INFORMATION TECHNOLOGY

DOI 10.20535/2411-1031.2020.8.2.222559
UDC 004[056.5+822]

ANATOLY GLADUN,
KATERINA KHALA,
IHOR SUBACH

ONTOLOGICAL APPROACH TO BIG DATA ANALYTICS IN CYBERSECURITY DOMAIN

Information security is a dynamic field in which methods and means of protection against threats and their destructive component are rapidly changing and improving, which is a challenge for organizations and society as a whole. Therefore, information systems related to cybersecurity require a constant flow of knowledge from internal and external sources, the volume of which is constantly growing. The introduction of big data sets in the field of cybersecurity provides opportunities for application for the analysis of data containing structured and unstructured data. The application of semantic technologies to search, selection of external big data, and description of knowledge about the cybersecurity domain require new approaches, methods, and algorithms of big data analysis. For selecting relevant data, we are offered a semantic analysis of metadata that accompanies big data and the construction of ontologies that formalize knowledge about metadata, cybersecurity, and the problem that needs to be solved. We are proposed to create a thesaurus of problems based on the domain ontology, which should provide a terminological basis for the integration of ontologies of different levels. The cybersecurity domain has a hierarchical structure, so the presentation of formalized knowledge about it requires the development of the hierarchy of ontologies from top to bottom. For building a thesaurus of problem, it is proposed to use an algorithm that will combine information from information security standards, open natural information resources, dictionaries, and encyclopedias. It is suggested to use semantically marked Wiki-resources, external thesauri, and ontologies to supplement the semantic models of the cybersecurity domain.

Keywords: big data analytics, cybersecurity, ontology, thesaurus, unstructured information, metadata, wiki technologies, semantic similarity.

Problem Statement. For today's, the issue of information security (IS) is a major problem in the activities of every organization or individual. The sphere of IS is a very dynamic area that requires constant monitoring of information coming from both internal and external information resources. Using data from networks and computers, analysts can extract useful information from data using analytical methods and processes. The dynamicity of the IS area generates big data that need to be processed in batch or transactional mode for rapid IS decision-making. Big data Analytics is used for decision making in information security systems [1]. The decisions can be made more judiciously using the results of the analysis, including actions that need to be taken and recommendations for improving policies, guidelines, procedures, tools, and other aspects of network processes. Big data analytics is a new analytics technology that enables the collection, storage, processing, and visualization of vast amounts of data. Such analytics take into account the main characteristic properties of big data [2].

The analysis of other works shows that, despite the high interest in big data, their analytics for cybersecurity, and availability of various technological means of their storage and processing, there are currently no relevant methods for selecting a pertinent subset of external big data blocks based on semantic description of metadata suitable for problem-solving. This is due to the unstructured nature of big data, its complexity, and diversity. The traditional metadata is the technical information that

characterizes the time of content creation, its volume, formats, etc., but does not relate to the content of the information contained in this data. Often, metadata are their meta-descriptions in natural language for many blocks of big data, ie annotations or explanations that require further semantic processing to make an appropriate decision about the choice of big data. There may be no problems with the selection of internal big data, as they have been accumulated within the organization in the process of performing cybersecurity tasks. The difficulties arise with external blocks of big data, which are needed for a complete data set and obtaining a quality result based on analytics. Creating an ontology of the cybersecurity domain will solve the problem of analyzing natural language annotations and the relevance of selected blocks of big data to solve the problem of IS. The ontology aligns terminology with those knowledge structures that are related to modern IS standards.

Analysis of recent researches and publications. The current situation regarding the storage, exchange, and processing of information, characterized by the intensive introduction of technology, the spread of local, corporate and global networks in all spheres of life of a civilized state, creates new opportunities and quality of information exchange. In this context, the question arises that information is protected and secure at the same time. The need to address this issue is actualized by the factors listed in [3].

Thus, we are faced with the task of determining the relationship of the basic concepts used in the search and navigation of resources related to the categories of "information security", "IS" and "cybersecurity".

In the work [4] it is proposed to use a thesaurus approach to formalize domain terminology. The information security thesaurus reflects a wide range of essential properties, features, and relationships inherent in this specific type of security. We agree with researchers [5] - [7], who distinguish three groups of terms of information security theory:

– terms that define the scientific basis of IS. This group includes terms that are used in many fields of knowledge and are unambiguous, semantically unified, and stylistically neutral;

– terms that define the subject basis of IS. This group of terms denotes concepts and their relationship with other concepts in the field of information security as a special field of knowledge;

– terms that determine the nature of the activities to ensure IS. This group includes terms denoting objects, phenomena, processes, their properties, and relations characteristic of this sphere.

But a more universal approach to big data analytics requires the use of ontological analysis and the creation of an IS orthology. Today, there are many developments in this area, but the issues of their mutual integration and compatibility with applicable systems remain open. The ontological model of IS allows to formally define the relationship between the basic concepts related to the categories of "information security", "IS" and "cybersecurity".

The IS thesaurus is integrated with the concepts of information security, application security, network security, Web security, and security of critical information infrastructure [8], [9]. Application security is defined as application software products, as well as information and software resources and processes involved in their life cycle. Network security involves the design, implementation, and use of networks within an organization, between organizations, between organizations, and users. Internet security refers to Internet services and related information and communication technology systems and networks.

For using big data as an external source of information, such as from the Web, you must first filter out the desired pertinent set of data sets that will be used for big data analytics. Preliminary cleaning of data and their preparation for analysis can be performed based on metadata analysis (eg, annotations). Metadata – data about data or data elements, possibly also their data descriptions, data on data ownership, access paths, access rights, and data changes during their processing.

For a data set to be considered big data, it must have one or more characteristics, the so-called "5V" characteristics: volume, speed, diversity, certainty, value [10].

Reliability and value are very important characteristics for obtaining high-quality results of big data processing. Reliability refers to the quality or accuracy of data that may cause data processing to eliminate erroneous data and noise. Value is defined as the usefulness of data for the enterprise and the characteristic is related to reliability, because of the higher the accuracy of the

data, the greater their usefulness. The value also depends on the data processing time, as the analytical results have a certain shelf life. Also in [11] identified the main problems that exist today in big data technology and need to be addressed. Analysis of scientific publications [12] shows that the question of the relevance of metadata used in big data is more acute than ever, so new strategies and approaches are being developed today.

The purpose of the article is to develop a thesaurus of the task to find the appropriate big data (an existing problem for solving), the new terms of which will supplement the terminological set of the domain ontology and establishing semantic connections between them. And also develop the generation algorithm to build such thesaurus with using of information security standards, open dictionaries on information security and encyclopedias, and descriptions of competencies of IS specialists are proposed. Thesaurus IS will allow displaying the most important knowledge of the area for the task, but the time for its processing and construction will be much less than the time of comparing ontologies with unstructured natural language text. Such thesaurus will significantly speed up the analysis of descriptions of learning outcomes.

The main material research. The introduction of big data sets in the field of IS opens up opportunities for the analysis of very large sets containing both structured and unstructured data. The availability of big data sets has created difficulties that we have to deal with not only in terms of semantics and analytics but also in terms of data management, storage, and distribution. However, an ontological approach based on analytical data provides a practical basis for addressing the semantic challenges presented by data sets. The life cycle of big data analytics (see Fig. 1) is nine stages [2]:

1. Analysis of the problem.
2. Data identification.
3. Collecting and filtering data.
4. Data transformation.
5. Checking and cleaning data.
6. Aggregation and presentation of data.
7. Data analysis.
8. Data visualization.
9. Use of analysis results.

Analysis of the problem. The IS big data analytics lifecycle begins with the rationale, motivation, and purpose of the analysis. This analysis allows you to determine the type of big data to be used (batch, transactional, internal, external).

Data identification. The data identification stage defines the data sets required for analytical calculations and their sources. Using a wider range of data sources can increase the likelihood of finding hidden patterns and correlations.

Collecting and filtering data. In this step, data is collected from all data sources that were identified in the previous step. The data is then filtered to remove corrupted data or data that is not relevant for analysis purposes.

Metadata (see Fig. 2) can be added to large internal data or external data to improve knowledge about them, their classification, and queries. Examples of added metadata include the size and structure of the dataset, source information, creation or collection date and time, and language-specific information. It is very important that the metadata is machine-readable and passed on to subsequent stages of analysis.

Data transformation (data extraction). This step is designed to transform big data into a format that is used by the underlying analytics software.

Data validation and cleaning. Incorrect data can distort and falsify analysis results. This step is designed to create complex validation rules and remove any known invalid data. Big data solutions often get redundant data across different datasets.

Aggregation and presentation of data. This stage is designed to integrate multiple datasets together to achieve a unified view. Data can be spread across multiple datasets, requiring datasets to be combined through common fields, such as date or ID.

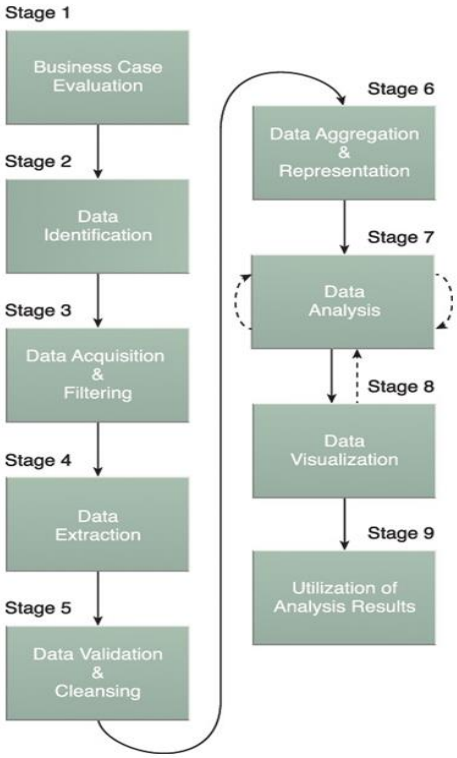

Figure 1 – The nine stages of the big data analytics lifecycle [2]

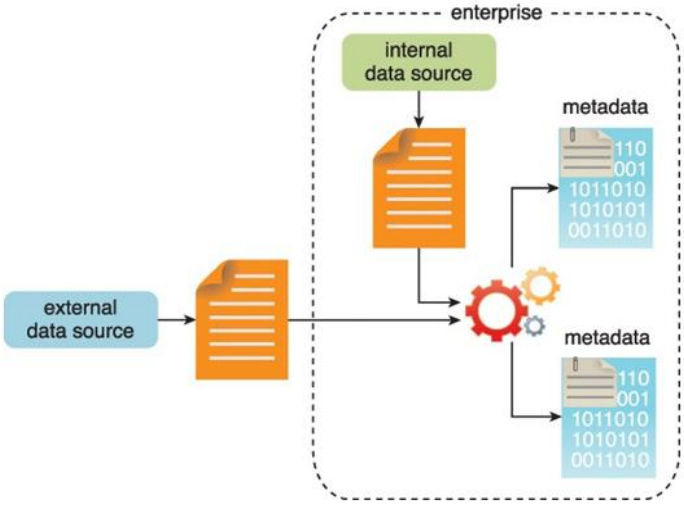

Figure 2 – Metadata is added to data from internal and external sources [2]

This step can be complicated by differences in:

– data structure – although the data format may be the same, the data structure model may be different;

– semantics – a value marked differently in two different datasets can mean the same thing, for example, “surname” and “last name”.

– data analysis. The data analysis phase is about performing the actual analysis task, which usually includes one or more types of analytics. This stage can be iterative, the analysis is repeated until a matching pattern or correlation is found.

Data visualization. The ability to analyze huge amounts of data and come up with useful insights doesn't matter if the analysts are the only ones who can interpret the results.

The semantic approach is used at all stages of the big data life cycle. Ontologies are widely used now in distributed intelligent applications to explicitly describe the domain knowledge system or information resource. Domain ontologies and task thesauri are the main semantic elements of metadata analysis. In the general case, ontology is an agreement on the shared use of concepts that provides the means of domain knowledge representation and agreement about their understanding. IS ontologies now become an active provider of data element relationships that can use machine learning and artificial intelligence algorithms to adapt to changes in the environment [13].

To create an ontology of the entire IS domain, it is necessary to integrate existing ontologies and improve them.

Unified IS ontology (UCO) [7]. Is designed to support the integration of knowledge in IS systems and should unify the most widely used information security standards. The ontology includes and integrates disparate data and knowledge schemes from different IS subsystems and is the most commonly used IS standards for sharing and sharing. The UCO can serve as a knowledge core for the IS domain.

A detailed description of the knowledge about the IS domain requires the development of a hierarchy of ontologies, starting from the top level to the bottom. The top-level ontology includes the basic concepts of the domain, which have previously been defined in ontologies on this topic. Below are mid-level ontologies that focus on the user, events, network operations, and geospatial data related to IS. Lower-level ontologies describe specific IS domains that require an industry-specific solution.

In the field of IS, a large number of ontologies have already been created that reflect various individual aspects of this subject area. For example, researchers have developed application ontologies to identify and classify network attacks: an ontology for distinguishing network security status [14]; ontology of intrusion detection [4]; ontology for automated classification of network attacks [15]; ontology for predicting potential network attacks [16].

Other ontologies can provide an adaptive vocabulary that can improve behavioral 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 [17] and the standards of this subject area.

This information, provided in Web Ontology Language (OWL), can be reused and integrated into a variety of applications. In Fig. 3, a fragment of such an ontology of upper level IS is given.

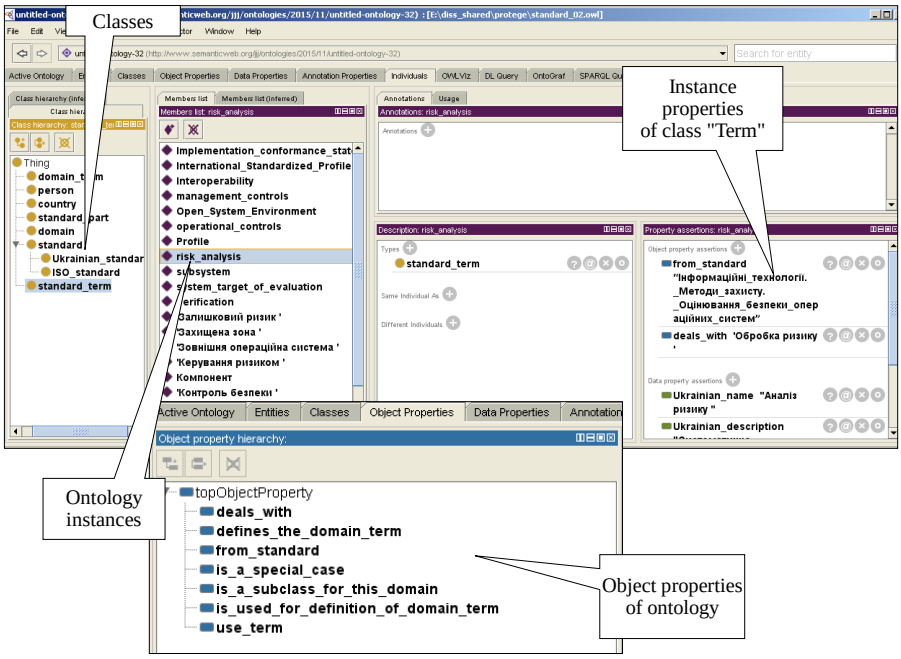

Figure 3 – Information security ontology (fragment) [18]

It is easier than from unstructured National League (NL) documents to extract information from those information resources (IR) that contain semantic markup. Examples of such IPs are semantized Wiki resources. Links between Wiki pages for which the content is explicitly defined can be used to build the ontology For example, on the portal of the Great Ukrainian Encyclopedia [19]. For this, you can use the pages of the category "Information security systems".

Ontology is a knowledge base that describes facts that are always assumed to be true within a particular community based on the generally accepted meaning of the thesaurus. Since thesaurus is a special case of ontology, which allows representing concepts so that they become suitable for machining and automated processing. It can be considered as a model of the logical-semantic structure of domain terminology. In the work [20] it is proposed to use a thesaurus approach to formalize the terminology of the subject area in the field of IS. Thesaurus IS reflects a wide range of essential properties, features, and relationships inherent in this specific type of security.

The task thesaurus is a special case of the subject area ontology, which contains only ontological terms (classes and instances), but does not describe (or limitedly describes) the semantics of the relationship between them to analyze natural language texts. It can be automatically generated by the ontology of the subject area and natural language description of the problem [21]. A simple thesaurus of the task is a thesaurus based on the terms of one ontology of the subject area. A compiled thesaurus of the task is a thesaurus based on the terms of two or more ontologies of the subject area.

Formal models either of ontologies or of thesauruses include as the basic concept the terms and connections between these terms. The collection of the domain terms with the indication of the semantic relations between them is a domain thesaurus. A formal model of thesaurus is based on formal model of ontology:

$$Th = < T_{th}, R_{th}, I > \quad (1)$$

where $T_{th} \subseteq T$ – finite set of the terms;

$R_{th} \subseteq R$ – finite set of relations between these terms;

I – additional information about terms (this information depends on specifics of thesaurus goals and can contain, for example, the weight of term or its definition).

Task thesaurus has the simpler structure because it does not include ontological relations and has additional information about every concept – it's weight $w_i \in W, i = \overline{1,n}$. Therefore, formal model of task thesaurus is defined as a set of ordered pairs $Th_{task} = < (t_i \in T_{th}, w_i \in W), \varnothing, I >$ with additional information in I about source ontologies.

The user has to formalize task if he/she needs the personified processing of information. The domain of task is formally characterized by domain ontology, and the task itself can be characterized formally by use of task thesaurus or informally – by its NL description, keywords, or example documents. The task thesaurus can be either built by the user manually or generated automatically by analysis of available NL documents and other IRs.

For construction of the task thesaurus, every IR is described by not empty set of the textual documents connected with this IR – text of content, meta descriptions, results of indexing etc. If IR contains multimedia content then this content can be transformed into text (by speech and text recognition methods etc.) methods. The algorithm of IR thesaurus generation has the following steps:

1. Formation of initial non-empty set A of the textual documents a_i connected with this IR as an input data for the algorithm. $A = \{a_i\}, i = \overline{1,n}$. Each of documents a_i from the set A has the coefficient of importance (for example, metadata of video are more important than the recognized speech) that allows defining the different weight of document elements for IR thesaurus).

2. IR dictionary construction. For every, a_i the set of words $D(a_i)$ is constructed. $D(a_i)$ is a dictionary that contains all words that occurred in the document. Dictionary of A is formed as a sum of the $D(a_i)$: $D_{IR} = \bigcup_{i=1}^{n} D(a_i)$.

3. Generation of IR thesauruses (see Fig. 4). With the use of domain ontology IR, thesaurus T_{IR} is created as a projection of the set of ontological concepts X into the set D_{IR}. $T_{IR} \subseteq X$. This step of processing is aimed to remove stop-words and terms from other domains that are not interesting for the user. The main problem deals with semantic connection of NL fragments (words) from T_{IR} with concepts from the set X of domain ontology O. This problem can be solved by linguistic methods that use lexical knowledge bases for every NL and is beyond the scope of this article. Each word from the thesaurus is necessary to link with one of the ontological terms. If the relationship is lacking the word is considered as a stop-word or marking element (for example, HTML tag) and should be rejected.

The group of the IR thesaurus words terms connected with one ontological term named *the semantic bunch* $R_j, j = \overline{1,n}$ is considered as a single unit: $\forall p \in T_{IR} \in R_j$, where $R_j = \{p \in D_{IR} : Term(p) = x_j \in X\}$. It allows to integrate processing of semantics of the documents written in various languages and, thus, to ensure the multilinguistic analysis of the Internet IR.

If user doesn't define domain ontology O then we consider that user domain of interests has no restrictions and therefore we don't remove any elements from IR dictionary: $T_{IR} = D_{IR}$.

Users can generate task thesauri based on IR thesauri by such set-theoretic operations as sum, intersection, and complement of sets. For example, thesaurus of some domain can be formed as a sum of thesauri of IRs pertinent to this domain. The weight of term for set sum operation is defined as a sum of its weight in every IR with the importance of IR s_{IR_j}: $\forall p \in T = \bigcup_{j=1}^{m} T_{IR_j} \exists w(p) = \sum_{j=1}^{m} w_{IR_j}(p) \cdot s_{IR_j}$. If user has to create thesaurus for some subset o domain then operations of set intersection and complement are used.

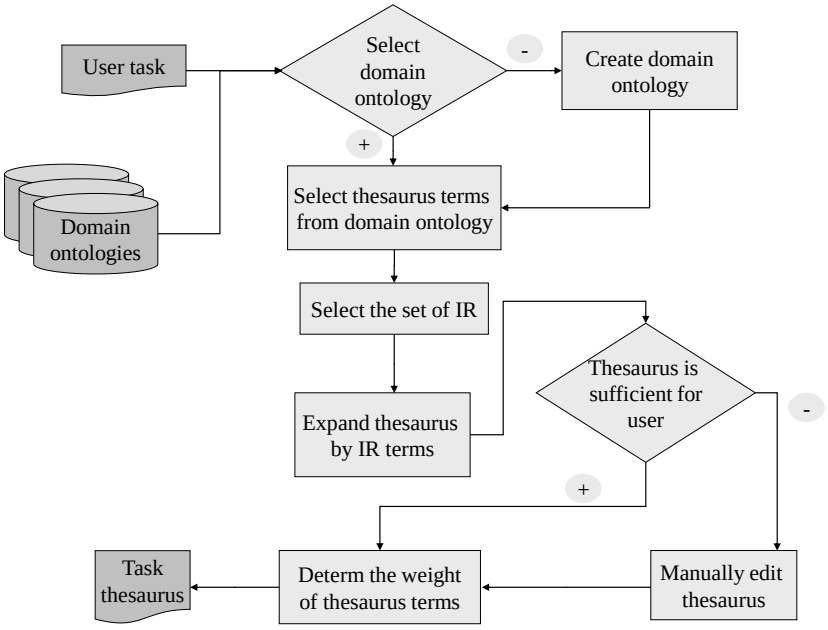

Figure 4 – Generalized algorithm of task thesaurus generation

The theoretic basis of ontology-based thesaurus generation is semantic similarity estimations. Semantically similar concepts (SSC) are a subset of the domain concepts that can be joined by some relations or properties. If domain is modeled by ontology then SSC is a subset of the domain ontology concepts. There are several ways to build SSC which can be used separately or together. The user can define SSC 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.

SSC 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 SSC concept has a weight (positive or negative) which determines the degree of semantic similarity of the concept with the initial set of concepts. The work [22], [23] are classified methods of semantic similarity measuring and their software realizations. Methods are grouped by parameters used in estimations and differ within the groups by calculation of these parameters.

For example, ontology is considered as a directed graph where concepts are interconnected by universal and domain-specific relations, mainly taxonomic (is-a). The simplest way to estimate SS between concepts is to calculate the minimum path length that connects the corresponding ontological nodes using "is-a" relation. The longer path between concepts means the major semantic distance between them. If we define a path $path(c_1, c_2) = l_1, ..., l_k$ as a set of links that connect the concepts c_1 and c_2 where $| path(c_1, c_2) | = k$ is the length of this path, then by analysis of all possible paths between and we can define the semantic distance between them as the minimum value of this length:

$$SS_{Rada} = \min | path(c_1, c_2) | \quad (2)$$

Despite the simplicity of such estimation, the assumption that different edges of the ontological graph reflect the same semantic distances which do not always correspond to domain causes many problems.

Other estimations are based on the analysis of the path between concepts and their depth in the hierarchy. For example, Wu and Palmer [24], [25] define the SS estimation between the concepts as follows:

$$SS_{WP} = \frac{2H}{N_1 + N_2 - 2H} \quad (3)$$

where N_1 and N_2 are the number of "is a" relations between c_1 and c_2 respectively to the lowest common generic object c;

H is the number of "is a" connections between c and the taxonomy root.

Measures of similarity based on information content [26], [27] determine the similarity of two concepts is defined as the information content of their lowest common generic object:

$$SS_{re} = IS(LCS(c_1, c_2)) \quad (4)$$

SS estimation parameters from various approaches (for example, from (2) - (4)) 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.

Conclusions. Unstructured and large amounts of information resources, complex hierarchical structure of knowledge of the IS domain cause the need to apply ontological analysis to the processing of Big Data related to information security. Therefore, the application of big data analysis methods to the construction of ontologies in the domain of IS is justified and appropriate. The task thesaurus was proposed as a dynamic element of model that is based on domain ontology that represents more stabile aspects of user interests. Simple structure of task thesaurus provides it's fast and efficient processing, and use of domain ontologies for their generation causes to avoid loss of important information. Semantic similarity estimations provide the theoretical basis for generation of task thesaurus as a set of concepts similar to user current task. The similarity is an important and fundamental concept in many fields.

The prospects of automates generation of ontology-based task thesauri depend on accessibility of pertinent domain ontologies and well-structured, trusted, and actual IRs that characterize user information needs and interests. Therefore, we can find information resources where such parameters are defined explicitly and can be processed without additional pre-processing. Semantic Wiki the where the relationship between concepts and their characteristics are defined through semantic properties correspond with such conditions.

REFERENCE

[1] S. Grimes, "Unstructured Data and the 80 Percent Rule", *Clarabridge, Bridgepoints*, 2008. [Online]. Available: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/. Accessed on: Aug 1, 2020.

[2] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall, Crawfordsville USA: ServiceTech Press, 2016.

[3] O. Savas, J. Deng, *Big Data Analytics in Cybersecurity*. New York, USA: CRC Press, 2018.

[4] L. Obrst, P. Chase, and R. Markeloff, "Developing an Ontology of the Cyber Security Domain", *In Proc. 7th Inter. Conf. on Semantic Technologies for Intelligence, Defense, and Security*, Fairfax, 2012, pp. 49-56.

[5] Z. Syed, A. Padia, T. Finin, L. Mathews, and A. Joshi, "UCO: A unified IS ontology", *in Proc. AAAI Conf. Artificial Intelligence for Cyber Security*, Phoenix, 2016, pp. 1-8.

[6] P. Bhandari, and M. S. Guiral, "Ontology Based Approach for Perception of Network Security State", *in Proc.of Recent Advances in Engineering and Computational Sciences*, Chandigarh, 2014, pp. 1-6.

[7] I. V. Diorditsa, "Representation of IS policy terminology in the texts of legal acts of Ukraine", *Scientific herald of the International Humanities University. Jurisprudence*, vol 1, no. 29, pp. 64-67, 2017.

[8] R. van Heerden, L. Leenen, and B. Irwin, "Automated classification of computer network attacks", *in Inter. Conf. on Adaptive Science and Technology*, South Africa, 2013, pp.157-163, doi: 10.1109/ICASTech.2013.6707510.

[9] M. Ushold, and M. Gruninger, "Ontologies: Principles, Methods and Applications", *Knowl. Eng. Rev. CUP*, vol. 11, no. 2, pp. 93-155, 1996, doi: 10.1017/S0269888900007797.

[10] P. Buneman, S. Davidson, M. Fernandez, and D. Suciu, "Adding structure to unstructured data", *in Proc.of Inter. Conf. on Database Theory*, Delphi, 1997, pp. 336-350.

[11] Y. Zhang, J. Ren, J. Liu, C. Xu, H. Guo, and Y. Liu, "A survey on emerging computing paradigms for big data", *CJE*, vol. 26, no. 1, pp.1-12, 2017, doi: 10.1049/cje.2016.11.016.

[12] K. Smith, L. Seligman, and A. Rosenthal, "Big Metadata: The Need for Principled Metadata Management in Big Data Ecosystems", *in Proc. Confe. Data analytics in the Cloud*, Snowbird, 2014, pp. 72-84. [Online]. Available: https://dl.acm.org/doi/10.1145/2627770.2627776, doi: 10.1145/2627770.2627776. Accessed on: Aug. 15, 2020.

[13] T. Takahashi, and Y. Kadobayashi, "Reference ontology for cybersecurity operational information", *The Computer Journal, OUP*, vol. 58, no. 10, pp. 2297-2312, 2015. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8205615, doi: 10.1093/comjnl/bxu101. Accessed on: Aug. 15, 2020.

[14] A. Salahi, and M. Ansarinia, "Predicting Network Attacks Using Ontology-Driven Inference", *IJICTR, IGI Global*, vol. 4, no. 2; pp. 27-35, 2012. [Online]. Available: http://arxiv.org/ftp/arxiv/papers/1304/1304.0913.pdf. Accessed on: Aug. 15, 2019.

[15] A. Oltramari, L. F. Cranor, R. J. Walls, and P. D. McDaniel, "Building an Ontology of Cyber Security", *in Proc. 9th Inter. Conf. on Semantic Technologies for Intelligence, Defense, and Security*, Fairfax, 2014, pp. 54-61, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.3593&rep=rep1&type=pdf. Accessed on: Aug. 15, 2020.

[16] J. A. Wang, and M. Guo, "OVM: An Ontology for Vulnerability Management", *in Proc. 5th Annu. Conf on Cyber Security and Information Intelligence Research*, Knoxville, 2009, pp. 1-4, doi: 10.1145/1558607.1558646.

[17] A. Y. Gladun, O. O, Puchkov, I. Yu. Subach, and K. O. Khala, *English-Ukrainian dictionary of terms on information technology and cybersecurity*. Kiev, Ukraine: NTUU KPI named by Igor Sikorsky, 2018.

[18] Protégé 5.0. [Online]. Available: https://protege.stanford.edu/. Accessed on: Aug. 24, 2020.

[19] Great Ukrainian encyclopedia. [Online]. Available: https://vue.gov.ua/. Accessed on: Aug. 10, 2020.

[20] A. Y. Gladun, and J Rogushina, "Mereological aspects of ontological analysis for thesauri constructing", *JIBS Buildings and Environment, Nova Scien. Publish.*, New York, pp. 301-308, 2010.
[21] A. Y. Gladun, and J. Rogushina, "Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources", *IJISA, MECS Press*, no. 1, pp. 11-20, 2012. [Online]. Available: http://www.mecs-press.org/ijisa/ijisa-v4-n1/IJISA-V4-N1-2.pdf, doi: 10.5815/ijisa.2012.01.02. Accessed on: Aug. 15, 2020.
[22] Y. E. Sachuk, "Professional training of specialists in IS and information protection: thesaurus and ontology", *Problems of engineering and pedagogical education*, no. 59, pp. 35-40, 2018.
[23] J. Rogushina, "Use of Similarity of Wiki Pages as an Instrument of Domain Representation for Semantic Retrieval", *in Proc. Conf. Open Semantic Technologies for Intelligent Systems*, Minsk, 2020, no. 4, pp. 111-116.
[24] Z. Wu, and M. Palmer, "Verbs semantics and lexical selection", *in Proc. 32nd Annu. Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, Stroudsburg, 1994, pp. 133-138, doi: 10.3115/981732.981751.
[25] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *in Proc. 14th Inter. Joint Conf. Artificial Antelligence*, vol. 1, 1995, pp. 448-453. [Online]. Available: https://arxiv.org/pdf/cmp-lg/9511007.pdf. Accessed: Aug. 22, 2020.
[26] A. Gladun, and K. Khala, "Using ontological models for formalized knowledge assessment", *Scient. Jour. Computer Means, Networks and Systems*, no. 27, pp. 67-73, 2019.
[27] S. Pryima, A. Gladun, and J. Rogushina, "Ontological Analysis of Outcomes of Non-formal and Informal Learning for Agro-Advisory System: AdvisOnt", *CCIS, Springer*, vol. 1309, pp. 3-17, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-62015-8_1, doi: 10.1007/978-3-030-62015-8_1.

The article was received 02.09.2020.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] S. Grimes, "Unstructured Data and the 80 Percent Rule", *Clarabridge, Bridgepoints*, 2008. [Online]. Available: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/. Accessed on: Aug. 1, 2020.
[2] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers &* Techniques. Prentice Hall, Crawfordsville USA: ServiceTech Press, 2016.
[3] O. Savas, J. Deng, *Big Data Analytics in Cybersecurity*. New York, USA: CRC Press, 2018.
[4] L. Obrst, P. Chase, and R. Markeloff, "Developing an Ontology of the Cyber Security Domain", In *Proc. 7th Inter. Conf. on Semantic Technologies for Intelligence, Defense, and Security*, Fairfax, 2012, pp. 49-56.
[5] Z. Syed, A. Padia, T. Finin, L. Mathews, and A. Joshi, "UCO: A unified IS ontology", in *Proc. AAAI Conf. Artificial Intelligence for Cyber Security*, Phoenix, 2016, pp. 1-8.
[6] P. Bhandari, and M. S. Guiral, "Ontology Based Approach for Perception of Network Security State", *in Proc.of Recent Advances in Engineering and Computational Sciences*, Chandigarh, 2014, pp. 1-6.
[7] І. В. Діордіца, "Представлення термінології політики ІС у текстах правових актів України", *Науковий вісник Міжнародного гуманітарного університету. Юриспруденція*, том 1, № 29, с. 64-67, 2017.
[8] R. van Heerden, L. Leenen, and B. Irwin, "Automated classification of computer network attacks", *in Inter. Conf. on Adaptive Science and Technology*, South Africa, 2013, pp.157-163, doi: 10.1109/ICASTech.2013.6707510.
[9] M. Ushold, and M. Gruninger, "Ontologies: Principles, Methods and Applications", *Knowl. Eng. Rev. CUP*, vol. 11, no. 2, pp. 93-155, 1996, doi: 10.1017/S0269888900007797.
[10] P. Buneman, S. Davidson, M. Fernandez, and D. Suciu, "Adding structure to unstructured data", *in Proc.of Inter. Conf. on Database Theory*, Delphi, 1997, pp. 336-350.

[11] Y. Zhang, J. Ren, J. Liu, C. Xu, H. Guo, and Y. Liu, "A survey on emerging computing paradigms for big data", *CJE*, vol. 26, no. 1, pp.1-12, 2017, doi: 10.1049/cje.2016.11.016.

[12] K. Smith, L. Seligman, and A. Rosenthal, "Big Metadata: The Need for Principled Metadata Management in Big Data Ecosystems", *in Proc. Confe. Data analytics in the Cloud*, Snowbird, 2014, pp. 72-84. [Online]. Available: https://dl.acm.org/doi/10.1145/2627770.2627776, doi: 10.1145/2627770.2627776. Accessed on: Aug. 15, 2020.

[13] T. Takahashi, and Y. Kadobayashi, "Reference ontology for cybersecurity operational information", *The Computer Journal, OUP*, vol. 58, no. 10, pp. 2297-2312, 2015. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8205615, doi: 10.1093/comjnl/bxu101. Accessed on: Aug. 15, 2020.

[14] A. Salahi, and M. Ansarinia, "Predicting Network Attacks Using Ontology-Driven Inference", *IJICTR, IGI Global*, vol. 4, no. 2; pp. 27-35, 2012. [Online]. Available: http://arxiv.org/ftp/arxiv/papers/1304/1304.0913.pdf. Accessed on: Aug. 15, 2020.

[15] A. Oltramari, L. F. Cranor, R. J. Walls, and P. D. McDaniel, "Building an Ontology of Cyber Security", *in Proc. 9th Inter. Conf. on Semantic Technologies for Intelligence, Defense, and Security*, Fairfax, 2014, pp. 54-61, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.664.3593&rep=rep1&type=pdf. Accessed on: Aug. 15, 2020.

[16] J. A. Wang, and M. Guo, "OVM: An Ontology for Vulnerability Management", *in Proc. 5th Annu. Conf on Cyber Security and Information Intelligence Research*, Knoxville, 2009, pp. 1-4, doi: 10.1145/1558607.1558646.

[17] А. Я. Гладун, О. О., Пучков, І. Ю. Субач, та К.О. Хала, *Англо-український словник термінів з інформаційних технологій та кібербезпеки*. Київ, Україна: НТУУ КПІ імені Ігоря Сікорського, 2018.

[18] Protégé 5.0. [Online]. Available: https://protege.stanford.edu/. Accessed on: Aug. 24, 2020.

[19] Велика українська енциклопедія. [Електронний ресурс]. Доступно: https://vue.gov.ua/. Дата звернення: Серп. 10, 2020.

[20] A. Y. Gladun, and J Rogushina, "Mereological aspects of ontological analysis for thesauri constructing", *JIBS Buildings and Environment, Nova Scien. Publish.*, New York, pp. 301-308, 2010.

[21] A. Y. Gladun, and J. Rogushina, "Use of Semantic Web Technologies and Multilinguistic Thesauri for Knowledge-Based Access to Biomedical Resources", *IJISA, MECS Press*, no. 1, pp. 11-20, 2012. [Online]. Available: http://www.mecs-press.org/ijisa/ijisa-v4-n1/IJISA-V4-N1-2.pdf, doi: 10.5815/ijisa.2012.01.02. Accessed on: Aug. 15, 2020.

[22] Ю. Є. Сачук, "Професійна підготовка фахівців із кібербезпеки та захисту інформації: тезаурус та онтологія", *Проблеми інженерно-педагогічно ї освіти*, № 59, с. 35-40, 2018.

[23] Ю. Рогушина, "Использование подобия Wiki-страниц в качестве инструмента представления предметной области для семантического поиска", *на конф. Открытые семантические технологии для интеллектуальных систем*, Минск, № 4, с. 111-116, 2020.

[24] Z. Wu, and M. Palmer, "Verbs semantics and lexical selection", *in Proc. 32nd Annu. Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, Stroudsburg, 1994, pp. 133-138, doi: 10.3115/981732.981751.

[25] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *in Proc. 14th Inter. Joint Conf. Artificial Antelligence*, vol. 1, 1995, pp. 448-453. [Online]. Available: https://arxiv.org/pdf/cmp-lg/9511007.pdf. Accessed: Aug. 22, 2020.

[26] А. Гладун, та К. Хала, "Використання онтологічних моделей для формалізованої оцінки знань", *Зб. наук. пр. Комп'ютерні засоби, мережі та системи,* № 27, с. 67-73, 2019.

[27] S. Pryima, A. Gladun, and J. Rogushina, "Ontological Analysis of Outcomes of Non-formal and Informal Learning for Agro-Advisory System: AdvisOnt", *CCIS, Springer*, vol. 1309, pp. 3-17, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-62015-8_1, doi: 10.1007/978-3-030-62015-8_1.

АНАТОЛІЙ ГЛАДУН,
КАТЕРИНА ХАЛА,
ІГОР СУБАЧ

ОНТОЛОГІЧНИЙ ПІДХІД ДО АНАЛІТИКИ ВЕЛИКИХ ДАНИХ У ДОМЕНІ КІБЕРБЕЗПЕКИ

Інформаційна безпека – динамічна сфера у якій швидко змінюються і удосконалюються як методи і засоби захисту від загроз, так і їх деструктивна складова, що є викликом для користувачів, організацій і всього суспільства в цілому. Тому інформаційні системи, пов'язані зі забезпеченням кібербезпеки потребують постійного надходження знань як із внутрішніх, так із зовнішніх джерел, обсяг яких постійно зростає. Введення наборів великих даних у сферу забезпечення кібербезпеки відкриває можливості застосуваня для аналізу джерел, що містять як структуровані, так і неструктуровані дані. Застосування семантичних технологій до пошуку, відбору зовнішніх великих даних та опису знань про домен кібербезпеки потребує нових підходів методів та алгоритмів аналітики великих даних. Для вибору релевантних даних пропонується семантичний аналіз метаданих, які супроводжують великі дані та побудова онтологій, які формалізують знання про метадані, про кібербезпеку та про задачу, яка потребує вирішення, тобто для ефективного вирішення задачі структурування даних під час аналізу великих даних. Основними перевагами онтологій є здатність здійснювати семантичний пошук, надання загального спільного словника та обмін знаннями в області, а також сприяння семантичній інтеграції та взаємодії між різнорідними джерелами знань. В якості інструментів аналізу пропонується створення тезаурусу задачі на основі онтології домена, який має забезпечити термінологічну базу для інтеграції онтологій різних рівнів. Домен кібербезпеки має ієрархічну структуру, тому і подання формалізованих знань про нього потребує розроблення ієрархії онтологій починаючи від верхнього рівня до нижнього. Для побудови тезаурусу задачі запропоновано використати алгоритм, що дозволить об'єднати інформацію із стандартів інформаційної безпеки, відкритих природомовних інформаційних ресурсів, словників та енциклопедій. Для поповнення семантичних моделей домену кібербезпеки запропоновано використати семантично розмічені Wiki-ресурси, зовнішні тезауруси та онтології.

Ключові слова: аналітика великих даних, кібербезпека, онтологія, тезаурус, неструктуровані дані, метадані, wiki-технологія, семантична подібність.

Gladun Anatoly, candidate of technical sciences, associate professor at the cybersecurity and application of information systems and technologies academic department, Institute of special communication and information protection National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

ORCID: 0000-0002-4133-8169.

E-mail: glanat@yahoo.com.

Khala Katerina, researcher, International research and training center for information technologies and systems under National Academy of Sciences and Ministry of Education and Science of Ukraine, Kyiv, Ukraine.

ORCID: 0000-0002-9477-970X.

E-mail: cecerongreat@ukr.net.

Subach Ihor, doctor of technical science, associate professor, head at the cybersecurity and application of information systems and technologies academic department, Institute of special communication and information protection National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

ORCID: 0000-0002-9344-713X.

E-mail: igor_subach@ukr.net.

Гладун Анатолій Ясоновий, кандидат технічних наук, доцент кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна.

Хала Катерина Олександрівна, науковий співробітник, Міжнародний науково-навчальний центр інформаційних технологій та систем Національної академії наук України та Міністерства освіти та науки України, Київ, Україна.

Субач Ігор Юрійович, доктор технічних наук, доцент, завідувач кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", Київ, Україна.