Assessment Methodology for the Categorization of ICT System Users Security Awareness

K. Solic, B. Tovjanin and V. Ilakovac
J.J. Strossmayer University, School of medicine/, Osijek, Croatia
kresimir@mefos.hr, btovjanin@mefos.hr, vilakov@mefos.hr

Abstract

The ICT system's users can significantly affect overall security level of the system, but problem is that most security solutions do not take into consideration user as possible critical security component of the system.

In this work assessment methodology is proposed to evaluate users' awareness regarding security issues. For purpose of collecting data on ICT system user's awareness special questionnaire was developed based on previously defined ontology domain regarding e-mail users' behavior.

The cluster analysis method was applied in order to group users into categories regarding level of their awareness about security issues. Cluster analysis gave six clusters of users on which Chi-square analysis was applied in order to detect potential relationship between level of awareness and gender, age, professional qualification and number of e-mail addresses used. The variables used to predict group membership were identified by applying discriminant analysis.

The evaluation and categorization of users' awareness should help in developing new concepts of security solutions with taking into consideration user as component of the ICT system.

I. Introduction

As ICT system evolves so the new security issues arise and new security technical solutions are going to be developed. This cycle seems to be hard to disassemble and many technical solutions are applicable only for definite period of time. Problem can be that these solutions do not focus on the user as critical security component of the ICT system.

User can significantly affect security level of the system [1] and should be taken in consideration and maybe even as starting point for proposition on a new possible concept of (global) security. Some solutions are proposed in some resent studies with subject of security in information and communication systems, with special emphasis on Internet security. One solution proposes concept *"trust on Internet"* [2], while the other proposes *"neighborhood watch"* organized as social network [3]. Both solutions target users and their awareness of security issues.

Also users are rarely taken into consideration in security guidelines. The *German national IT security guideline* takes user into consideration but it points out only user's usage of e-mail communication system as potential security critical issue and recommends that every user has to have basic understanding/awareness of system's security issues [4].

In this work assessment procedure is proposed to evaluate users' awareness regarding security issues. The evaluation and categorization of users' awareness should help with development of new concepts of security solutions with taking into consideration user as component of the ICT system.

II. Assesment Methodology

In order to begin evaluation over some object first step should be to define and describe that object. Proposition is to use ontology, because in recent years ontology structure is mostly used do define domain(s) in area of information technologies [5]. *OWL ontology* has been chosen to formally define knowledge about some domain of interest by defining concepts and relations between them [6]. If ontologies that define particular domain already exist, one can choose among them.

Questionnaire was developed with questions mapping particular classes in ontology structure in order to gather data for evaluation.

Cluster analysis procedure was chosen in order to categorize ICT system's user's awareness regarding security issues. This procedure is frequently used for example in economic field related to marketing for categorization of customers [7]. Cluster analysis seeks to identify homogeneous groups of cases or individuals in a population, where the optimal number of groups, the properties of segments and group membership are unknown in advance. This means that a cluster analysis is used as exploratory technique.

Drawing *dendogram*, also known as *tree diagram*, is a common way to visualize the cluster analysis's progress by displaying the distance level at which there was a combination of objects and clusters. It is possible to define number of clusters by tracking differences between distance levels in previous and next step of algorithm [8].

Discriminant analysis was applied on groups and grouping variables in order to evaluate quality of clustering and to identify variables that have significant influence on group membership.

For detailed analysis of each identified group external variables can be defined. Those several additional variables can identify gender, age, working place, professional qualification, technical background, etc.

III. Case Analysis

For purpose of collecting data on ICT system user's awareness special questionnaire was developed that was based on previously defined ontology domain regarding e-mail users' behavior. There are several ontologies defined regarding domain of ICT system or its parts, but even most detailed ontology fails to cover ICT system's users and their behavior [9]. Hence in this work ontology from previous research that formally defines behavior of e-mail system's user was used.

The organization of ontology's classes is presented in Fig.1. Instances in ontology are values representing grades from poor to excellent; some subclasses that represent questions have all five possible entities and some only two or three depending on associated answers. The questionnaire has few basic questions about gender, age, working place, number of e-mail addresses used; and questions regarding e-mail user's behavior which maps each case in the ontology and comprise following topics:

- quality of password
- criticism towards collocutor
- way of usage of e-mail address
- security issues regarding e-mail system
- way of access to e-mail system

Free, open source software tool, ontology editor Protégé 4.1 (Stanford, California, USA) was used for definition and description of the ontology.

Data variables collected by questionnaire were divided on external variables and dependent variables. External variables were gender, age, professional qualification and number of e-mail addresses in usage; and were used to analyze groups after categorization. Dependent variables were collected from answers regarding ICT system users' awareness of security issues and were used for categorization. Answers on those questions were ordinal data in scale from one to five (poor, indifferent, average, good and excellent), but some questions had only two or three possible answers.

From 18 questions 12 were discarded for the cluster analysis because of several reasons: questions with binary data are meaningless for cluster analysis; questions that correlate had to be reduced before cluster analysis and also relatively small size of dataset (n=306) was additional reason for discarding questions [8].

In total six questions with possible answers were selected (Fig.1, Tab.1):

1. Do you use free e-mail services (like gmail, yahoo and hotmail) and if yes in what manner?

   a) for professional communication

   b) for private purposes

   c) only for periodical usage

If answer is "No" grade equals *indifferent* meaning very secure, and for the rest of answers grades are as follows: "a" equals *average*, "b" equals *good* and "c" equals *excellent*.

2. Do you use web browser for accessing e-mail system and if answer is **sometimes** or **yes**, do you take care of the place where from you make connection?

   a) only from home or office PC

   b) sometimes from public places as well

Five possible answers give all five possible grades. Answer "No" equals *excellent*, because it is more secure to use e-mail client software tool. Combination of answers "Yes + a" equals *good*, combination of answers "sometimes + a" equals *average* and combination

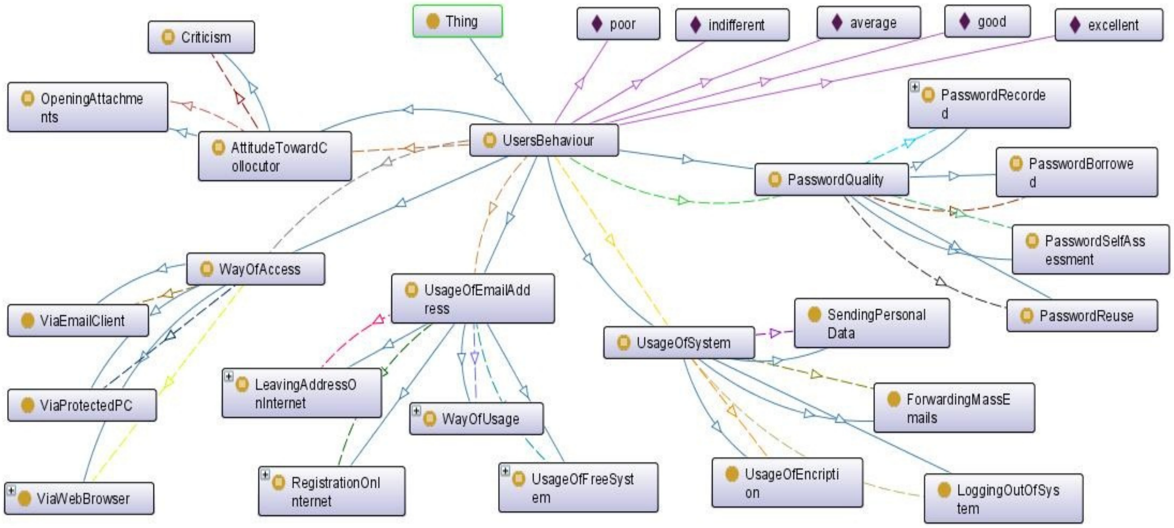

Figure 1. Formaly defined domain of knowledge in ontology editor Protégé 4.1

TABLE I. AVERAGE ANSWER GRADES PER GROUP

Selected questions with covered subjects	**Group 1 /n=45**	**Group 2 /n=42**	**Group 3 /n=63**	**Group 4 /n=46**	**Group 5 /n=63**	**Group 6 /n=47**	***p*****
Q2 (usage of free e-mail services) /mean±SD	3.11±0.78	3.07±0.64	3.21±0.77	3.20±0.72	3.30±0.85	3.09±0.62	0.578
Q5 (way of access) /mean±SD	1.09±0.29*	3.26±0.59	3.24±0.53	2.50±1.28	2.49±1.15	2.60±1.14	<0.001
Q10 (attachments from unknown senders) /mean±SD	5.00±0.00	4.95±0.22	4.98±0.13	5.00±0.00	1.86±0.35*	4.96±0.20	<0.001
Q12 (sending private/sensitive data) /mean±SD	4.27±0.45	4.33±0.48	4.48±0.50	4.50±0.51	3.27±1.58	1.02±0.15*	<0.001
Q14 (logging off the system) /mean±SD	5.00±0.00	5.00±0.00	5.00±0.00	2.43±0.78*	4.41±1.01	4.49±1.04	<0.001
Q15 (quality of password) /mean±SD	3.80±1.08	5.00±0.00*	2.65±0.77	3.30±1.46	3.35±1.32	3.40±1.46	<0.001

*significant influence from particular question on particular group

**One Way ANOVA test; *p* is significant at level <0.05

“sometimes + b” equals *indifferent*. Combination of questions “Yes + b” equals grade *poor*, because it is the most unsecure combination.

3. Are you opening e-mail attachments sent to you from unknown persons?

With possible answers “No”, “Sometimes” and “Yes” the examinee can get grades equal to *excellent, indifferent* or *poor*.

4. Are you sending your personal or sensitive data over e-mail?

Possible answers were: “Yes”, “On an exceptional basis”, “No” and “Don’t know” and matching grades are equal to *poor, average, excellent,* and *indifferent*.

5. Do you log of from the e-mail system after finishing your work?

Possible answers for this question were “Yes”, “Mostly yes”, “No” and “Don’t know” with matching grades *excellent, good, poor* and *indifferent*.

6. Please self-assess your password quality?
   a) excellent (combination of small letters, capital letters and numbers)
   b) average
   c) poor
   d) don’t know

Grades for questions “a”, “b” and “c” are the same as given answers and grade for answer “d” is *indifferent*.

The examinee would get the lowest possible grade if he/she gave multiple answers.

Hierarchical method was used as the most common approach to cluster analysis [8]. Also Euclidean distance measure of (dis)similarity was chosen because of ordinally scaled data and Ward’s method was chosen algorithm because there are no outliers and the aim was to have

TABLE II. DISTRIBUTION OF GENDER, AGE, PROFESSIONAL QUALIFICATION AND NUMBER OF ADDRESSES

	Group 1 /n(%)	**Group 2 /n(%)**	**Group 3 /n(%)**	**Group 4 /n(%)**	**Group 5 /n(%)**	**Group 6 /n(%)**	***p****
gender							
male	19 (42)	19 (45)	18 (29)	14 (30)	24 (38)	21 (45)	**0.341**
female	26 (58)	23 (55)	45 (71)	32 (70)	39 (62)	26 (55)	
age							
<=25	27 (60)	18 (43)	29 (46)	28 (61)	39 (62)	22 (47)	**0.099**
30<>45	16 (36)	20 (48)	24 (38)	12 (26)	21 (33)	23 (49)	
>=45	2 (4)	4 (9)	10 (16)	6 (13)	3 (5)	2 (4)	
professional qualification							
secondary school	24 (53)	21 (50)	38 (60)	27 (59)	41 (65)	26 (55)	**0.687**
university education	21 (47)	21 (50)	25 (40)	19 (41)	22 (35)	21 (45)	
number of e-mail addresses in usage							
one	12 (27)	11 (26)	23 (36)	16 (35)	28 (44)	16 (34)	**0.099**
two	20 (44)	20 (48)	31 (49)	20 (43)	30 (48)	19 (41)	
three	7 (16)	10 (24)	6 (10)	9 (20)	4 (6)	10 (21)	
> 3	6 (13)	1 (2)	3 (5)	1 (2)	1 (2)	2 (4)	
total	**45 (100)**	**42 (100)**	**63 (100)**	**46 (100)**	**63 (100)**	**47 (100)**	

*Chi-square Test; *p* is significant at level <0.05

similarly sized clusters [7]. Standardization of variables is needed when values are in different scales or variance differs significantly, which is not the case in this example [8].

After the categorization and evaluation of clusters each group of users was analyzed regarding awareness of security issues in combination with additional variables (gender, age, professional qualification and number of e-mail addresses used).

Clustering and statistical calculations were done with software tool Statistica 10.0 (StatSoft, Tulsa, OK, USA).

IV. Results

The number of clusters is defined when examining dendogram (Fig.2) which graphically presents result of the cluster analysis. The steps in which Ward's algorithm can be stopped should be detected from resulted dendogram, depending on number of clusters and distance between then. In this analysis algorithm was stopped between 26% and 41% of the whole clustering procedure because it presents quite big distance between groups and results in similarly sized clusters. Bigger distance in dendogram presents difference between groups and it is presented graphically as higher jump. This procedure results in six clusters representing six groups of users.

Classification of discriminant analysis showed that a 98.7% of originaly grouped cases were correctly classified and canonical discriminant functions gave variables that significantly influenced on group membership (Tab.1). Overlapping between groups was only 1.3%.

None of the questions had significantly influenced on the Group 3 and also question Q2 has equally influenced on all six groups (p=0.578).

Only Group 2 has value "excellent" for variable that has significant influence on clustering analysis and can be called *"excellent password quality group"*. Groups 1, 4, 5 and 6 are "poor" or "indifferent" in related variable with significant influence while group 3 is "average" in a way regarding all six variables.

Users that belong to the *"excellent password quality group"* have their password graded as excellent which is significantly different from the grades of passwords of users of the other five groups.

In the first group that can be called *"less secure access group"* users prefer less secure way of access to the e-mail system, which significantly differs comparing them to the other users.

While the users of the *"group of average awareness"* are average regarding answers to the all six questions, users that belong to the forth group, "*forgettable group*" do not log off the system after finishing working with it.

Fifth group can be called *"naive group"* because these users are not critical to unknown collocutors and sixth group can be called similarly, for example *"security critical group"* because users from that group are sending personal and sensitive data by e-mail as plain text.

Distributions of external variables were not significantly different between categorized groups. However sample size is relatively small and *p* value is close to level of significance for external variables: *age* and *number of e-mail addresses used* (Tab.2).

V. Conlusion

According to results it seems that cluster analysis, in combination with ontology and associated questionnaire, can be applicable method for categorization of the ICT

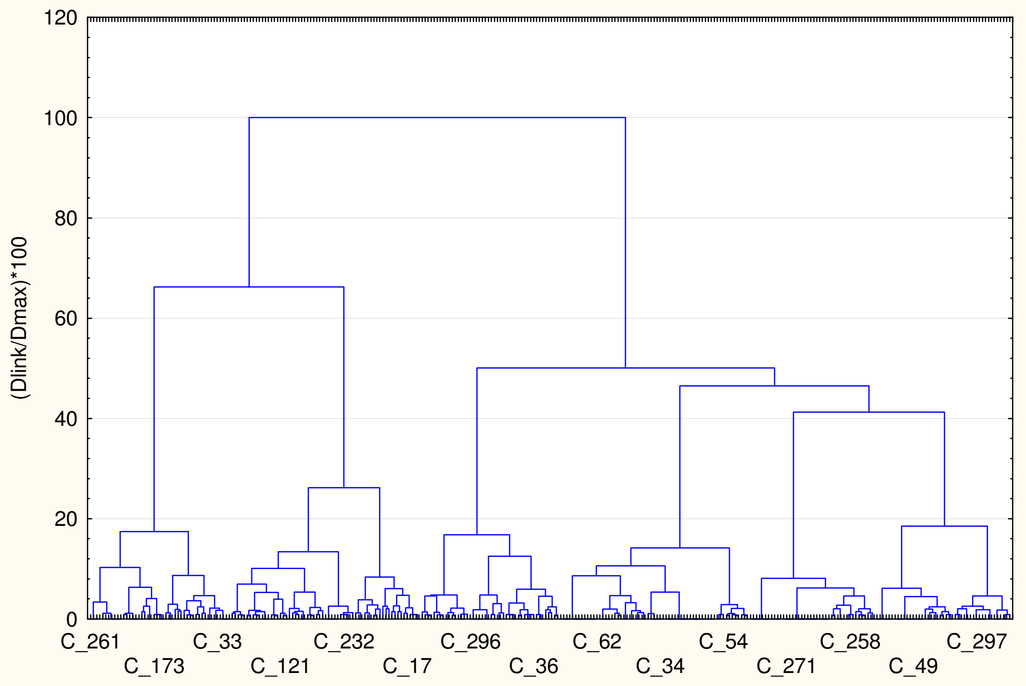

Figure 2. Tree diagram graphically presents result of the cluster analysis

system's users' awareness regarding security issues.

Formalizing domain of interest by using OWL ontology concept allows high level of flexibility (upgrade and update) and reuse of domain with possible coupling with other ontologies.

For upgraded or newly defined ontology new questionnaire can be easily made by mapping questions to the subclasses and answers to the possible instances.

Clustering analysis is flexible exploratory method and allows repetition of categorization on bigger sample size of either general or specific types of users (e.g. users of particular organization's ICT system). The evaluation and categorization of users' awareness should help in developing new concepts of security solutions.

Drawback of this study is in cluster analysis. As it is exploratory method researchers do not know what results to expect and if the resulting groups will differentiate enough for further statistical analysis.

Specific solutions can target different groups of users regarding results of the cluster analysis. For example specific solution can be developed for specific company after categorization and analysis of its employees and depending on company's security requirements which differs from bank, hospital, ICT company or some hotel.

From example analysis in this work *"naive group"* and *"security critical group"* need urgent attention and development of some solution in order to influence on security awareness level of these users.

Possible future work may be in applying this assessment methodology on students from different faculties. Analysis may identify what are differences in security awareness between students of different study fields and possible differences between students at the start and at the end of their studying period.

REFERENCES

[1] K. Solic, D. Sebo, F. Jovic, V. Ilakovac, "Possible Decrease of Spam in the Email Communication", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, L. Ordanic, 2011, pp. 170-173.

[2] S. Gros, M. Golub, V. Glavinic, "Using Trust on the Internet", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, V. Dragsic, Eds. 2008, pp. 118–123.

[3] S.J. Lukasik, "Protecting Users of the Cyber Commons", Communications of the ACM, vol. 54, 2011, pp. 54-61.

[4] IT Security Guidelines. Federal Office for Information Security, Bonn, Germany. 2007. URL: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/guidelines/guidelines_pdf.pdf (14/01/2012).

[5] A. Klaic, N. Hadjina, "Methods and Tools for the Development of Information Security Policy – A Comparative Literature Review", IEEE MIPRO / D. Cisic, Z. Hutinski, M. Baranovic, M. Mauher, L. Ordanic, 2011, pp. 190-195.

[6] M. Horridge, "A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools", The University of Manchester. 2011. URL: http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf (14/01/2012).

[7] E. Mooi, M. Sarstedt, "A Concise Guide to Market Research" (Cluster Analysis, Chapter 9), Springer, 2011, pp. 237-284.

[8] J.D. Jobson, "Applied Multivariate Data Analysis", Springer, 1992, pp. 518-616.

[9] S. Fenz, A. Ekelhart, "Formalizing Information Security Knowledge", ASIACCS, Y. Mu, P. Ogunbona, 2009, pp. 183-194.