Alin ZAMFIROIU, PhD
E-mail: zamfiroiu@ici.ro
The Bucharest University of Economic Studies
The National Institute for Research & Development in Informatics
Cristian CIUREA, PhD
E-mail: cristian.ciurea@ie.ase.ro
The Bucharest University of Economic Studies

A MODEL FOR USERS' PROFILE RECOGNITION BASED ON THEIR BEHAVIOR IN ONLINE APPLICATIONS

***Abstract**: In this article we propose a users' recognition model in online applications based on their behavior characteristics. For establishing the user profile recognition, there are defined the behavior characteristics in the case of online applications and these characteristics will be followed within some testing applications. Based on the identified characteristics, there are performed measurements for each user. The measurements are saved in a database and by calculation of an average, a profile is achieved. In the case of future user' authentications in the application it is compared the behavior measured in the current session with the profile saved in the database. The purpose of this paper is to determine patterns of user behavior analysis and to compare the current behavior with the profile from the database using the Euclidean distance calculation formula.*

***Keywords:** Euclidean distance, model, analysis, users, behavior, online applications.*

JEL Classification: C81, C88, O35

1. Introduction

Nowadays, the expansion of web and mobile applications and the increased use of information and communication technologies in all areas of activity have generated concerns about user behavior research in online software (Roth, 2014), (Yampolskiy, 2007). The developments in the mobile devices market have greatly increased the Internet traffic on mobile devices and the access degree to mobile web applications at the expense of the personal computers.

In (Bent et al, 2017) is described the way in which mobile devices can offer personalized services to users, based on personal data collected via embedded sensors along with monitoring the daily activities.

The users' privacy on social network applications is a hot topic in both industry and academic fields (Li et al, 2015), because nowadays collecting users personal information in web and mobile applications is a sensitive issue. (van Dam and van de Velden, 2015) presents a framework for data collection in order to obtain individual specific information and propose methodology to explore user profiles. In (Caruccio et al, 2015) is shown that predicting the users' goals can be extremely useful in web applications, such as e-commerce, online entertainment, and many others.

In (Guo et al, 2017) is presented a Markov model fordetecting the evolution properties of online user preference diversity, and the (Loyola et al, 2012) describes an ant colony optimization-based algorithm to predict web usage patterns.In (Kassak et al, 2016) is considered that the prediction of user's behavior in web applications can increase user experience, in order to provide better content. The same idea is described in (Hawalah and Fasli, 2015), where user's interests and preferences are stored in user profiles to provide tailor-made services.

Based on (John et al, 2012), the user profile tracking in cloud computing can be achieved by web usage mining techniques. Some different solutions to store large string data types and determine the optimal length of useful information is described in (Ivan et al, 2008), (Özen et al, 2017) and the economic impact of implementing cloud computing technology to provide personalized educational services is presented in (Boja et al, 2013).

According to (Shi et al., 2011), entering a new password in order to authenticate a user on a mobile device is difficult. Therefore, there is necessary to recognize the user behavior based on previous visits of the application. Thus, at every user authentication, the application registers its behavior, and once the model is established, at each future user authentication the application establishes a score calculated based on actual behavior and the behavior from previous authentications (Figure 1).

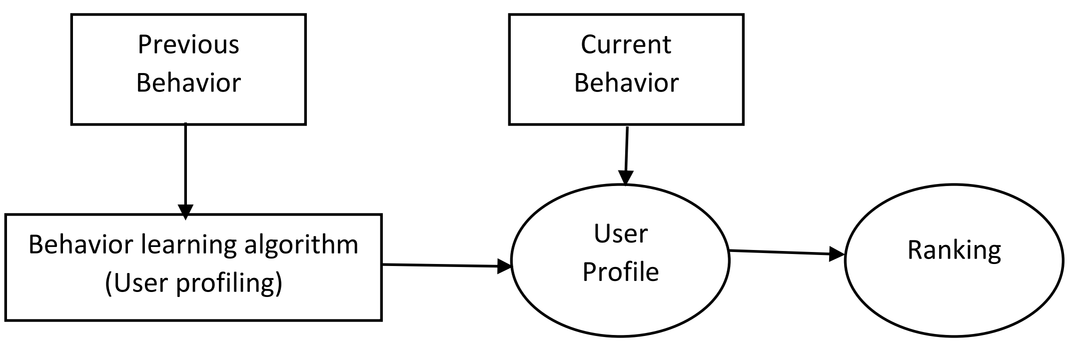

Figure 1. Structure of the algorithm proposed by Shi (Shi et. al., 2011)

Depending on the score resulting from running the algorithm, it determines whether or not the user is the rightful owner of the account. If the user is not the rightful owner, it is logged off and asked to log in again in the application.The existence of historical data about user behavior in the application helps to increase the security levels within the application, the protection of users' personal data and other sensitive data from the application.

2. Characteristics of the online applications users' behavior

In order to establish the users' profile, we proposed a predefined set of behavioral characteristics. For web applications, we take into account the following characteristics:

- ULCNr – **U**sed **L**eft**C**licks **N**umbe**r** within the application for push buttons or links at a time; this feature is influenced by how the application is deployed; so if the application requires pressing more buttons, then the value of this indicator will be very high;
- NLCNr –**N**o-functional **L**eft **C**licks **N**umbe**r** with the application; many people when reading content from the application they use to work with mouse clicks and give for selecting text to highlight or read on-screen cursor;
- RCNr –**R**ight **C**licks **N**umbe**r** made in the application using the contextual menu (if the application has contextual menus);
- CZ –**C**ursor **Z**one – the area where the user hover the mouse when don't use it; to determine this feature is preferred that the entire surface of the application to be divided into sectors defined and numbered so as to retain these features may not use the exact position in pixels, but only sectors that define the area; so depending on the application, the area can be divided into two or more sections;
- SM –**S**croll **M**ode of the content inside of the page with much text; according to user preferences; this can be scrolled by one of the following ways:
    - using the mouse scroll wheel; it is a specific method for users who mainly use the mouse and is more practical for them;
    - with the mouse using the *VerticalScroll* control on the right side of the page;
    - using keys *PageUp* and *PageDown*, is a specific method for users who use a lot the keyboard to interact with the computer and who want to run larger sections of the page;is a method used when the user reads very much, so to scroll after reading one or more paragraphs;
    - using *Up* and *Down* keys of the keyboard, is a method used to progress step by step, because it is in contrast with the scrolling

method by*PageDown* and *PageUp* keys, this method scrolls less, only few lines;

- SS –**S**croll **S**peed of the page, this feature is influenced by user's reading speed, so the user is directly characterized on the basis of personal characteristics; this feature is affected by the page content and information presented in the page;
- TS –**T**yping **S**peed **-** is a very important feature, however, is applicable only when the user has to enter text and should introduce enough text so that profiling is based on this characteristic in order to be as fair and accurate as possible (IT4fans, 2010);
- PT –**P**ressing **T**ime, in milliseconds; this feature is directly influenced by the typing speed characteristic; a user who has a very high typing speed will hold a key less time; also this analysis on key press time can be extended to track how long each key is pressed;
- DM –**D**elete **M**ode of the words written above (Genistra, 2009); there are several ways to erase and each user has a preferred one:
    - using the *DELETE* key, some users prefer to position the mouse cursor to the beginning of the word and use the *DELETE* key to delete the word;
    - by using the *Backspace* key, some users prefer to position the cursor at the end of the word and use the *Backspace* key to delete the word;
    - by selecting the word to be deleted using the contextual menu by pressing the right mouse click, then select the delete option from the contextual menu;
    - by selecting the word to be deleted and use the *DELETE*key or *Backspace* key to delete the entire word;
- TSM –**T**ext **S**electing **M**ode; text selection can be performed using the mouse or via *Shift + LEFT / RIGHT* keys; this feature depends very much of the user type: if a user prefers to use frequently the mouse or the keyboard;
- TCM –**T**ext **C**opying **M**ode **-**the user can choose one of the following options:
    - select the text and use the shortcuts *Ctrl + C*, *Ctrl + X* to copy it or to cut it;
    - select the text and use the contextual menu activated by right**-**clicking;

  this feature is also influenced by the type of users who use the keyboard or mouse to interact with online applications;
- CM –**C**apitalization **M**ode - methods which can write the capitalized letters are:

  - by pressing the *Caps Lock* before pressing the corresponding capitalized letter, and after pressing again the *Caps Lock* key to return to typing non-capitalized;
  - by pressing *SHIFT*key together with pressing the desired letter to be capitalized;
  - writing normally (non-capitalized) and then to transform in upper-case the desired characters after entering the full text.
- CKU –**C**ontrol **K**eys **U**sed (SHIFT, ALT, CTRL), on the left or on the right of the keyboard; each user utilizes these keys depending on its orientation, if he is right-handed or left-handed, and depending on the keys to be used with control keys; so if using a key on the left side, then it will be used control keys on the left; if using a key on the right side, then use the control key from the right;
- NM –**N**avigation **M**ode in the application; for example if the application allows using the BACK button, then use the BACK button of the browser or use *Alt + Backspace* keys combination to return to the previous page in the application;

In this way is created the crowd of 14 characteristics forthe user behavior:

$$MCh = \{\mathrm{ULCNr, NLCNr, RCNr, CZ, SM, SS, TS, PT, DM, TSM, TCM, CM, CKU, NM}\}$$

These characteristics are used to create the users' profile based on their behavior and their recognition in the online applications.

3. Building user profiles based on behavioral characteristics

To create the user profile it can be used all characteristics from MCh set or a partial set from this set.

Because the full set of characteristics can be considered very large, and these measurements may complicate the interaction with the online application, it can be considered a subset of characteristics. This subset consists of random characteristics, the characteristics chosen by the measurement capacity or by application area.

If from the multitude of characteristics MCh, we take into account the characteristics for user interaction with the mouse, then it is realized the subset:

$$SMCh^{mouse} = \{\mathrm{ULCNr, NLCNr, RCNr, CZ, SM, SS, DM, TSM, TCM, NM}\}$$

If from the multitude of characteristics MCh, we take into account the characteristics for user interaction with the keyboard, then will be realized the subset:

$$SMCh^{keyboard} = \{TS, \mathrm{PT, DM, TSM, TCM, CM, CKU, NM}\}$$

We can also consider the subset of characteristics that are represented by user actions:

$$SMCh^{action} = \{\text{ULCNr}, \text{RCNr}, \text{SM}, \text{SS}, \text{TS}, \text{PT}, \text{DM}, \text{TSM}, \text{TCM}, \text{CM}, \text{CKU}, \text{NM}\}$$

or the subset of characteristics which are the times when users do not interact with concrete actions of online application:

$$SMCh^{inactivity} = \{\text{NLCNr}, \text{CZ}\}$$

This subset includes a very low number of characteristics: only two characteristics.
The subset of characteristics for the user interaction with the online application is realized via the keyboard and is representative and the metrics for these features can easily be made and measured.
For all the characteristics of this set of characteristics will be performed measurements for all the user sessions within the online application.
To realize a profile, the measured values must be numeric values. In the case of characteristics for which measurements are not numerical values, it will be applied a conversion to carry out measurements.
For typing speed characteristic (TS) and the pressing time(PT) the measured values are numeric. For all other six characteristics there is considered the encoding from table 1.

Table 1.Coding modes for non-numerical values

Characteristic	Coding mode
DM	$= \begin{cases} 25 & DELETE \\ 50 & BACKSPACE \\ 75 & Contextual\ \ menu \\ 100 & Selection \end{cases}$
TSM	$= \begin{cases} 50 & MOUSE \\ 100 & SHIFT + LEFT / RIGHT \end{cases}$
TCM	$= \begin{cases} 50 & CTRL + C, CTRL + X \\ 100 & Contextual\ \ menu \end{cases}$

CM	$= \begin{cases} 50 & CapsLock \\ 75 & SHIFT \\ 100 & after\ \ typing \end{cases}$
CKU	$= \begin{cases} 50 & LEFT \\ 100 & RIGHT \end{cases}$
NM	$= \begin{cases} 50 & application\ \ BACK \\ 75 & browser\ \ BACK \\ 100 & ALT + BackSpace \end{cases}$

In Table 2, the measurements for *t* sessions of the user U are presented.

Table 2.Measurements made for a user session

Session	TS	PT	DM	TSM	TCM	CM	CKU	NM
S_1^u	TS_1^u	PT_1^u	DM_1^u	TSM_1^u	TCM_1^u	CM_1^u	CKU_1^u	NM_1^u
S_2^u	TS_2^u	PT_2^u	DM_2^u	TSM_2^u	TCM_2^u	CM_2^u	CKU_2^u	NM_2^u
…	…	…	…	…	…	…	…	…
S_i^u	TS_i^u	PT_i^u	DM_i^u	TSM_i^u	TCM_i^u	CM_i^u	CKU_i^u	NM_i^u
…	…	…	…	…	…	…	…	…
S_t^u	TS_t^u	PT_t^u	DM_t^u	TSM_t^u	TCM_t^u	CM_t^u	CKU_t^u	NM_t^u

These measurements are saved in a table, where are saved also all the measurements for all sessions performed from that application. For each user it will be developed a profile obtained by the arithmetic mean of all sessions performed by that user, calculated with the formula:

$$TS_m^u = \frac{\sum_{i=1}^{t} TS_i^u}{t}$$

The same formula applies to all features of the subset $SMCh^{keyboard}$, following the achievement of the average yield of the U user's profile. This profile is saved in a new table that holds the average values for each user and the number of sessions used for this average, as the format of table 3.

Table 3.Profiles of multiple users

User	Number of sessions	TS	PT	DM	TSM	TCM	CM	CKU	NM
U^1	T^1	$TS_m^{U^1}$	$PT_m^{U^1}$	$DM_m^{U^1}$	$TSM_m^{U^1}$	$TCM_m^{U^1}$	$CM_m^{U^1}$	$CKU_m^{U^1}$	$NM_m^{U^1}$
U^2	T^2	$TS_m^{U^2}$	$PT_m^{U^2}$	$DM_m^{U^2}$	$TSM_m^{U^2}$	$TCM_m^{U^2}$	$CM_m^{U^2}$	$CKU_m^{U^2}$	$NM_m^{U^2}$
…		…	…	…	…	…	…	…	…
U^j	T^j	$TS_m^{U^j}$	$PT_m^{U^j}$	$DM_m^{U^j}$	$TSM_m^{U^j}$	$TCM_m^{U^j}$	$CM_m^{U^j}$	$CKU_m^{U^j}$	$NM_m^{U^j}$
…		…	…	…	…	…	…	…	…
U^q	T^q	$TS_m^{U^q}$	$PT_m^{U^q}$	$DM_m^{U^q}$	$TSM_m^{U^q}$	$TCM_m^{U^q}$	$CM_m^{U^q}$	$CKU_m^{U^q}$	$NM_m^{U^q}$

These profiles are used to recognize the user based on its behavior.

If the user U^j will record a new session $S_i^{U^j}$, then measurements are performed for all of the characteristics from subset $SMCh^{keyboard}$ and it is desired that these measurements are able to contribute for changing the existing profile and the data from the database will be amended according with the formula:

$$new(TS_m^{U^j}) = \frac{old(TS_m^{U^j}) * T^j + TS_i^u}{T^j + 1}$$

This formula applies to all characteristics of the subset $SMCh^{keyboard}$, after that is changing the number of sessions in the table, based on the following formula:

$$new(T^j) = old(T^j) + 1$$

Similarly measurements and profiles will be realized for any particular subset or complete set of characteristics for user behavior.

In order that the profile to be representative for a user, it is recommended that the number of sessions based on which the profile is made to be as high.

4. Using Euclidean distance for recognizing users

It is considered a new session conducted by an unknown user U. Measurements for the characteristics from the subset $SMCh^{keyboard}$ are presented in table 4.

Table 4.Measurements of the unknown user U

S^U	TS^U	PT^U	DM^U	TSM^U	TCM^U	CM^U	CKU^U	NM^U

It will be considered a point in Euclidean space

$$U(TS^U, PT^U, DM^U, TSM^U, TCM^U CM^U, CKU^U, NM^U)$$

The profiles of the users from the database will be considered also as the points in Euclidean space (table 5):

Table 5.Euclidean points for the profiles from the database

User	Euclidean point
U^1	$U^1(\mathrm{TS}_m^{U^1}, \mathrm{PT}_m^{U^1}, \mathrm{DM}_m^{U^1}, \mathrm{TSM}_m^{U^1}, \mathrm{TCM}_m^{U^1}, \mathrm{CM}_m^{U^1}, \mathrm{CKU}_m^{U^1}, \mathrm{NM}_m^{U^1})$
U^2	$U^2(\mathrm{TS}_m^{U^2}, \mathrm{PT}_m^{U^2}, \mathrm{DM}_m^{U^2}, \mathrm{TSM}_m^{U^2}, \mathrm{TCM}_m^{U^2}, \mathrm{CM}_m^{U^2}, \mathrm{CKU}_m^{U^2}, \mathrm{NM}_m^{U^2})$
…	…..
U^j	$U^j(\mathrm{TS}_m^{U^j}, \mathrm{PT}_m^{U^j}, \mathrm{DM}_m^{U^j}, \mathrm{TSM}_m^{U^j}, \mathrm{TCM}_m^{U^j}, \mathrm{CM}_m^{U^j}, \mathrm{CKU}_m^{U^j}, \mathrm{NM}_m^{U^j})$
…	…..
U^q	$U^q(\mathrm{TS}_m^{U^q}, \mathrm{PT}_m^{U^q}, \mathrm{DM}_m^{U^q}, \mathrm{TSM}_m^{U^q}, \mathrm{TCM}_m^{U^q}, \mathrm{CM}_m^{U^q}, \mathrm{CKU}_m^{U^q}, \mathrm{NM}_m^{U^q})$

In order to identify the unknown user U is calculated the Euclidean distance between the point *U* and the point of the *q* corresponding profiles from the database, using the formula:

$$ED^j = d(U, U^j),\ j = \overline{1,q}$$

We have detailed for the characteristics from the subset $SMCh^{keyboard}$, where the formula is:

$$ED^j = \sqrt{\begin{aligned}&(TS^U - \mathrm{TS}_m^{U^j})^2 + (PT^U - \mathrm{PT}_m^{U^j})^2 + (DM^U - \mathrm{DM}_m^{U^j})^2 + \\ &+ (TSM^U - \mathrm{TSM}_m^{U^j})^2 + (TCM^U - \mathrm{TCM}_m^{U^j})^2 + (CM^U - \mathrm{CM}_m^{U^j})^2 + \\ &+ (CKU^U - \mathrm{CKU}_m^{U^j})^2 + (NM^U - \mathrm{NM}_m^{U^j})^2\end{aligned}} \qquad j = \overline{1,q}$$

With these values is created the set MED of the Euclidean distances between the point determined by measurements made for the user U and the profiles from the database.

$$MED = \{ED^1, ED^2, ..., ED^j, ..., ED^q\}$$

To identify the user *U* is enough to determine the minimum of the MED set. The user is identified by the user index for the minimum determined of the MED set.

5. Case Study: MOODLE web application users

In order to validate the recognition model based on user behavior within the application, a number of nine students who worked with Moodle application have been selected. For they were made measurements and realized nine profiles, presented in Table 6:

Table 6.Moodle user application profiles

UT	TS	PT	DM	TSM	TCM	CM	CKU	NM
1	94	63	25	100	50	50	50	100
2	11	99	25	100	50	50	50	50
3	70	57	50	50	50	75	100	75
4	48	24	75	100	100	100	50	50
5	33	22	100	100	50	100	100	75
6	47	28	100	50	50	75	100	75
7	63	54	75	50	100	50	100	75
8	23	7	100	50	50	50	50	50
9	60	76	25	50	50	50	50	50

The user U is considered with the measurement from the table 7:

Table 7.Measurements for the unknown user

U	24	6	25	100	50	75	100	50

For these measurements, it is calculated the Euclidean distances and is obtained the MED set shown in table 8:

Table 8. MED set with calculated Euclidean distances

User	Euclidean distances
U1	117.3627
U2	109.284
U3	92.0163
U4	95
U5	84.92349
U6	98.80789

U7	112.1383
U8	106.0754
U9	108.7244

By calculating the minimum from this set, the unknown user is identified with the user *U5*, which have the minimum: 84.92349.
If is considered the set of characteristics formed only with the first three characteristics, TS, PT and DM, the Euclidean distances calculated are shown in table 9:

Table 9.The set of Euclidean distances calculated for the first three characteristics

User	ED
U1	90.27181
U2	93.90421
U3	73.08899
U4	58.30952
U5	77.21399
U6	81.47392
U7	79.52987
U8	75.01333
U9	78.71467

In Figure 2 are represented in 3D space ten points awarded to ten users and Euclidean distances between the user U and the user profiles from database.
The graph from the Figure 2 was performed using Plotly 2.0 platform (Plotly, 2016).

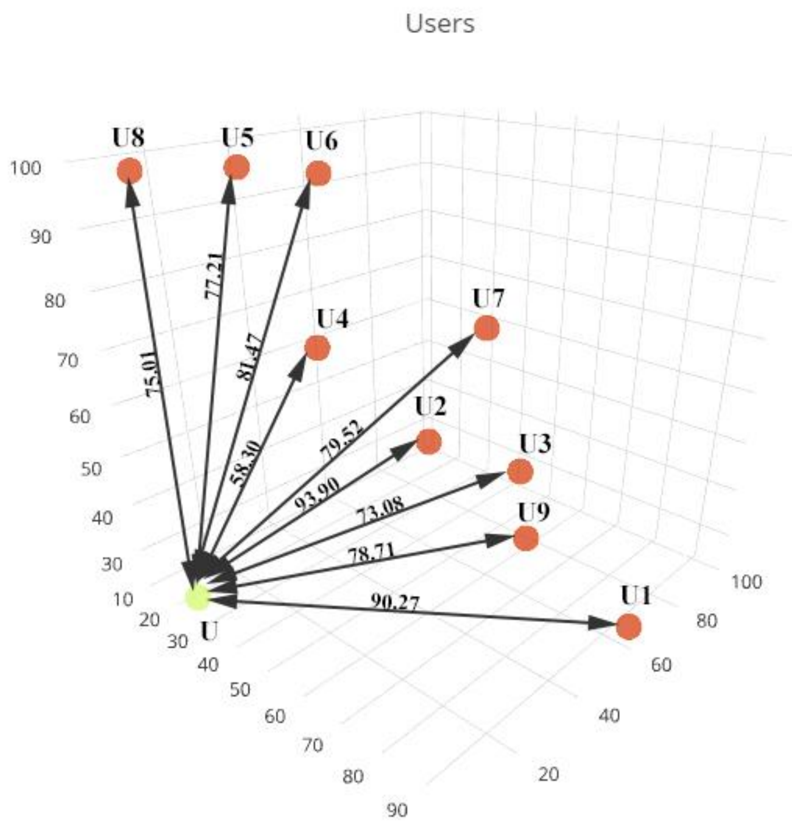

Figure 2. Euclidean distances between the user U and the user profiles from database

In this case, the unknown user is identified with the user U4, as the calculated Euclidean distance is minimal: 58.30952.

6. Conclusions

In this paper are identified behavioral characteristics for users in online applications. Based on these characteristics and measurements taken for these characteristics was built a model for determining the user profile and a user recognition model based on measurements made and profiling practices stored in the database.

In a future research we intend to optimize and realize a new model to be used in extreme situations, so that the user can be determined by other criteria, not only bythe minimum from the MED set.

Another existing solution includes weighting the characteristics, and thus achieving the highest differences between users.

ACKNOWLEDGMENT

This work was been carried out as part of the project: PN 16 09 01 02 – Cercetări privind autentificarea online în cadrul aplicațiilor software bazată pe comportamentul utilizatorilor (Researches on the online authentication on software applications based on users behavior).

REFERENCES

[1] **Bent, O., Dey, P., Weldemariam, K., Mohania, M. K. (2017),** ***Modeling User Behavior Data in Systems of Engagement****; Future Generation Computer Systems* 68, pp. 456-464;

[2] **Boja, C., Pocatilu, P., Toma, C. (2012),** ***The Economics of Cloud Computing on Educational Services****; Proceedings of 3rd World Conference on Learning, Teaching and Educational Leadership (WCLTA)*, Brussels, Belgium, October, 25-28;

[3] **Caruccio, L., Deufemia, V., Polese, G. (2015),** ***Understanding User Intent on the Web through Interaction Mining****; Journal of Visual Languages and Computing* 31, pp. 230-236;

[4] **van Dam, J.-W., van de Velden, M. (2015),** ***Online Profiling and Clustering of Facebook Users****; Decision Support Systems* 70, pp. 60-72;

[5] **Genistra (2009),** Available: http://www.genistra.com/blog/2009/02/12/whats-the-difference-between-delete-and-backspace/

[6] **Guo, Q., Ji, L., Liu, J.-G., Han, J. (2017),** ***Evolution Properties of Online User Preference Diversity****; Physica A* 468, pp. 698–713;

[7] **Hawalah, A., Fasli, M. (2015),** ***Dynamic User Profiles for Web Personalisation****; Expert Systems with Applications* 42, pp. 2547-2569;

[8] **IT4fans (2010),** Available: http://www.it4fans.ro/3360/cat-de-eficient-iti-folosesti-tastatura.html

[9] **Ivan, I., Boja, C., Doinea, M. (2008),** ***Determining the Optimal Length of Useful Information****; Economic Computation and Economic Cybernetics Studies and Research*, Vol. 42, Issue: 1-2, pp. 23-34.

[10] **John, J. M., Mini, V., Arun, E. (2012),** ***User Profile Tracking by Web Usage Mining in Cloud Computing****; International Conference on Modelling, Optimization and Computing (ICMOC 2012),* Procedia Engineering 38, pp. 3270-3277;

[11] **Kassak, O., Kompan, M., Bielikova, M. (2016),** ***Student Behavior in a Web-based Educational System: Exit Intent Prediction****; Engineering Applications of Artificial Intelligence* 51, pp. 136-149;
[12] **Li, K., Lin, Z., Wang, X. (2015),** ***An Empirical Analysis of Users' Privacy Disclosure Behaviors on Social Network Sites****; Information &Management*52, pp. 882-891;
[13] **Loyola, P., Roman, P. E., Velasquez, J. D. (2012),** ***Predicting Web User Behavior Using Learning-Based Ant Colony Optimization****; Engineering Applications of Artificial Intelligence* 25, pp. 889-897;
[14] **Özen, G.Z., Mehmet T., Rayimbek S.(2017).*****Hadoop framework implementation and performance analysis on a cloud***; *Turkish Journal of Electrical Engineering & Computer Sciences* 25.2, pp. 705-716.
[15] **Plotly 2.0 (2016),** Available: https://plot.ly/create/
[16] **Roth, J. (2014),** ***On Continuous User Authentication via Typing Behavior***; *IEEE Transactions on Image Processing*, July 28;
[17] **Shi, E., Niu, Y., Jakobsson, M., Chow, R. (2011),*****Implicit Authentication through Learning User Behavior***; *Information Security*. Springer Berlin Heidelberg, pp. 99-113;
[18] **Yampolskiy, R. V. (2008),** ***Action-based User Authentication***; *Int. J. Electronic Security and Digital Forensics*, Vol. 1, No. 3.