Background Data Mining for an Information Security Awareness Education Program - Reasons, Possibilities, Methodologies

Jeno Duchon

duchon.jeno@nav.gov.hu

Abstract: Nowadays we experience that the information is a very important thing. The information helps our days but sometimes it cause dependency. So the structure and the function of it is very complex. This is the main reason why we have to learn live with it. We have to develop out IT skills and competencies. But not enough just read and learn about the information. We have to change our attitude with the information if we want to be successful in the information society. Development of our information security awareness is the part of this process. But this part is a very important because for this not enough enroll to an IT security course. For this we have to change our learners attitude while we update their knowledge and forming their competencies. All of this require a highly complex training program. For plan this program we have to know a lot of input data and where are we able to find a lot of system usage date? The answer is: log files. So we if we use web mining tools and methods we are able to reach a lot of information about our user's behavior in our IT system.

Keywords: data mining, Moodle, users behaviour, security awareness, education

1. Background of an Information Security Awareness Educational Program

1.1. Why is it important to develop information security awareness?

We live in the information society where the information has value. However, the type of value is not only economic type but also it is social or power type.

In the information society the information itself and related activities, such as the communications, the data collections, the data processing, play a much more main role than before, and it significantly defines the relationship among the people, the

culture, the functioning of state and other organizations, and the definition of time and space. [1]

We could say that we live in digital prosperity. [2] Like in any other case, we need to look after for this welfare. It is not enough just enjoy the benefits. We have to ensure maintain this prosperity and this is a serious and responsible task.

In a welfare society, everyone has a responsibility to learn to handle and use various public utilities (for example water, gas, electricity). In this case the information is no exception from these utilities. In a welfare information society, the information is one of public utilities so we need to learn how to manage, process and filter the information. We can say that, we must develop every citizen's conscious using of information which includes the knowledge of information security and creating the appropriate attitudes and methods against the information vulnerability. [2]

1.2. IT security awareness development program is not equal teaching IT security knowledge

The developing of IT security awareness is not equal to a simple training. It is not enough to teach information security. The purpose of a security awareness program is that individuals are able to recognize IT security concerns and respond accordingly. The goal is here creating a behavior which allows the persons are able to pay attention IT security problem, and with wich they are able to perceive the dangers and are able to react consciously. [3]

Therefore, beyond the simple knowledge-transfer we are also need additional tools in order to produce the desired behavior change. We have to make a change of attitudes and that cannot be done through a simple educational program. A complex program is required.

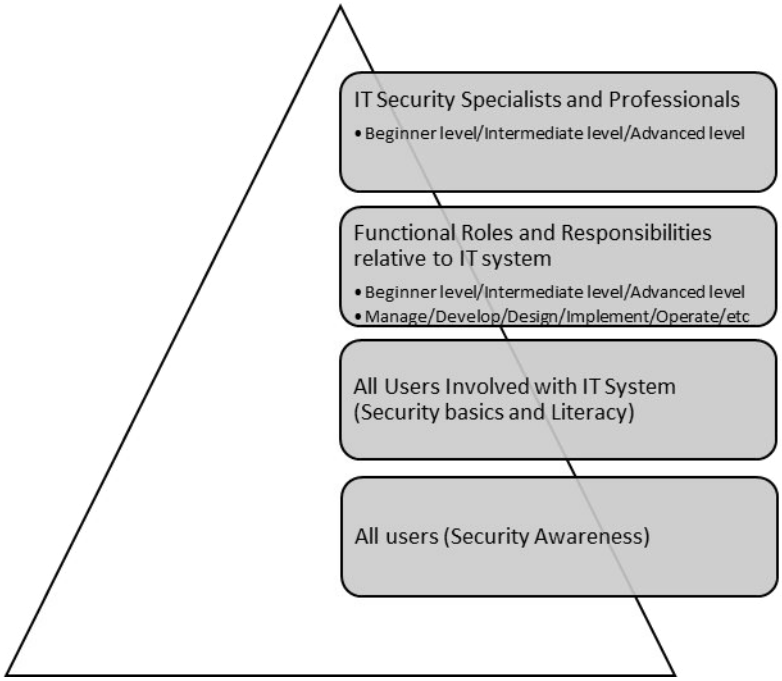

Figure 1.
Structure of a complex educational program
(adapted from [3])

1.3. The way to a good educational program

1.3.1. The factors of attitude

In order to understand how we can make a change of attitude, it is important to see what other factors affect the attitude of a person: [4]

Behavior intentions: Every person have a personal intention about own behavior in a certain situation under certain condition;

Real behavior: This means that actual behavior what the person are given in an actual situation. Expected this behavior will be different than the person's intention behavior;

Cognitions: Ideas, convictions and knowledge about how to need behave in a given situation;

Affective responses: that factors which affect the person's actual mood, social sensitivity in a given situation

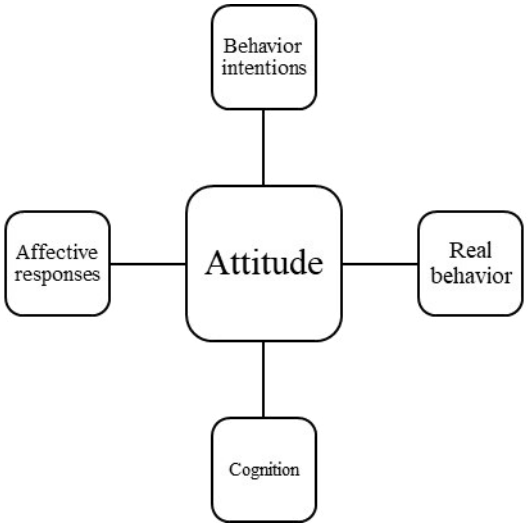

Figure 2.
An attitude system
(adapted from [4])

As we can see on the Figure 1 all factors are interrelated each other and it follows that, if one factor of them is changing, that cause changing the others.

1.3.2. The main areas of IT security what we have to focus on

During security awareness program development, always we have to highlight the areas which we would like to concentrate. This is influenced by environmental factors and needs (eg organizational needs). On the other hand, we need to focus on areas of IT security that affect the users in any case. Based on these the following areas can be identified where the security practices are recommended: [5]

Passwords

Patches and Updates

E-mail use and Antivirus software

Firewall, Spyware, and Popups

Backups

Physical Security

1.3.3. How to measure the awareness

In case of educational programs, we have opportunity to receive and adapt existing programs but lot of institutes develop their own unique program. [6] Regardless of whether, we develop a unique program or adapt an existing program suite, we should do measure the current level of security awareness of our affected target group. However, this level is not a simple indicator. We can not summarize our person's awareness with a grade or index. We have to find the weaknesses and the strengths of our employees' security knowledge and security problem management. For this we can apply a variety of research methods using the following five-step process: [6]

Knowledge about information security

Attitude towards information security

Normative belief towards information security

Intention for Information security

Information security behavior

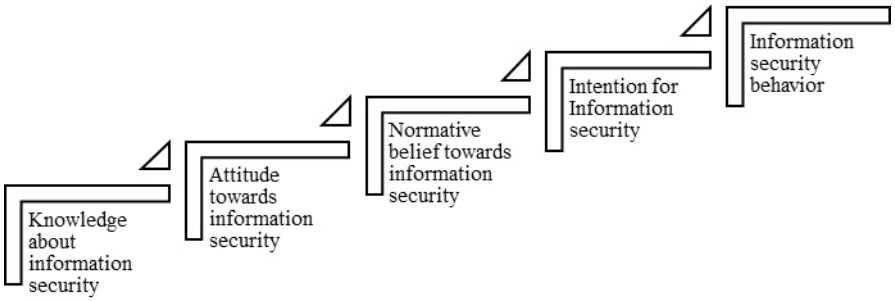

Figure 3.
Five step ladder model for measuring security awareness
(adapted from [6])

There is an important question about every measurement: what do we want to measure at all? Assume that, we should measure the global information security awareness level of the organization. To achieve this, we have to measure awareness level at each region and there are many problems that can be identified in each region. Based on this we can build up a tree structure of problems as we

can see in Kruger and Keaney article [7] where the authors' work relied on work of Belton and Stewart [8].

In this technique, the problem tree's root is the general level of IT security awareness and the next sublevel is the regional level. Such regions may be understood as organizational levels of the workplace or scope of job activities. Each regions are examined in three additional dimensions: knowledge (what you know), attitude (what you think) and behavior (what you do). It is important to note that, these three items are appeared when we define the awareness or these are connected to the five-step ladder model on Figure 2 because the figure's first steps are contain these. In the structure of tree every dimensions are subdivided into 6 focus areas and additional factors and subfactors are assigned to every focus areas:

Adhere to policies

Keep passwords secret

Email and internet

Mobile equipment

Report security incidents

Actions and consequences

We can see that the previously outlined safety areas are also displayed here, which also indicates to us which areas should be focused on the survey and the educational program.

1.3.4. How to change the attitude

Aforementioned, the assessments of the actual situation have a very important role and the expectations of organization with the safety. We could also summarize these areas that need to be focus in general. However, we have found that, security awareness can not be developing within a simple course, in this case a much more complex solution is needed, in which we can mark three main objectives: [4]

1. Directly changing the behavior (ignoring existing knowledge an attitudes) In this case we transmit same knowledge to each participant (eg in case of frontal training that describes the content of the information security rules of company);
2. Changing the attitudes of people through behavioral change. In this case, we have to build on previous experiences. The roleplay is very important in this level. It help to resolve incompatibility conflicts and establishes the self-opinion.
3. Attitude change through persuasion. In this case, we also build on past experience, but the persuasion gets an emphasis role.

In this point, it might be worth we are starting thinking about what tools we can use for this methodologies. Here we can use a wide range of information broadcast tools. We are able to combine the traditional techniques with ICT tools to make it more effective. For example, we have the following options: [6]:

Education with attendance form

Education with e-learning methods (eg in a LMS environment)

E-mail messages

Group discussion

Newsletter articles

Posters

Video games

Of course, every opportunities from aforementioned list have different efficiencies in the information broadcast changing the real behavior. It is important to underline, we don't have to choose one from the tools above, but it is advisable to apply them simultaneously and side by side.

2. Data mining as a measurement method

And now it is worthwhile to go back to the question: How do we explore the current level of awareness, and how we track the changing of this?

We can do a variety of ways to investigate information security awareness for a particular environment. On the one hand, we can do a lot of interviews. On the other hand, we can ask many people with different questionnaires. If we use the internet we are able to find a lot of templates what is good base for the investigation. (eg. [9] [10] [11])

However, these methods do not always give a true feedback about the real situation so it should be supported by other research methods, such as tracking the activity in an IT environment and analysing the stored user activity data.

The web mining basically covers three topics include:

Web content mining

Web structure mining

Web usage testing

The processing of datafiles what was created by an information system is classified into the topic of web usage testing. [12] This process is the data mining

which can help us to gain valuable information from a large amount of data. [13] Based on these we can say that the data mining can be considered to a kind of methodology of the web mining. [14]

2.1. Log files in Moodle system

We have seen before that the establishment of security awareness involves the behavior and attitude change and in this process, there are given an important role to the investigation of actual behavior characteristics. The log files of information systems contain the imprints of user behavior patterns, so we can conclusion from them by data mining.

If we would like to develop an educational program for upgrade employees of an organization, we have to design some questionnaire but we have to examine log files of that system what employees are using. For example, we are able to examine the log files of organization's educational system where every users learn about security. In this case, the real activities of students in the training environment may be differ from given student answers in the course exam. In addition to the learning outcomes we get information about the using frequency of each course's items, we check when these were opening, downloading. So, we can gain information about learning time and we can draw their behavior within a course, within an IT system. [15]

Moodle is a one of the most popular learning management system. There are more than 91000 Moodle registered sites in 233 country. [16] The National Tax and Customs of Hungary is used Moodle system for support their education and the examined security course was built within this system. Therefore, we should list what kind of data were extracted from the log files [17]:

The exact date of the activity performed in year/month/day and hour/minute.

The IP address of the computer that the activity was performed.

The full name of the person whose performing the activity

The activity itself.

Information metadata what was recorded for the activity.

In Summary, we can say that we are able to read more from the logs than the pure educational data (eg. quiz results, activity scores or grades, etc.). We get information about the participants of course when, how and how many times have reached the course items. [18]

2.2. Analyzing a log file

The accumulated data in the log files are very large numbers. To analyze the data, we need a suitable method, such as the following [17] [19]:

1. Pre-Processing: Collecting, processing, or producing data files.

   Select data: Which data we will use and how we will group these.

   Production of summary tables: Assigning data to the participants (eg How many times has a participant solved the test successful?; How many times has a participant reached an resources? ; etc.)

   Data discretization: We generate data ranges from the discrete data.

   Data Transformation: We have to convert the available data to ann other form what can be interpreted by the data mining algorithms.

2. Data mining: Selecting a data mining algorithm from the following:

   Statistical algorithm: Each individual element is grouped and given a quantitative value.

   Decision tree based algorithm: It represents a hierarchical structure of the conditional system. In this case to reach a leaf item from the root item all conditions must be met.

   Rule based algorithm: The decision has been by IF-THEN procedures. Each rule consists two parts: First part is the condition (IF) and the second part is a consequence (THEN). The condition part can be complex and based on logical algebra.

   Fuzzy logic based algorithm: There is a complex sets of rules which define a number of parameters for each variable.

   Artificial neural network based algorithm: This attempts to model biological neural networks.

3. Post processing: Interpretation of data produced by the above mentioned algorithms.

Of course, analysis is supported by a number of data mining tools which can be commercial (DBMiner, SPSS Clementine, DB2 Intelligent Miner) and publicly available (Weka, Keel). [19]

3. Data mining through one case – An example

3.1. Background of case

The National Tax and Customs Administration of Hungary has a lot of educational program within which the employees have to get to know an internal regulation. We have seen it before, these type of trainings are very important but these have got only role of knowledge broadcast and knowledge growing. They build up the theoretical knowledge to behavior which are expected by the organization. As we have seen before there is advisable to examining the current attitudes and behavior especially in case of a security policy teaching course.

In order to show how the web mining works in an attitude investigation we choose a Moodle course from the NTCA LMS system: Description of Guarding Safety Regulation. The form of this course is e-learning. Every participants have to get to know the contents of all regulation and then they have to solve a quiz and declare about learning of rules. Every employees must know the central regulation and the local regulation what extend the central rules with local specialties. So within the course every user was able to reach two resources (pdf file) which contain everything about the central and the local regulation and two quizzes.

3.2. Step 1: Pre-processing

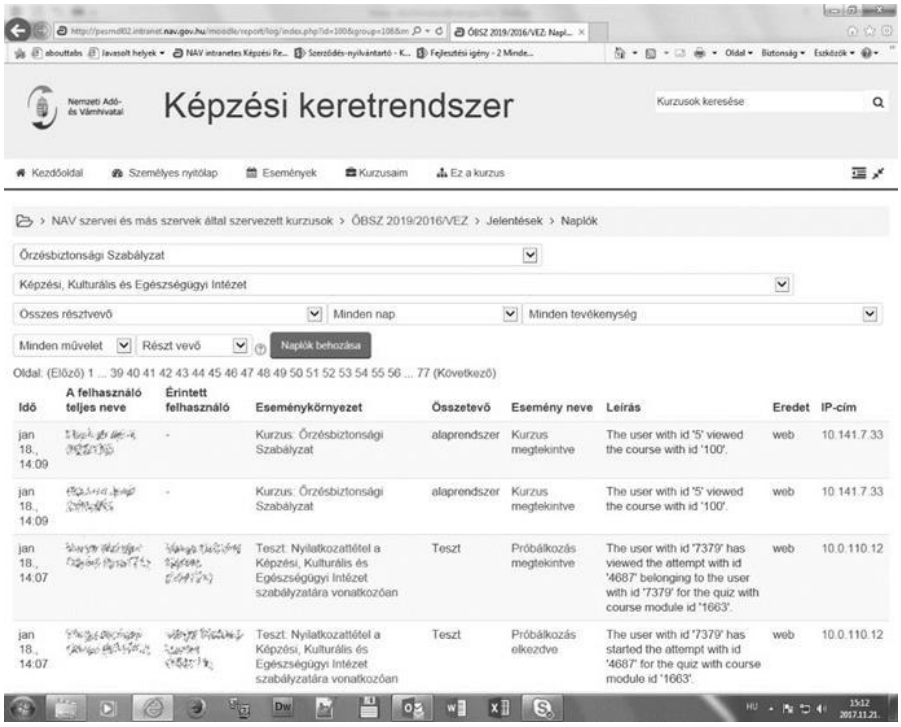

Figure 4.
Statistics on the use of the course elements

The figure 4 contains an example about this course log file (we have seen the ingredients of this file before).

First, we have to clean these data. We filter all items because we just need that rows what was written by Quiz and Resources module.

Then we have to do data transformation for the calculations. We transform every date and time record to minutes passed since 1 January 2001. It was necessary because with this we can examine the time spent between each activity. Next step, we restructured the data structure with a script. This script produced a user object. This object stored every user's attempt on the selected item of the course. With this we get a user base dataset where every rows contain is one user and their data.

```
<textarea rows="10" cols="100" id="input"></textarea>
<button id="start">start</button>
<textarea rows="10" cols="100" id="output"></textarea>
<textarea rows="10" cols="100" id="output2"></textarea>
</textarea>

$(document).ready(function(){
      $("#start").click(function(){
              var lines = $('#input').val().split('\n');
              var users = new Array();
              for(var i = 0;i < lines.length;i++){
              var lineArray = lines[i].split('\t');
              var user = new Array();
              objIndex = users.findIndex((user => user.TASZ ==
              lineArray[4]));
              user.TASZ = lineArray[4];
              user.times = [];
              user.times.push(lineArray[2]);
              if(objIndex < 0) {
                      users.push(user);
              }else{
              if(users[objIndex].times.indexOf(lineArray[2])<0){
                      users[objIndex].times.push(lineArray[2]);
              }
              }
              var outputString = new String();
              var output2String = new String();
              $.each(users, function(index, user) {
                      outputString += user.TASZ+"\t";
                      output2String += user.TASZ+"\t";
                      var beforeTime = 0;
                      $.each(user.times, function(index, time) {
                              outputString += time+"\t";
                              if (index == 0) {
                                      output2String += "0\t";
                              }else{
                                       output2String += time -
                                       user.times[index-1]+"\t";
                              }
                      });
                      outputString += "\n";
                      output2String += "\n";
              });
              $("#output").val(outputString);
              $("#output2").val(output2String);
      });
});
```

Figure 5.
Data transformation script

3.3. Step 2: Selecting a statistical algorithm for data mining

With the pure data, we are able to run some statistics test. First, we can verify that how many person open the resource content (it means they used the education content about the security regulation) and how many employees just filled the quiz without open the resources files. Second, we can calculate the descriptive statistics in relation to elapsed time between open the quizzes and open the resources. Of course, we can do an frequency examination on the data. With crosstabs, we can view the details on the elapsed time between opening contents and the participants' reached grade on the quizzes. We have the opportunity some complex examination on the preprocessed data. Before these we have to do normality test with for example Kolmogorov–Smirnov test. Of course we can search correlation between our variables too.

3.4. Step 3: Post processing

Within this study is not goal to deduction of conclusion about the behavior of employees of NTCA in terms of security awareness. Within this structure, we would like to give you some examples of data analysis. Therefore every examples below use one directorate data (NTCA Training Healthcare and Cultural Institute).

3.4.1. Some comperative statistics

To the course there was enrolled 282 employees from the Institute. Pre-processing the log files we get that 204 user solve the exam quiz about the central regulation and 202 user solve the exam quiz about the local regulation. From these participants there were 75 person who opened resource of central regulation from the course page 66 person did same with the resource of local regulation.

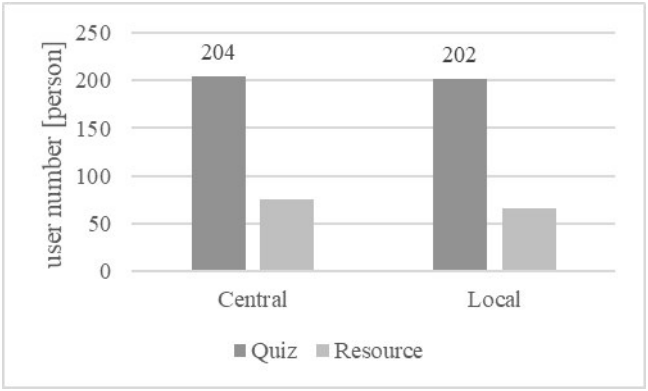

Figure 6.
Statistics on the use of the course elements (n=205)

We can see that just quarter of participants used the learning contents about the regulations but if we check the submitted cognition statement, we will discover that, every user signed it so they say that they have known the regulations.

In this point we have a conjecture: We have found a bad behavior pattern, because the employees could solve the quiz without got knowledge about the regulations. This conjecture is growing strong when we are drawing the difference on each users between the first open of the resource and the first quiz solving. (difference = first resource opening - first quiz solving)

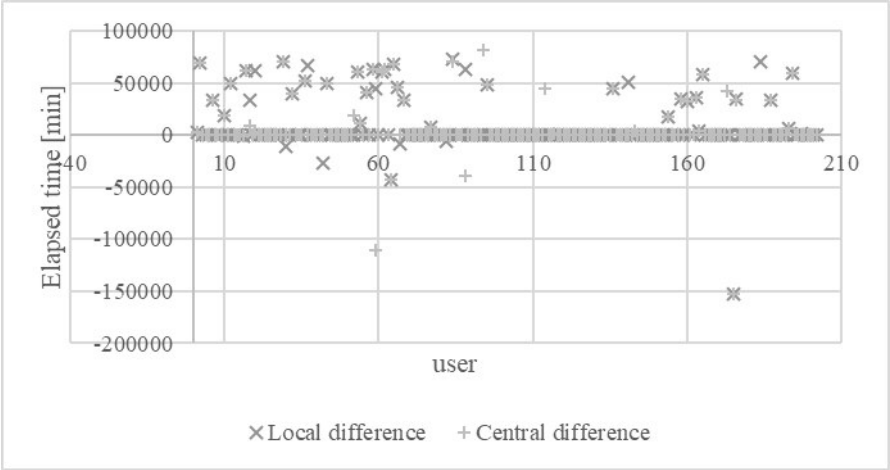

Figure 7.
Elapsed time between course elements (n=205)

We see that there are a lot of minus value and very rarely when the elapsed time is big between the resource opening and quiz solving. This is confirmed by the crosstab of differences.

Elapsed time between resource and quiz (central) * Elapsed time between resource and quiz (local) Crosstabulation

Count

		Elapsed time between resource and quiz (local)				
		<1 hour	1 hour > < 1 day	1 hour > < 1 week	> 1 week	Total
Elapsed time between resource and quiz (central)	<1 hour	154	0	1	5	160
	1 hour > < 1 day	0	4	0	0	4
	1 hour > < 1 week	2	0	4	2	8
	> 1 week	4	0	0	29	33
Total		160	4	5	36	205

Figure 8.
Elapsed time between resources and quizzes

3.4.2. Searching behavior pattern

In this point, expedient to examine our variables are normal distribution or not.

One-Sample Kolmogorov-Smirnov Test

		Elapsed time (central)	Elapsed time (local)	Quiz point (central)	Quiz point (local)
N		205	205	205	205
Normal Parameters[a,b]	Mean	6007,8146	7416,0634	95,5011	94,8544
	Std. Deviation	23599,30505	23152,01136	9,47702	14,00083
Most Extreme Differences	Absolute	,396	,420	,400	,439
	Positive	,396	,420	,317	,357
	Negative	-,375	-,340	-,400	-,439
Kolmogorov-Smirnov Z		5,664	6,019	5,721	6,278
Asymp. Sig. (2-tailed)		,000	,000	,000	,000

a. Test distribution is Normal.

b. Calculated from data.

Figure 9.

Examination normality

If we run a One-Sample Kolmogorov-Smirnov test, we experience that elapsed time between the resource opening and quiz solving neither in the central nor in the local in the variable not show normal distribution (K-S=5,664, p=0,001; K-S=6,019, p=0,001) Interestingly, we run this test on the result of user and we can see that these variable are neither normal distribution. (K-S=5,721, p=0,001; K-S=6,278, p=0,001)

Could be interesting if we compare the user's result with the elapsed time between resource and quiz.

Elapsed time between resource and quiz (central) * Quiz result (central) Crosstabulation

Count

		Quiz result (central)			
		Pass	Good	Excellent	Total
Elapsed time between resource and quiz (central)	<1 hour	3	40	117	160
	1 hour > < 1 day	1	0	3	4
	1 hour > < 1 week	0	5	3	8
	> 1 week	0	9	24	33
Total		4	54	147	205

Figure 10.

Compare ellapsed time with quiz result

We can see that the most users were successful in the quiz without learning the regulation because they opened the resource not long before the quiz. So they didn't have enough time to learn the connected knowledge.

Now it would be important examining there are any connection between our variables.

Correlations

		Elapsed time between resource and quiz (central)	Elapsed time between resource and quiz (local)	Quiz result (central)	Quiz result (local)
Elapsed time between resource and quiz (central)	Pearson Correlation	1	,824**	-,026	,065
	Sig. (2-tailed)		,000	,713	,352
	N	205	205	205	205
Elapsed time between resource and quiz (local)	Pearson Correlation	,824**	1	-,001	,082
	Sig. (2-tailed)	,000		,990	,241
	N	205	205	205	205
Quiz result (central)	Pearson Correlation	-,026	-,001	1	,020
	Sig. (2-tailed)	,713	,990		,773
	N	205	205	205	205
Quiz result (local)	Pearson Correlation	,065	,082	,020	1
	Sig. (2-tailed)	,352	,241	,773	
	N	205	205	205	205

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 11.

Correlations between variables of elapsed time and variables of quiz result

We examined the elapsed time between open the resource and solve the quiz events, and we experienced that the correlation between the central and the local variables are strong. (r=0,824, p= 0,001) Between the result of quizzes in central exam and local exam are not any connection. This means that the users follow same learning strategy in the central and local course elements but it did not lead to the same result.

3.5. Conclusion

Based on the examinations presented as an example we can make the following observations:

The users followed similar strategies for complete the course in the central and the local regulations.

The users' goal was to successfully complete the course. The result achieved was not important.

Many users didn't want to know anything about the regulation, just "put a pipe next to the course".

We have seen that in the factor of attitude, if we change any factor the attitude will be changing to. We have examined the actual behavior of our users (for more serious conclusions, a more detailed examination is required) but now we can

suspect that if we modify this course that will be inducing a behavior changing in the user.

For example, if we modify the learning resources to an interactive e-learning application or we will using some gamification elements within the course (eg. badges, coin gathering or storyline, etc.), maybe the users goal will be change and this case they don't want to just finished the course. If the users are able to sink into the learning content, we think their behavior will change unnoticed.

4. Summary

In this study, our goal was that we investigate how help for us the data mining if we would like to build a successful educational program for upgrading employees' IT security awareness in an organization. First, we collected why it is important and then we investigated how we can measure the actual states. The next section we briefly described data mining as a measurement tool and methodology and the last we show some examples with real data.

References

[1] R. Pintér, "Úton az információs társadalom megismerése felé," in Az információs társadalom - Az elmélettől a politikai gyakorlatig, Budapest, 2007.

[2] Z. Nyikes, "A biztonságtudatosság a digitális kompetencia tükrében," XXI. Fiatal Műszakiak Tudományos Ülésszaka, 2016.

[3] M. Wilson and J. Hash, "Building an information technology security awareness and training program.," NIST Special publication, vol. 800, p. 50, 2003.

[4] M. E. Thomson and R. von Solms, "Information security awareness: educating your users effectively," Information management \& computer security, vol. 6, no. 4, pp. 167-173, 1998.

[5] C. P. Garrison and O. G. Posey, "Computer Security Awareness of Accounting Students," 2006.

[6] B. Khan, K. S. Alghathbar, S. I. Nabi and M. K. Khan, "Effectiveness of information security awareness methods based on psychological theories," African Journal of Business Management, vol. 5, no. 26, pp. 10862-10868, 2011.

[7] H. A. Kruger and W. D. Kearney, "A prototype for assessing information security awareness," Computers & Security, vol. 25, pp. 289-296, 2006.

[8] V. Belton and T. Stewart, Multiple Criteria Decision Analysis - An Integrated Approach, Springer Science & Business Media, 2002.

[9] Academic Frontier Project for Private Universities, “Survey on Internet Security Awareness,” 03 2009. [Online]. Available: http://www.kansai-u.ac.jp/riss/en/shareduse/data/17_E_questionnaire.pdf. [Accessed 24 06 2017].

[10] SANS, “Security Awareness Planning Kit,” 2016. [Online]. Available: https://securingthehuman.sans.org/resources/planning. [Accessed 25 06 2017].

[11] Security Mentor, “Security Awareness Survey,” 2015. [Online]. Available: https://www.securitymentor.com/resources/awareness-survey. [Accessed 25 06 2017].

[12] C. D. K. B. A. B. &. J. S. P. Desikan, “Web Mining For Self-directed E-learning,” in Data Mining in E-Learning, S. V. C. Romero, Ed., Southampton, Boston, WITPress, 2006, pp. 21-37.

[13] W. Klösgen and J. M. Zytkow, Handbook of data mining and knowledge discovery, New York, NY, USA: Oxford University Press, Inc, 2002.

[14] B. M. a. J. S. R. Cooley, “Web mining: information and pattern discovery on the world wide web,” in 9th IEEE International Conference on Tools with Artificial Intelligence, Vols. Newport Beach, California, Newport Beach, California, 1997, pp. 558-567.

[15] P. Tóth, “Online learning behavior and web usage mining,” Transactions on Advances in Engineering Education, vol. Vol. 10, pp. 71-81, 2013.

[16] Moodle.org, “Moodle Statistics,” 2017. [Online]. Available: https://moodle.net/stats/. [Accessed 14 11 2017].

[17] C. Romero, S. Ventura and E. García, “Data mining in course management systems: Moodle case study and tutorial,” Computers & Education, vol. 51, p. 368–384, 2008.

[18] A. El-Halees, “Mining students data to analyze e-Learning behavior: A Case Study,” 2009.

[19] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero and S. Ventura, “Web usage mining for predicting final marks of students that use Moodle courses,” Computer Applications in Engineering Education, vol. 21, no. 1, pp. 135-146, 2013.