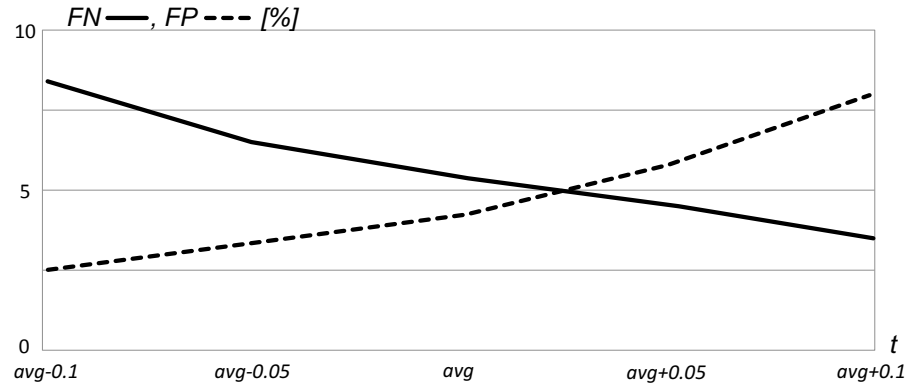

Fig. 2. The influence of individually adjusted values of t on FN and FP ratios.

Table 6. Number of files as a training set for different portions of the database used for obtaining average profiles.

user	100% of data	75% of data	50% of data	25% of data
user 1	26	20	13	7
user 2	5	4	3	1
user 3	4	3	2	1
user 4	17	13	9	4
user 5	6	5	3	2
user 6	4	3	2	1
user 7	19	14	10	5
user 8	5	4	3	1
user 9	16	12	8	4
user 10	21	16	11	5

stored work time.

The obtained results presented in Tab. 7 show not very strong influence of the amount of data on the output in the assumed periods of user activity (approximately 1 week to 1 month). Therefore, a general profile of a user activity can be based on periods containing only several days of work.

5. CONCLUSIONS

The experimental analysis of the method presented in this article confirmed initial assumptions of the authors. Based on observations of the results it can be concluded that the proposed method of profiling users of computer systems concerning multiple aspects of their activity appropriately distinguishes users when analyzing longer periods of time (several hours of activity). Therefore, the method is more suitable for use in intrusion detection in off-line mode, when the activity analysis is performed not in the real time mode but after collecting the activity data. Additionally, the proposed method can analyze whether there is the same user working continuously or whether the habits of the user are changing. Overall results are promising, however the future research should focus on two important aspects: collecting larger user activity database and profiles obtained from short periods of user's activity.

Table 7. The influence of individually adjusted values of t parameter on number of alarms and efficiency of the method.

% of database	$alarms$	$false\,alarms$	$false\,negative$	$true\,positive$	FP	FN	TP	E
100%	1106	36	37	1070	3,25%	3,34%	96,75%	93,40%
75%	1091	37	53	1054	3,39%	4,79%	96,61%	91,82%
50%	1100	41	48	1059	3,73%	4,34%	96,27%	91,94%
25%	1109	46	44	1063	4,15%	3,97%	95,85%	91,88%

6. ACKNOWLEDGMENTS

The work described in this article has been partially financed by the project "DoktoRIS - Scholarship program for innovative Silesia" co-financed by the European Union under the European Social Fund.

BIBLIOGRAPHY

[1] AGRAWAL A., User Profiling in GUI based Windows Systems for Intrusion Detection, Master's Projects, Paper 303, 2013.
[2] AHMED A., TRAORE I., A New Biometrics Technology based on Mouse Dynamics, University of Victoria, 2003.
[3] BLEHA S., SLIVINSKY C., HUSSEIN B., Computer-access Security Systems using Keystroke Dynamics, Pattern Analysis and Machine Intelligence, 1990.
[4] DE VEL O., ANDERSON A., CORNEY M., MOHAY G., Mining E-mail Content for Author Identification Forencics, ACM SIGMOD, 2001.
[5] GARG A., RAHALKAR R., UPADHYAYA S., KWIAT K., Profiling Users in GUI Based Systems for Masquerade Detection, The 7th IEEE Information Assurance Workshop, 2006.
[6] LUKATSKY Alex, Wykrywanie włamań i aktywna ochrona danych, wyd. Helion, Gliwice, 2005.
[7] MONROSE F., RUBIN A., Authentication via Keystroke Dynamics, New York University, 1997.
[8] PORWIK P., SOSNOWSKI M., WESOŁOWSKI T., WRÓBEL K., Computational Assessment of a Blood Vessels Compliance: A Procedure Based on Computed Tomography Coronary Angiography, LNAI, Springer, 2011, Vol. 6678/1, pp. 428-435.
[9] PUSARA M., BRODLEY C., User Re-Authentication via Mouse Movements, Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security, 2004.
[10] RICHARDSON Robert, 2010/2011 Computer Crime and Security Survey, New York, 2010.
[11] SCHONLAU M. et al., Computer intrusion: detecting masquerades, Statistical Science, 2001, Vol. 16, pp. 58-74.
[12] SCHONLAU M., Masquerading user data, http://www.schonlau.net.
[13] STARZYNSKA W., Statystyka praktyczna, wyd. PWN, Warszawa, 2005.
[14] STOLFO S., HERSHKOP S., WANG K., NIMESKERN O., A Behavior-based Approach To Securing Email Systems, Columbia University, 2003.
[15] YAMPOLSKIY Roman, Direct and Indirect Human Computer Interaction Based Biometrics, University at Buffalo, Buffalo, 2007.