Heterogeneous, Semi-Structured Data Analysis in Information Security

Maria A. Poltavtseva[1], Petr D. Zegzhda[2]
Information security of computer systems dep.
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russia
[1]Maria.poltavtseva@ibks.icc.spbstu.ru,
[2]zeg@ibks.ftk.spbstu.ru

Elizaveta A. Zaitzeva[3]
OOO "Neobit"
St. Petersburg, Russia
[3]elizavetazaytceva@mail.ru

***Abstract*— Solving information security problems one need to analyze and evaluate the similarity of heterogeneous semi-structured objects sets. The authors propose to use for this an approach on the basis of case analysis. The data is represented as a set of objects (use cases), each of which is described by a variable set of properties. The paper proposes measures to evaluate the similarity of individual objects and precedents, presents the results of the similarity evaluation method experimental testing. The authors present the architecture of the analysis system and the method of its learning.**

Keywords— component; analysis of heterogeneous semi-structured objects; case analysis; information security

I. INTRODUCTION

Information security is a multi-disciplinary area, combining organizational actions and working with technical protecting information facilities. To assess information objects security, information security professionals use a variety of methods and tools, including not only the analysis of technical factors and monitoring results, but also, the estimate of social and organizational data. Examples of such tasks include assessment of the presence of leaks and data in open sources, testing for penetration, network swapping analysis.

Due to such subject areas variety, the researcher has to deal not only with poorly formalized or incomplete data, but also with data of different semantics, i.e. heterogeneous data (technical, social, etc.) and perform their joint analysis. Thus, the task of analyzing this kind heterogeneous objects is highly relevant for the current situation in the information security industry.

This article is devoted to the analytical systems construction for working with heterogeneous, poorly structured data, including questions of their joint analysis, similarity assessment and learning.

II. RELATED WORKS

Solving various analytical problems in the field of information security is a highly urgent task. For the last twenty years, a large number of works have been devoted to it, including ultimate works in the field of large data [1] and artificial intelligence [2].

In terms of the knowledge base and analyzing information representing, a number of approaches, methods and techniques can be distinguished [3]. First, systems based on rules are widely used. This class includes most of expert systems [4,5] and many solutions in the field of information security [6]. In such systems, the knowledge base is set expertly in the form of a set of rules, ontologies, frames, or other representation reducible to the rules system. The analytical conclusion is based on the initial data application to the knowledge base rules predicates and generation of conclusion chains.

Secondly, it is necessary to note machine-learning systems. These systems are based on various methods of Data Mining and widely used for solving various information protection problems [7,8]. The methods of Data Mining consist in the large sets training data application to reveal the subject domain laws.

Third, precedent systems [9]. In the information security field, we can mention [10] which analyses security incidents, but its methods are not extended to complex systems analysis and penetration testing.

Analyzing approaches to the construction of analytical systems in the field of information security, we can note the wide application of systems based on expert knowledge. In recent times, approaches using machine learning have also been popular. Precedent analysis, in turn, is not so popular that it is conditioned by the complexity of formalizing multi-disciplinary precedents.

When choosing an approach to build an analytical system, it is necessary to note the following:

1. The variety of modern devices, software, and, most importantly, attack variants, the development of APT attacks, make it impossible for the expert to formalize and maintain in the actual state a subject area, no matter how considerable it is.
2. In many problems that operate with heterogeneous, poorly structured data, there are not enough training

Project is financially supported by the Ministry of Education and Science of the Russian Federation, Federal Program "Research and Development in Priority Areas of Scientific and Technological Sphere in Russia for 2014-2020" (Contract No. 14.578.21.0231; September 26, 2017, the unique identifier of agreement RFMEFI57817X0231).

978-1-5386-5824-6/18/$31.00 ©2018 IEEE

sets (for example, in the field of penetration testing [11]).

Therefore, we propose to use a precedent approach to constructing an analytical system with using mathematical statistics elements and modeling.

III. THE APPROACH TO BUILDING A DECISION SUPPORT SYSTEM FOR SECURITY EVALUATION

The input data of the system for analyzing heterogeneous, weakly structured data are the results of evaluating the objects of the domain. Depending on the specific task, it can be different combinations of the technical devices operation results, monitoring and scanning systems, analysis of social networks, internet resources, messages, network traffic and other sources. The main properties of the input data are:

1. Semantic and syntactic nonuniformity (heterogeneity).
2. Weak degree of structuring (in case of account natural-language texts, graphic images, video and other similar types of information).
3. Incompleteness.

Semantic heterogeneity is generated by the data sources heterogeneity, and weak structuring – both the source and the nature of the incoming information. Incompleteness is due to the difference in data sets (including formalized ones) from one source under different circumstances. For example, information amount about a distributed system obtained as a result of scanning can be significantly different.

Based on their properties in the analytic system we propose to formalize this kind of data in the form of a set of objects characterized by simple properties [11]. According to [12], in this case the set of objects is specified as, where, where O is the set of objects; C – properties of objects; Unknown – the fact that property is known, but its value is not defined; Value – the value of the property; RefO – property – reference to the object. In this case, we can talk about the representation of an object as exclusively a set of properties ("bag of attributes").

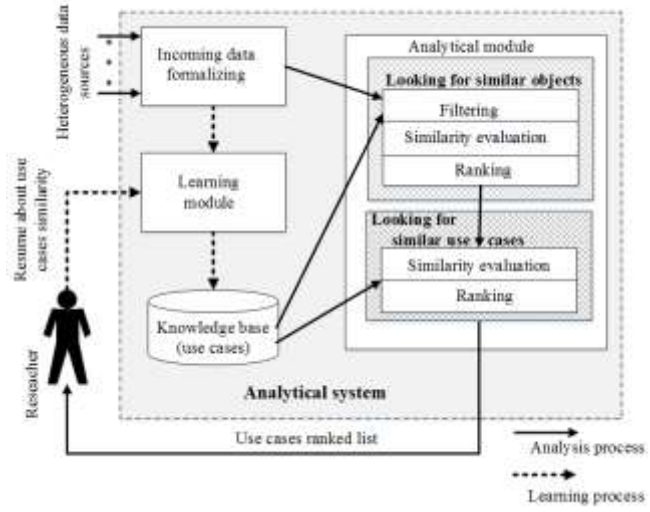

Fig. 1. Structure of the evaluating security decision support system

The output of the analytical system should be recommendatory proposals for the situation in question. Depending on the problem specific nature, it could be: the presence of a dangerous situation, possible system vulnerabilities, and others. In general, the task is to search the knowledge base and build conclusions based on stored and incoming information.

From this point of view, precedent $p = \{o \mid o_j \in O\}$, where O is the set of objects. That is, a precedent is a set of objects that describes a specific example of a domain, situation, attack, etc. By analogy with the natural language texts analysis [13] the approach can be called "bag of objects".

The structure of the system constructed in accordance with the chosen approach includes the basic blocks in the form of a knowledge base, an input data formalization block, an analysis block, and a training unit (Fig. 1).

IV. DATA ANALYZING AND ESTIMATION THE USE CASES SIMILARITY

The analytical system knowledge base is formed based on a well-known set of precedents. For tasks in which obtaining such a set is difficult, it is proposed to apply the precedent modeling methods based on expert estimates and statistical data with subsequent correction in the process of application through the training unit [14].

The corresponding ETL module upon entering the system performs input data formalization in accordance with the chosen model of their object representation.

The analysis unit performs two stages of analytical processing. At the first stage, object properties preliminary filtering of is carried out (for example, based on an evaluation of the property significance [11]) and an estimation of the objects similarity is carried out using an similarity integral measure.

First, binary property vectors are defined for objects, where 1 means that the object has the corresponding property, 0 is its absence. The vector coordinates are defined as the ordered properties corresponding to intersection of the objects properties set:

$$\left.\begin{array}{l} o_1 \to C_1 = \{c_1^1, ..., c_1^{k1}\} \\ o_2 \to C_2 = \{c_2^1, ..., c_2^{k2}\} \end{array}\right\} \to \overrightarrow{C_{1,2}} = C_1 \cup C_2$$

o1, o2 – the objects described by the property sets;

c_j^i – the j-th object i-th property;

The two objects similarity is defined as the degree of the semilarity of the binary vectors describing them. Thus, the evaluation of the correspondence of two objects is formed by the presence of properties $s^1 \in \{0,1\}$, where 1 – objects completely coincide, and 0 – completely do not coincide.

If both objects have a property c^i, and in both cases its value is defined, you can compare objects by the this property

value. At this stage, only the coincidence of properties was taken into account. As a result, for any property, an estimate is obtained, where 1 – properties completely coincide, and 0 – completely do not coincide. If objects have n common defined properties, then the evaluation of the two objects correspondence by the properties value has the form

$$s^2 = \frac{\sum_{i=1}^{n} s_i^2}{n}$$

The measure of the objects similarity is defined as $s_i = \alpha * s_i^1 + \beta * s_i^2$, where the weight coefficients sum is: $\alpha + \beta = 1$. For properties that cannot be compared by value $\beta = 0$.

The general measure of the two objects similarity: $s = \frac{\sum_{i=1}^{m} s_i}{m}$, where $m = \left|\overrightarrow{C_{1,2}}\right| = |C_1 \cup C_2|$.

Precedents similarity is determined based on their objects similarity. Since each precedent is an unordered objects set, at this stage the precedents similarity is evaluated depending on the following tasks:

A. Security estimating by testing for penetration.

Since the most important in this case is the possibility of penetration into the system, the precedents similarity is determined by the most similar objects that can be the attack entry point, that is, for precedents P_i and P_j the precedents similarity measure:

$$sim_{i,j} = \max(s_{k,l}) \quad (1)$$

where $s_{k,l}$ is the measure of the similarity of objects $o_k \in P_i, o_l \in P_j$.

B. Similarity estimation between objects of informatization.

In this case, the general similarity of systems is important, calculated as the average measure of the objects similarity, taking into account the number of objects of the analyzed precedent:

$$sim_{i,j} = \frac{\sum s_{k,l}}{K} - s_{k,l} \quad (2)$$

a measure of the similarity of objects $o_k \in P_i, o_l \in P_j$, K – objects number in the input precedent. The objects number in the precedent from the knowledge base is not taken into account, since it is important not only the overall total similarity between systems, but also the coincidence of the input case with any part of another precedent.

The results of the analysis block approbation are given in Table 1. The table lists the precedent objects types. They are not taken into account in the analysis and are given to show the quality of the assessment of the objects similarity. In the experimental check on the top of the ranked lists of similarity there were always objects of the same type with the original one. Also, both similarity measures are given for this example.

TABLE I. TABLE STYLES

Use case id	Type of object in test use case	Type of object in knowledge base use case	S	Sim (1)	Sim (2)
P5	IP	IP	1,00	1,00	0,33
P0	PC	PC	0,85	0,93	0,86
	PC	PC	0,85		
	Application	Application	0,93		
	Application	Application	0,83		
	Application	Application	0,85		
	Application	Application	0,79		
	Network	Network	0,85		
	E-mail	E-mail	0,89		
P2	PC	PC	0,75	0,93	0,57
	PC	PC	0,75		
	Application	Application	0,93		
	Application	Application	0,83		
	Application	Application	0,85		
	Application	Application	0,79		
	Network	Network	0,85		
P1	PC	PC	0,75	0,85	0,78
	PC	PC	0,75		
	Network	Network	0,85		

An important issue is the training the system in the process of use. This is necessary to maintain its relevance and compliance with the modern picture of the subject area in the field of information security. However, for the chosen approach, training is a difficult task due to the heterogeneity of the objects, the incompleteness and, often, the subjectivity of the precedent setting by the expert. To solve this problem, the authors suggest using a probabilistic estimate of each object property value importance. In this case, each property is a tuple <Measure, Value>, where $Measure \in [0,1]$. If the testing Pg precedent is successfully compared with the existing Pa precedent in knowledge base, then for all the Pa precedent vertices similar to at least one Pg object, for all the similar attributes increments Measure:

Measure = Measure + (1- Measure) * 0.5.

For different properties values takes place a reduction:

Measure = Measure * 0.5 is proposed. The value is considered as significant if $Measure > \sigma$, where σ is the threshold value. Authors test this approach on a set of real precedents when security estimation.

V. CONCLUSION

The authors developed an approach to the analysis of heterogeneous objects in the security estimation problems solution. A precedent analysis has been chosen as a basis for constructing an analytical system. Measures of objects and precedents similarity are developed. The structure of the decision support system based on the proposed approach includes the components of data preprocessing, analysis and

training. The proposed similarity measures performed well in an experimental verification of the approach.

The authors see further development as the decision support system training block improvement, enhancement methods for estimating objects and precedents similarity, and introduction of their structure explicit account in the evaluation precedents similarity.

REFERENCES

[1] Tianmu L., Leiming Y. SIEM Based on Big Data Analysis. Cloud Computing and Security. 2017. Vol. 10602. Pp. 167-175. DOI https://doi.org/10.1007/978-3-319-68505-2_15

[2] Lee S., Shin Y. The Direction of Information Security Control Analysis Using Artificial Intelligence. Advances in Computer Science and Ubiquitous Computing. 2017. vol 474. Pp. 872-877 DOI https://doi.org/10.1007/978-981-10-7605-3_138

[3] Artificial Intelligence Techniques: A Comprehensive Catalogue. Ed. Bundy A. Berlin. Springer. 1997. P. 129 DOI https://doi.org/10.1007/978-3-642-60359-4

[4] Jackson P. Introduction to Expert Systems (3rd ed.). USA. Boston. MA. Addison-Wesley Longman Publishing Co. 1998. 542 p.

[5] Osipov G.S. Dynamic intelligent systems. Artificial intelligence and decision making. No. 1. 2008. Pp. 47-54. (In Russian).

[6] Azhmukhamedov I.M., Marienkov A.N. Security of computer networks on the basis of the analysis of network traffic. Bulletin of ASTU. Management, computer engineering and Informatics series. 2011. No. 1. Pp. 137-141. (In Russian).

[7] Dua S., Du X. Data Mining and Machine Learning in Cybersecurity. New York. Taylor and Francis Group. LLC. 2011. 248 p. DOI: 10.13140/RG.2.2.35197.26085

[8] Tsvetkova O.L., Ajdinyan A.R. Intellegent system evaluation information security of the enterprise from internal threats. Bulletin of computer and information technologies. No: 8. 2014. Pp. 48-53. (In Russian).

[9] Karpov L.E., Udin V.N. Adaptive case-based management based on classification of managed object States. The proceedings of ISP RAS. 2007. No. 2. Pp. 37-57. (In Russian).

[10] Zhukov V.G., Shalyapin A.A. Case based analysis if information security incidents. Siberian journal of science and technology. 2013. No. 2 (48). Pp. 19-24. (In Russian).

[11] Poltavtseva M.A., Pechenkin A.I. Intelligent Data Analysis in Decision Support Systems for Penetration Tests. Automatic Control and Computer Sciences. 2017. Vol. 51. No. 8. Pp. 985–991. DOI https://doi.org/10.3103/S014641161708017X

[12] Poltavtseva M.A., Zegzhda P.D. Heterogeneous semi-structured objects analysis. Intelligent Systems Conference 2018 6-7 September 2018. (Unpublished).

[13] Wallach H.M. Topic modeling: beyond bag-of-words. Proc. of the 23rd int. Conf. on Machine learning. Pittsburgh. ACM. New York. 2006. Pp. 977–984. DOI: https://doi.org/10.1145/1143844.1143967

[14] Zegzhda P.D., Poltavtseva M.A., Pechenkin A.I., Lavrova D.S., Zaitseva E.A. Heterogenious semi-structured objects case-based reasoning in information security. Problems of information security. Computer systems. No. 1. 2018. Pp. 17-32. (In Russian).