UDC 004[056.5+822] https://doi.org/10.15407/pp2021.02.034

A.Ya. Gladun, K.A. Khala

ONTOLOGY-BASED SEMANTIC SIMILARITY TO METADATA ANALYSIS IN THE INFORMATION SECURITY DOMAIN

It is becoming clear that one of the most important resources to combat cyberattacks is the processing of large amounts of data in the cyber environment. There is also a need to automate the tasks of searching, selecting and interpreting Big Data to solve operational information security problems. For analyzing Big Data metadata, the authors propose pre-processing of metadata at the semantic level. As analysis tools, it is proposed to create a task thesaurus based on the domain ontology, which should provide a terminological basis for the integration of ontologies of different levels. For building task thesaurus, authors proposed to use the standards of open information resources. The development of an ontology hierarchy formalizes the relationships between data elements for machine learning, and development of artificial intelligence algorithms to adapt to changes in the environment, which will increase the efficiency of big data analytics for the cybersecurity domain.

Keywords: big data analytics, information security, cyber security, ontology, thesaurus, unstructured data, metadata, semantic similarity.

Introduction

Maintaining the growth and efficiency of enterprises while protecting confidential information is becoming increasingly difficult due to the ever-increasing threat of cybersecurity. The rise of cyberattacks is of great concern to both businesses and individuals. Also, the amount of information processed around the world has grown significantly over time, prompting cybersecurity to become more sophisticated and to introduce new methods of processing large amounts of data.

The use of big data itself can be incredibly useful, as it can not only help block any potential cyberattacks, but also help analyze vast amounts of data much faster and easier.

Obviously, data corruption prevention is one of the biggest big data challenges in cybersecurity. To make the most of big data, you need to know how to analyze it properly and use it to make the wisest.

Then cybersecurity big data analytics comes forward. It allows security professionals to analyze much more information and data than traditional cybersecurity solutions. Security systems use big data to automate the calculation of operations as correlation rules, which have the ability to dramatically reduce the number of false positives generated by the system.

The rapid growth in the popularity of big data analytics contributes to machine learning and deep learning, which are subsets of artificial intelligence. These teaching methods can process large amounts of data collected by the system and identify patterns that may indicate a cyber-threat. The challenge of safe big data is to analyze and process very large amounts of data in a timely manner to respond more quickly to incidents and obtain meaningful information that can be used by cybersecurity professionals.

The data itself faces big data challenges, which creates difficulties at every step from data collection to visualization and use. Thus, there is a need for a semantic context to access data and use and interpret results. For a semantic context, the same term can be represented differently, and therefore the result will depend on the context itself. However, you can find different concepts that represent the same object, or data that share a definition that is different from another. It is semantic technologies used to eliminate inconsistencies, evaluate and identify new information from existing knowledge bases, so it is advisable to consider different approaches that combine semantic technology with big data.

So Big Data is being used effectively today to make decisions in information se-

curity and cybersecurity systems. Big Data analytics allows you to make more informed decisions, ensure regulated implementation, and make recommendations to improve policies, guidelines, procedures, tools, and other aspects of network processes. The use of semantic modelling methods in Big Data analytics is necessary for the selection and combination of heterogeneous Big Data sources, recognition of patterns of network attacks and other cyber threats, which must occur quickly to implement countermeasures [1].

Metadata role's for interpretation Big Data

Big Data analytics in information security needs to solve the tasks of external units of Big Data. These data are used to predict and stop cyberattacks. Attack prevention and threat intelligence are becoming important for securing information systems and technologies.

In order for a data set to be considered big data, it must have one or more of the following characteristics: volume; speed; diversity; certainty; value [2]. Volume is the volume of data sets, i.e. the amount of data generated; speed (speed of formation and transmission of data) covers the structure, behaviour and permutation of data sets; diversity (type of structured and unstructured data) encompasses the tools and methods used to process large or complex data sets. Reliability refers to the quality or accuracy of the data, which can cause data processing to eliminate errors and noise. Value is defined as the usefulness of data and it is intuitively related to reliability, because the higher the accuracy of the data, the greater their usefulness.

Metadata (see Fig.) for Big Data are blocks of data, both physically attached to big data and located externally from Big Data. These metadata provide information about the characteristics and structure of the Big Data sets: name; data origin, data source information; source; XML tags indicating the author and date of creation of the document; attributes indicating the size and formatting, control of the total amount; number of data set records; image resolution; brief description of data, etc. [3, 4].

It is important that metadata is machine-readable, as this helps maintain the origin of the data throughout the lifecycle of big data analytics, which helps to establish and maintain the accuracy and quality of the data.

Thus, there is a need for semantic analysis of Big Data metadata based on the development of methods for analyzing natural language (NL) metadata texts using the Big Data ontology, which formalizes the knowledge and features of the domain and allows for semantic processing, if necessary, other elements of big data metadata.

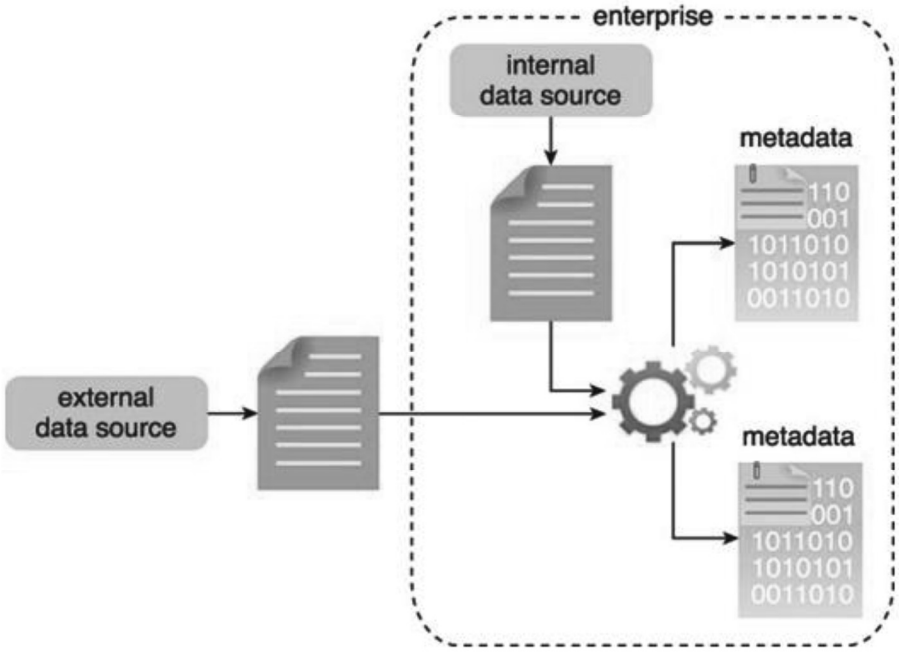

Fig.1. Metadata are adding to data from internal and external sources [1]

Ontological analysis for information security

The basis of the cybersecurity ontology is the need for a common language that includes basic concepts, complex relationships, and basic ideas. The most important feature of the cybersecurity ontology is that it illustrates the relationship between all the elements. By creating a correct and coherent cybersecurity ontology, cybersecurity professionals around the world can communicate effectively and develop a common understanding of important areas in this field.

Because cybersecurity ontologies are unique in that they cover the relationship between each entry in the ontology, this allows cybersecurity professionals to make faster and more accurate decisions. In addition, the ability to see the relationship between incidents, events and concepts provides valuable information.

Cybersecurity ontologies have become increasingly popular in recent years, as such a taxonomy allows cybersecurity professionals in different organizations or even in different countries to communicate faster and more efficiently, as well as to use their resources more efficiently. Also, ontologies can be very useful for describing critical vulnerabilities, risky vulnerabilities, and vulnerabilities that can harm organizations, employees, and regular users who use mobile devices.

Today, there are a large number of ontologies for information security that reflect various individual aspects of this subject area. For example, researchers have developed application ontologies to identify and classify network attacks: ontology for distinguishing network security status [5]; ontology of intrusion detection [6]; ontology for automated classification of network attacks [7]; ontology for predicting potential network attacks [8].

Other ontologies can provide an adaptive vocabulary that can improve behavioural analysis and help stop the spread of threats. Terms for such IS ontologies can be obtained from open sources, such as a dictionary of IS terms [9] and the standards of this subject area.

These ontologies describe the main artefacts of a cyber-attack to support the overall presentation of collected data and reuse, namely:

- the attacker's network environment, including its IP address, network size, range and name;
- the attacker's hosting environment, including information on the hosting operating system and its vulnerabilities, detected open ports, the level of the blacklist of the host in question based on its IP address, the number of virtual hosts;
- the type of organization where the attack will take place, as well as information about the location of the host with coordinates;
- type of attack based on its classification according to existing cyberattack dictionaries;
- date, day and time of the attack, taking into account the time zone of the attacker.

Clarifications or sub-concepts regarding countermeasures, assets, threats and vulnerabilities consist of a specific technical vocabulary. The dictionary was compiled from literature and security taxonomies. Ontology is implemented in OWL, where concepts are implemented as classes, relationships are implemented as properties, and axioms are implemented with constraints. In Fig. a fragment of such an ontology of upper level of cybersecurity is given.

To select and interpret the external blocks of Big Data, an ontology of the problem to be solved in the field of cybersecurity is used. The cybersecurity information system contains a hierarchical structure of interconnected ontologies: domain ontology, Big Data ontology, and task ontology. To select Big Data blocks, the task ontology can be replaced by a task thesaurus, which can be built by a Big Data ontology, you need to select a set of classes and a set of instances of classes. It is also advisable to highlight the relationship between attribute instances and their values. The following formal model was used to describe the ontologies of big data:

$$O_{BD} =< C, R, I, D_t, A > \quad (1)$$

which contains the following elements:

$C = C_C \cup C_{In}$ – a set of ontology concepts, where C_C a set of classes, C_{In} a set of class instances;

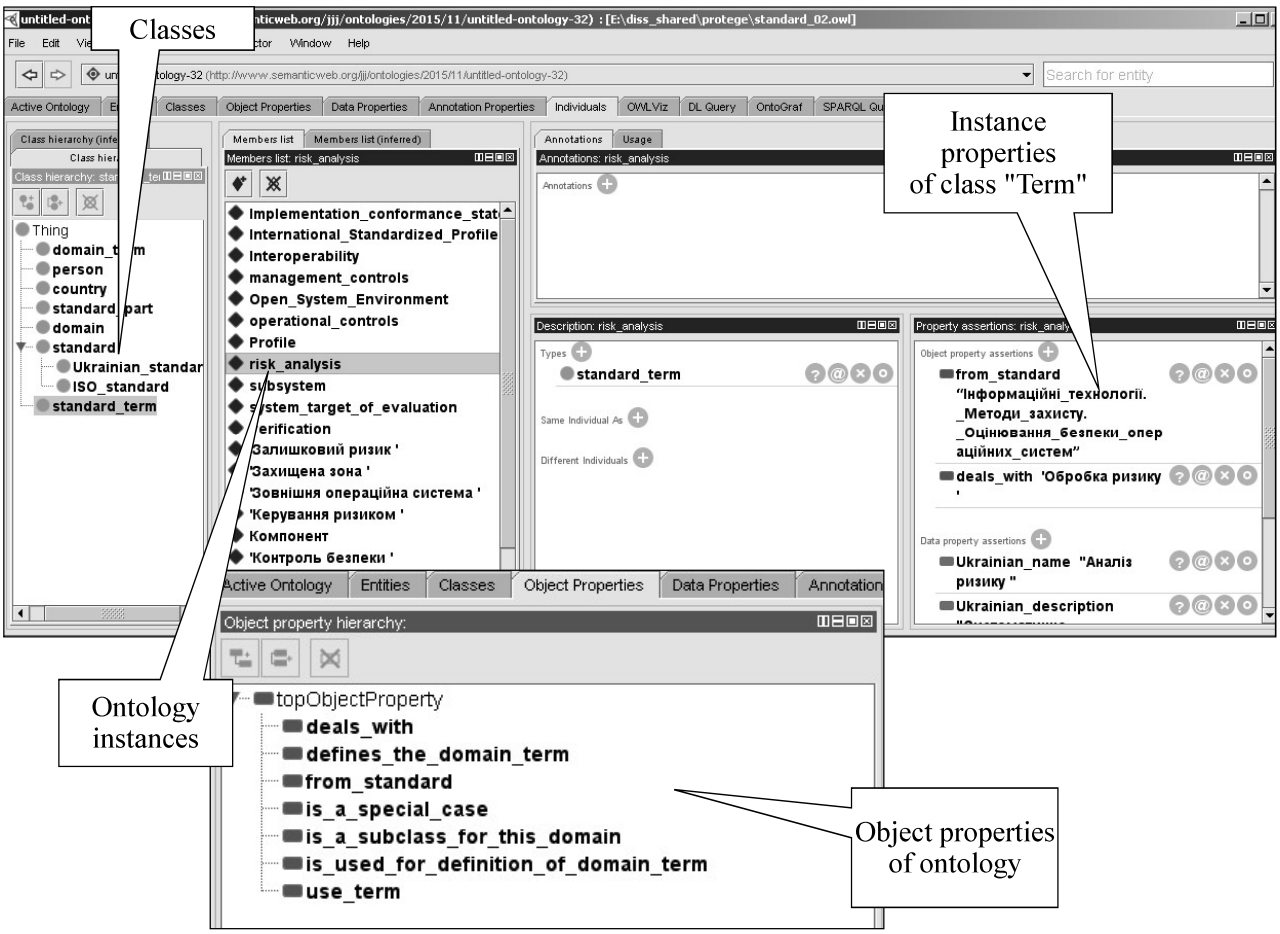

Fig. 2. The fragment of ontology of the cybersecurity [10]

$R = cr_{er} \cup \{or_i\} \cup or_{ie} \cup \{dr_j\} \cup dr_{er}$ – the set of relations between the elements of the ontology, where cr_{er} – hierarchical relations between ontology classes; $\{or_i\}$ – a set of object properties that establish relationships between instances of classes; or_{er} – hierarchical relations between ontology classes; $\{dr_j\}$ – a set of data properties that establish the relationship between instances of classes and values from D_t; dr_{er} – hierarchical relations between the properties of these instances of ontology classes;

$I = \{I_C \cup I_P\}$ – a set of characteristics that can be used for logical inference over the ontology;

D_t – a set of data types for dr_j ontology class values; A – a set of rules.

Task thesaurus is a special case of the subject area ontology, which contains only ontological terms, but does not describe the semantics of the relationships between them in order to analyze NL texts. It can be automatically generated from the ontology of the subject area and the description of the problem in NL [11]. In the task thesaurus for concepts and relationships, a weight is introduced that indicates the degree of significance of a concept or relationship that improves the quality of model processing. The formal model of the thesaurus has the form:

$$T = < C_t, R_t, Inf > \qquad (2)$$

where $C_t \subseteq C$ – final set of terms; and $R_t \subseteq R$ – the final set of relationships between these terms, Inf – additional information about timing (e.g. weight).

The task thesaurus has a simpler structure because it does not cover ontological relations and for each concept has additional information as a weight $w_i \in W, i = \overline{1,n}$. Then the formal model task thesaurus is defined as a set of ordered pairs $T_{task} = < (c_i \in C_t, w_i \in W), \varnothing, Inf >$ with more information in Inf regarding the source of the ontology. The algorithm of thesaurus generation for InRs has the following main stages:

Formation $Docs = \{doc_i\}, i = \overline{1,n}$, initial set $Docs$ from text documents doc_i related to InRs, where each document doc_i from the set $Docs$ has a weight factor W, which allows you to determine the importance of document elements for the InRs thesaurus.

Formation of the dictionary InRs $D(doc_i)$ for everyone doc_i, which contains all the words found in the document. Then the dictionary D_{Docs} formed as a sum $D(doc_i)$:

$$D_{Docs} = \bigcup_{i=1}^{n} D(doc_i).$$

Generation of InRs thesauri T_{res}, as projections of a set of ontological concepts C on the plural D_{Docs}. $T_{res} \subseteq C$. This processing step is aimed at removing terms from other non-user domains and stop words. The main problem at this stage is the semantic relationship of fragments of NL with T_{res} with the concepts of the set C domain ontology O_{BD}. It can be solved by linguistic methods that use lexical knowledge bases for each NL that go beyond this work. Each word from the thesaurus must be associated with one of the ontological terms. In the case of a lack of relations, the word is considered as a stop word or an element of notation, in which case it must be rejected.

A semantic *bunch* $R_{t_j}, j = \overline{1,n}$ is a group of thesaurus InRs associated with a single ontological term, used to train the semantics of documents written in different languages, and treated as $\forall p \in T_{res} \in R_{t_j}$, where $R_{t_j} = \{p \in D_{Docs} : Term(p) = c_j \in C\}$.

In the case where the domain ontology O_{BD} is not defined, we assume that the domain has no restrictions, and therefore does not remove any elements from the dictionary InRs: $T_{res} = D_{Docs}$.

Task thesaurus can also be generated based on InRs thesauri using set operations like sum, intersection and complement sets. Thus, a thesaurus of a particular domain can be formed as the sum of thesauri InRs related to that domain. The weight of a term for a given amount operation is defined as the sum of its weight in each InRs.

Methods for assessing semantic similarity for generating thesauri based on ontology

Semantic similarity is a field of research that is actively developing, which is based on an attempt to calculate the relationship between words, concepts, sentences, paragraphs and documents. The similarity between two words is a measure of the probability of their meaning, calculated on the basis of the properties of concepts and their relationships in the ontology. Semantic similarity plays a fundamental role in information management, especially for unstructured data that in addition comes from a variety of flexible sources.

Semantic similarity is used to encompass similarity measures that use an ontology structure or external sources of knowledge to determine similarities between entities within one ontology or between two different ontologies. Potential applications of these measures are knowledge identification and decision support systems that use an ontology.

Semantic similarity concepts are a subset of the domain concepts that can be joined by some relations or properties. If domain is modeled by ontology then Semantic similarity concepts is a subset of the domain ontology concepts. There are several ways to build semantically similar concepts, which can be used separately or together. The user can define Semantic similarity concepts directly (manually – by choosing from the set of ontology concepts or automatically – by any mechanism of comparison of ontology with description of user current interests that uses linguistic or statistical properties of this description).

Semantic similarity concepts can join concepts linked with initial set of concepts by some subset of the ontological relations (directly or through other concepts of the ontology). Each semantically similar concept has a weight (positive or negative) which determines the degree of semantic similarity of the concept with the initial set of concepts.

A lot of different approaches used now to quantifying the semantic distance between concepts are based on ontologies that contain these concepts and define their relations and properties.

Factors related to the hierarchy of ontologies can affect the measurement of semantic distance: path length, depth and local density. Similarity measures and taxonomy are interconnected by taxonomic connec-

tions, i.e. the position of concepts in the taxonomy, the number of hierarchical connections and the information content of concepts are considered.

Approaches to calculating semantic similarity can be classified into the following main categories:

• by structure - approaches based on the structure or calculation of edges, semantic similarity based on taxonomic relations of the ontology hierarchy (is-a, part-of). They calculate the length of the path connecting the terms and the position of terms in the taxonomy. Thus, the more similar the two concepts, the more connections between the concepts and the more closely they are related [12] [13].

the shortest path is a simple, powerful measurement designed first and foremost to work with hierarchies. Where Max is the maximum path length between C_1 and C_2 in the taxonomy, and SP is the short path connecting C_1 with C_2:

$$Sim(C_1, C_2) = 2 * Max(C_1, C_2) - SP \quad (3)$$

Hirst and St-Onge - measure HaS [14] calculates the relationship between concepts, using the distance between the nodes of concepts, the number of changes in the direction of the path connecting the two concepts, and the acceptability of the path. Where SP is the short path connecting C_1 to C_2, d is the number of direction changes, C and k are constants. Thus, the longer the path and the more direction changes, the smaller the Sim:

$$Sim_{HaS}(C_1, C_2) = C - SP - k * d \quad (4)$$

Wu and Palmer - WaP measure [15] calculates the similarity, taking into account the depth of the two concepts in the taxonomies of WordNet, as well as the depth of LCS (Least Common Subsumer (LCS), the formula:

$$Sim_{WaP}(C_1, C_2) = 2 \times \frac{depth(LCS(C_1, C_2))}{depth(C_1) + depth(C_2)} \quad (5)$$

• by terms of information content - approaches based on the content of information use, the information content of concepts to measure the semantic similarity between two concepts. The value of the information content of the concept is calculated based on the frequency of the term in this collection of documents.

Lin – Lin et al. [16] [17] proposed a measure based on an ontology bounded by hierarchical connections and corpus. This similarity takes into account the information between two concepts, such as Reznik [18], but the difference between them is in the definition:

$$Sim_{Lin}(C_1, C_2) = \frac{2 * \ln(p_{mis}(C_1, C_2))}{\ln(p(C_1)) + \ln(p(C_2))} \quad (6)$$

• by characteristics - characteristics-based approaches assume that each term is described by a set of terms that indicate its properties or characteristics. The degree of similarity between two terms is determined according to their properties or according to their relationship with other similar terms in the hierarchical structure. Tversky [19] takes into account the characteristics of terms to calculate the similarity between different concepts, ignoring the position and information content of terms in the taxonomy. Each term should be described by a set of words indicating its characteristics.

$$Sim_{tvsk}(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha |C_1 - C_2| + (\alpha - 1)|C_2 - C_1|} \quad (7)$$

where C_1 and C_2 represent the corresponding sets of descriptions of the two terms. $\alpha \in E[0,1]$ is the relative importance of unusual characteristics. The value of α increases with commonality and decreases with the difference between the two concepts. The definition of α is based on the observation that similarity is not necessarily a symmetric relationship.

Many measures take into account only the path length between concepts. The basic idea of such estimates is that the similarity of the two concepts is a function of the path length that connects concepts (by taxonomic relation "*is-a*") and their positions in the taxonomy. The same approach can be applied to arbitrary domain ontology where path between concepts can consist of all ontological relations.

Semantic similarity estimation parameters from various approaches (for example,

from (3)-(7)) can be used for generation of task thesaurus. We can consider such thesaurus as a set of concepts that have semantic distance from some initial set of concepts greater than some constant ones. In these estimations we can use different coefficients for universal and domain-specific relations R of domain ontology O_{BD}.

Conclusions

Prospects for automating the creation of thesauri based on ontologies depend on the availability of appropriate domain ontologies and well-structured, reliable InRs that characterize the needs and interests of users in information. Therefore, we can find InRs where such parameters are clearly defined and can be processed without additional pre-processing. Semantic Wiki, where the relationship between concepts and their characteristics is determined by semantic properties, meet the following conditions.

Big data is the best way to develop when it comes to cybersecurity, as identifying threats at the earliest opportunity becomes easier. Big data undoubtedly has advantages for any business that requires regular processing of large amounts of data. But despite this, the increasingly sophisticated methods used by cybercriminals are becoming increasingly difficult to combat. In large organizations with hundreds of employees, the system collects and analyzes huge amounts of data. Security professionals can use this information to predict trends and improve cybersecurity. With this in mind, it is safe to say that optimal approaches to cybersecurity should be used.

The semantic similarity was reviewed and its important role in the task of automating the creation of thesauri based on ontology was emphasized.

References

1. Erl T., Khattak W., and Buhler P.: Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall, ServiceTech press, 2016.
2. P. Buneman, S. Davidson, M. Fernandez, D. Suciu:Adding structure to unstructured data, In 6th International Conference on Database Theory, pp. 336-350. Delphi, Greece, 1997.
3. Smith K., Seligman L., Rosenthal A.: Big Metadata: The Need for Principled Metadata Management in Big Data Ecosystems. In Proceedings of the Company DanaC@SIGMOD, p. 46-55. Snowbird, UT, USA 2014.
4. Dey A., Chinchwadkar G., Fekete A., Ramachandran K.: Metadata-as-a-Service. In Proceedings of the 31st IEEE International Conference on Data Engineering Workshops, p.6-9. IEEE, Seoul, South Korea, 2015.
5. Salahi A., Ansarinia M.: Predicting Network Attacks Using Ontology-Driven Inference.In IJICTR, IGI Global, vol. 4, no. 2; pp. 27-35, 2012.
6. Bhandari P., Guiral M.S.: Ontology Based Approach for Perception of Network Security State. In Proc.of Recent Advances in Engineering and Computational Sciences, Chandigarh, pp.1-6, 2014.
7. Oltramari A., Cranor L.F., Walls R.J.: Building an Ontology of Cyber Security. In Proc. 9th Inter. Conf. on Semantic Technologies for Intelligence, Defense, and Security, Fairfax, pp. 54-61, 2014.
8. Wang J.A. and Guo M.,: OVM. An Ontology for Vulnerability Management. In Proc. 5th Annu. Conf on Cyber Security and Information Intelligence Research, Knoxville, pp. 1-4, 2009.
9. Gladun A.Y., Puchkov O.O, Subach I.Yu., and Khala K.O.: English-Ukrainian dictionary of terms on information technology and cybersecurity. Kiev, Ukraine: NTUU KPI named by Igor Sikorsky, 2018.
10. Protégé 5.0. [Online]. Available: https://protege.stanford.edu/. Accessed on: Nov 24, 2020.
11. Gladun A., Rogushina J.:Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources. International Journal of Intelligent Systems and Applications, №1, pp.11-20, 2012.
12. Rada R., Mili H., Bicknell E.: Development and application of a metric on semantic nets. In Proceedings of the IEEE transactions on systems, man, and cybernetics, p. 17–30, 1989.
13. Richardson R., Smeaton A., Murphy J.: Using WordNet as a knowledge base for measuring

semantic similarity between words. Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University, 1994.
14. Hirst G., St-Onge D.: Lexical chains as representations of context for the detection and correction of malapropisms. In Proceedings of the WordNet: An electronic lexical database, vol. 305, p. 305–332, 1998.
15. Wu Z., Palmer M.: Verbs semantics and lexical selection.In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, p. 133–138, 1994.
16. Lin D.: An information-theoretic definition of similarity. In ICML, vol. 98, p. 296–304, 1998.
17. Lin D.: Principle-based parsing without overgeneration. In Proceedings of the 31st annual meeting on Association for Computational Linguistics, p. 112–120,1993.
18. Resnik P.: Semantic similarity in a taxonomy. An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), vol. 11, p. 95–130, 1999.

Received: 07.05.2021

About authors:

Gladun Anatoliy Yasonovych,
Candidate of Technical Sciences, Senior Research Fellow.
Number of scientific publications in Ukrainian publications - 188.
Number of scientific publications in foreign publications - 75.
http://orcid.org/0000-0002-4133-8169,

Khala Kateryna Oleksandrivna,
Researcher.
Number of scientific publications in Ukrainian publications - 31.
Number of scientific publications in foreign publications - 8.
http://orcid.org/0000-0002-9477-970X.

Affiliation:

International Research and Training Center of Information Technologies and Systems of National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine.
03187, Kyiv,
Academician Glushkov Avenue, 40.
Tel: 044 502 6366.
E-mail: glanat@yahoo.com,
cecerongreat@ukr.net