Method of Detecting Cybersecurity Objects Based on OSINT Technology

Dmytro Lande[1,2], Olexander Puchkov[1] and Ihor Subach[1]

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prosp. Peremohy, Kyiv, 03056, Ukraine
[2]Institute for Information Recording of the National Academy of Sciences of Ukraine, 2, Mykoly Shpaka Str., Kyiv, 03013, Ukraine

Abstract

The information resources of the Internet contain a lot of hidden knowledge. This knowledge is contributed by users forming a kind of expert environment. In this regard, the main task of open source intelligence technologies (OSINT) is identification and extraction of hidden expert knowledge, their generalization, as well as further analytical processing. To achieve this purpose, methods of in-depth data analysis (Text Mining), linguistic and statistical methods, as well as methods of cluster analysis are used. The paper suggests a method of extracting concepts from the texts of messages of network sources related to the subject area of cybersecurity. These concepts are filtered according to statistical characteristics and ranking. A network of their relationships is created, clustered and visualized. To create a software implementation of the suggested approaches, the Perl programming language is used in the Linux OS environment, as well as software tools for graph modeling, analysis, and visualization - Gephi.

Keywords

OSINT, cybersecurity objects, time series, concept extraction, terms network, web resources

1. Introduction

Specialists working in a specific subject area usually know its main concepts and objects. However, with the passage of time, new concepts and new objects emerge. In the field of cybersecurity, various types of cyberattacks, hacker groups, destructive software, analytical groups, etc., can become such objects. New meaningful connections between such objects may appear, previous ones may disappear, which also requires additional analysis. In certain groups of objects, for example, criminal hacker groups, the centers and objects of special attention of cybersecurity specialists may shift. Thus, there is a task of constant information monitoring within the defined subject area.

Such information is widely available in social networks, on forums, the Internet (particularly, documents posted on websites), to the content of which OSINT can be applied [1, 2]. OSINT is one of the directions of intelligence, the essence of which is the search and analysis of information obtained from open sources, collection of information and its further analysis, formation of reports concerning the object of surveillance [3]. The role of OSINT in ensuring cybersecurity is determined by a number of aspects, including availability and efficiency, volume, quality, reliability, ease of further use, cost of obtaining, etc.

The process of OSINT planning and preparation is influenced by such factors as effective information support – information about the objects of information and cyberattacks is obtained from open sources. The availability, depth and scope of publicly available information allow us to find the

XXII International Scientific and Practical Conference "Information Technologies and Security (ITS-2022)", November 16, 2022, Kyiv, Ukraine
EMAIL: dwlande@gmail.com (D. Lande); iszzi@iszzi.kpi.ua (O.Puchkov); igor_subach@ukr.net (I. Subach)
ORCID: 0000-0003-3945-1178 (D. Lande); 0000-0002-8585-1044(O.Puchkov); 0000-0002-9344-713X(I. Subach)

necessary information without involvement of specialized means of intelligence, unnecessary technical and human methods of conducting intelligence. The possibility of massive monitoring of open sources of information in order to find targeted content, people and events leads to the necessity to use Big Data technologies, which are successfully developing nowadays. In addition, a sharp reduction in access time is achieved. As experience shows, competently collected pieces of information from open sources in total can be equivalent or even more significant than professional intelligence reports.

The objective of this work is to create and test the method for determining the main cybersecurity objects and connections between them based on the analysis of the meaningful component of the web-space, as well as formation, clustering and analytical processing of the formed networks of cybersecurity objects, analysis of the objects dynamics in the subject area. To achieve this goal, a number of tasks are solved, in particular, the targeted information collection, its processing, extracting the necessary entities from it, establishing connections between them, that is, forming a network, cluster analysis of objects network, identifying the centers of these clusters, etc.

2. Method description

A feature of this technique is the simplicity of its implementation when using a typical information retrieval system and a system for analyzing and visualizing graph structures.

A method is proposed for consideration, the essence of which is to perform such technological operations as expert creation of queries to existing information retrieval systems corresponding to the subject area. As a result of query processing, large arrays of relevant documents are created. Named entities (objects) belonging to different periods of time are extracted from the selected arrays. In the future, through the analysis of networks, the interconnections of objects are studied, individual clusters are determined.

Fig. 1 shows the main stages (chains) of the method including 1) obtaining information; 2) extraction of concepts – cybersecurity objects; 3) filtering concepts with the involvement of experts (or artificial intelligence tools); 4) formation of a cybersecurity objects network; 5) analysis (including clustering) and visualization of this networks; 6) visualization of the dynamics of the appearance of concepts in time.

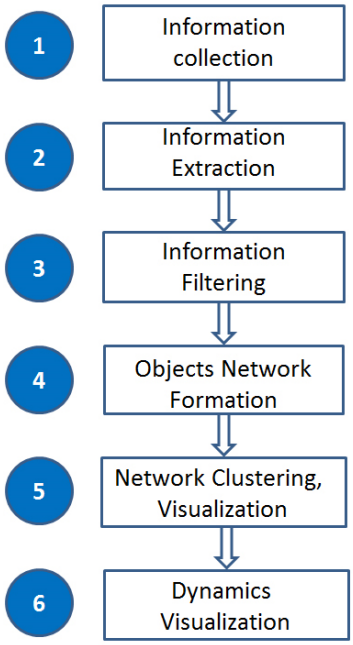

Figure 1: The main stages of detecting cybersecurity objects

The method, offered in this paper for the selection of named entities corresponding to cybersecurity objects, identification of connections and the study of the dynamics and identified named entities in information flows, involves the implementation of a number of stages.

2.1. Information collection

At the first stage of the information extraction method, an information array of documents relevant to the topic is formed. For this purpose, existing information and search systems, both public and corporate content monitoring systems, such as the Cyber Aggregator system [1], should be used.

It should be noted that the Cyber Aggregator system collects news from 12 social networks and provides users' access to it in search mode. It is also possible to download relevant information in RSS format [8].

Like most similar systems for aggregating information from social networks, the CyberAgregator system consists of three main parts: a server for collecting and primary processing of information, an information retrieval server (search engine) and an interface server from which the service is provided to users and other systems through the API .

Aggregation of information from social networks includes the following steps:

1) search for messages from social networks related to a common broad topic – the formation of an information flow from thematic messages;

2) determining the language of individual messages downloaded from social networks;

3)extracts from information messages, such concepts as keywords, persons, companies, geographical names, etc;

4) sentiment analysis of individual messages;

5) data formatting, conversion to standard formats (XML, JSON);

6) loading the received stream into full-text databases.

The CyberAgregator system provides the user with a web interface from which he can access the functions of information search and analysis.

The system user receives documents upon request both in the retrospective database (Search) and in the current information (Current), as well as for data analysis (Analysis).

As a result of a query search (Fig. 2), the user is provided with a list of relevant message titles with links to the full texts of these messages in the system, as well as to these messages in social networks.

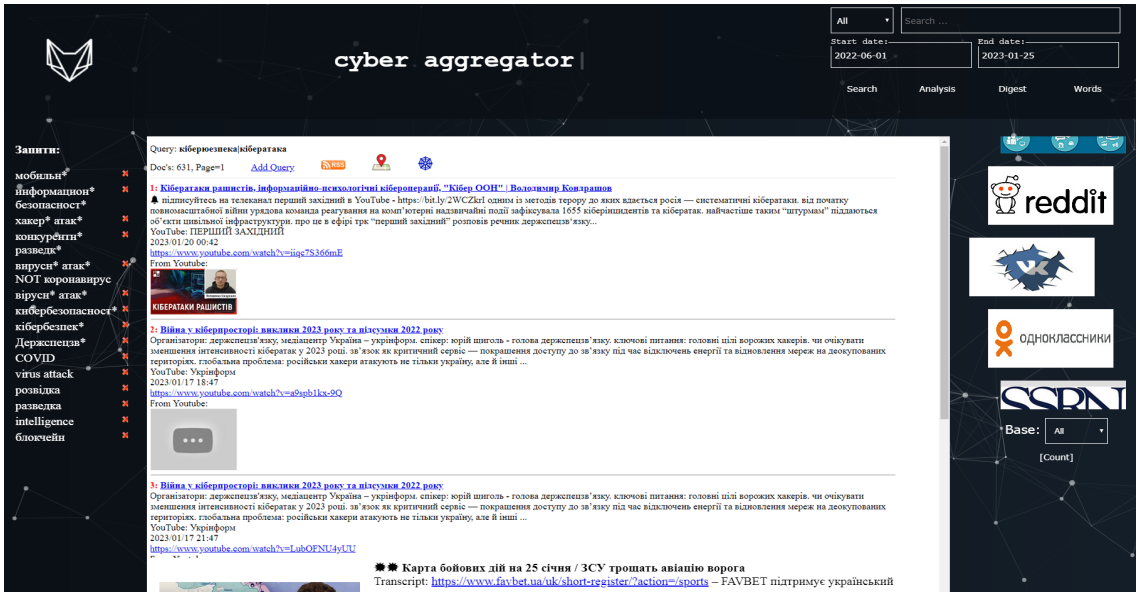

Figure 1:A fragment of the user interface in search mode

To obtain an information array of publications on cybersecurity, it is necessary to determine the necessary period of processing a thematic request to such a system, for example, a request for information selection was used:

"cybersecurity | cyberattack"

To simplify the extraction of named entities, requests in Cyrillic are used within the framework of the suggested method.

If a user finds documents that are relevant to their search query, they can save the query for future use by selecting the 'Add Request' command.

You can later display the found messages in RSS format (with subsequent loading of these results into the so-called RSS aggregators on an ongoing basis), as well as display search results with details on a geographical map, scalable both automatically and through settings.

As a result of processing such a request, an array of relevant documents is obtained, which is subjected to further processing.

2.2. Information extraction

At the second stage, on the basis of linguistic and statistical analysis, concepts from the subject area contained in the documents of the information array obtained at the first step are extracted. The main idea of recognizing named entities – cybersecurity objects is that nowadays most new concepts in Cyrillic messages are denoted by Latin letters (non-Cyrillic short phrases in the information array are taken into account), or by Cyrillic letters but in quotation marks.

The peculiarity of the given method is the simplicity of its implementation when using a typical information search system and a system of graph structures analysis and visualization.

Usually[4, 5], the detection of named entities (Named-entity recognition, NER) is carried out with the help of special software libraries (spaCy, Flair, FastText), the common disadvantages of which are the low speed of extracting concepts and the need for a complex stage of system training (it is known that names of cybersecurity objects are not always typical company or brand names).

The use of network information not in Latin encoding (Ukrainian, Russian, Chinese, or other languages) greatly simplifies the task of extracting cybersecurity objects, such as hacker groups, names of analytical centers, etc., which are mostly written in Latin encoding.

In particular, the spaCy library is interesting in that several pre-trained models are available in about 20 languages [6]. This means that in many cases it is not necessary to train your model to extract entities. The spaCy library is considered a "production class" framework because it is very fast, reliable, and comes with comprehensive documentation. Another popular Python entity detection framework is the Flair library [7], which is based on the PyTorch deep learning framework. It is gaining a lot of popularity as it achieves higher precision in many languages compared to spaCy. However, the increase in accuracy comes at the cost of speed reduction.

Within the framework of this work, the application of a new heuristic approach is proposed. The main idea of recognizing nominal entities – cybersecurity objects – is that, at present, most of the nominal entities of cybersecurity objects, such as hacker groups, names of analytical centers, etc., mentionedin messages from social networks are not in Latin coding(Ukrainian, Russian, Chinese, etc.);they are mainly indicated inLatin (non-Cyrillic short phrases in the information array are takeninto account), or in Cyrillic letters, but in quotation marks. This greatly simplifies the extraction task. In these cases, it is sufficient to detect short words or phrases in Latin encoding or in quotation marks. Obviously, the technical solution of such a problem does not require large resource and time costs (rather spaCy), including special machine learning.

At the same time, a dictionary of known named entities of cybersecurity objects, which are searched for in the information array, is also used to extract already known named entities.

2.3. Information filtering

At the third stage, the selected concepts are sorted by frequency and filtered by an expert specialist. Usually, the number of possible cybersecurity objects detected by this method does not exceed several thousand, that is why this operation does not take much time.

2.4. Objects network formation

At the fourth stage, a network of selected concepts is formed [9]. For this, undirected connections between concepts are defined. Connections can be established on the basis of different approaches, in particular, two concepts can be considered connected if they are included in the same segment of the document (sentence, paragraph, circle of N words, or the entire document) from the selected information array. Also, connections can be calculated as mutual correlations between time series of frequencies of occurrence of individual named entities per day.

A method is proposed that associates a nominal entity (a concept from the subject area of cybersecurity) with a dynamics vector corresponding to the distribution of documents by dates (days). More specifically, each day is assigned a number – the number of occurrences of the concept in publications covered by the content monitoring system. The dimension of this vector corresponds to the number of days, the length of the time interval during which the array of network publications [10] was analyzed. An example of time series corresponding to entities (names of criminal cybergroups is shown in Fig. 3a, 3b).

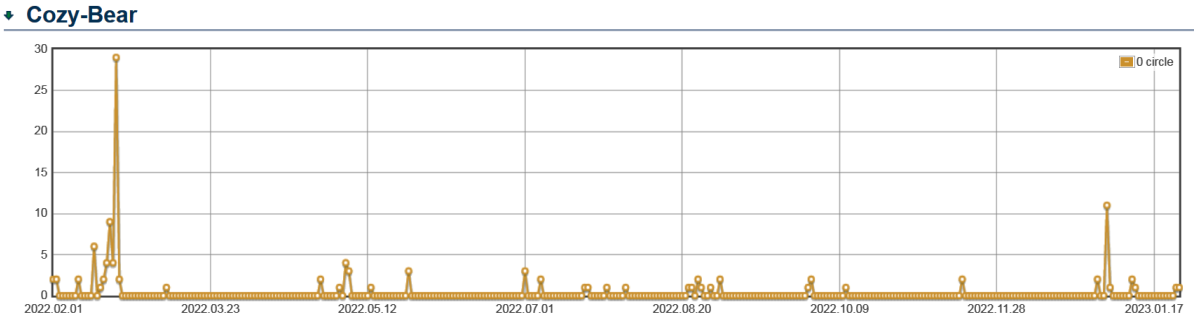

Figure 3a: Graph of the number of messages per day. Request Cozy Bear

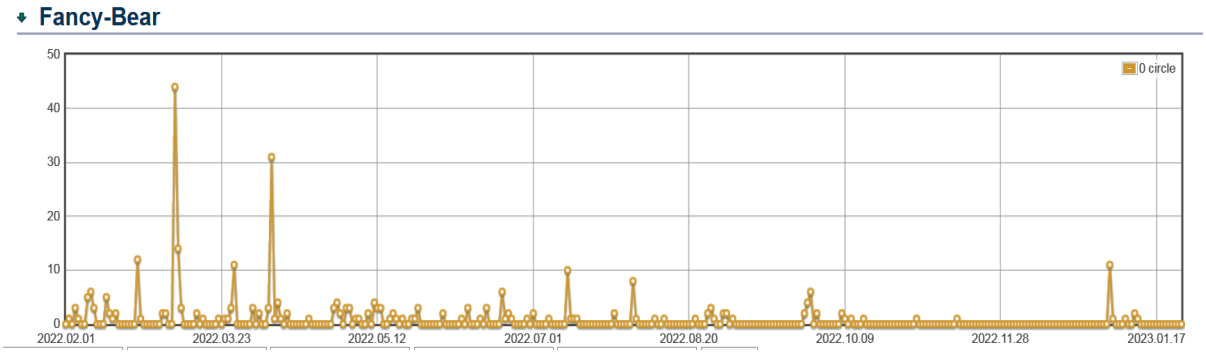

Figure 3b: Graph of the number of messages per day. Request Fancy Bear

To form a correlation network, a number of steps are performed, namely:1) for each entity, a request is generated to the content monitoring service (in our case, to the Cyber Aggregator system). The analysis period is also determined - the dimension of the corresponding time series - dynamics vectors; 2) as a result of query execution, a set of dynamics vectors corresponding to the given

nominal entities (concepts) is determined; 3) the set of maximum cross-correlations between the obtained vectors is calculated, the corresponding correlation matrix is formed with elements:

$$a_{ij}(m) = \max_{m} \frac{\sum_{k=1}^{n-m} w_{k+m}^{i} w_{k}^{j}}{\sqrt{\sum_{k=m+1}^{n} \left(w_{k}^{i}\right)^{2}} \sqrt{\sum_{k=1}^{n-m} \left(w_{k}^{j}\right)^{2}}}. \quad (1)$$

Each entity s_k from the set $S = \{s_k\}_{k=1}^{|S|}$ is assigned a vector parameter value $\overline{w^k} = \left(w_1^k, w_2^k, ..., w_n^k\right)$, where $n = |G|$ is the number of elements in the parameter set. The max function is used for the reasons that processes that are similar in nature can have behavior that is close in dynamics, but possibly with a time shift; 4)the adjacency matrix is formed in accordance with formula (1) and this matrix is saved in a file in CSV format. Due to the fact that there are links between all nodes in the adjacency table, the links are ignored, the value of which is less than some selected threshold. The choice of this threshold depends entirely on the experience of the analysts.

Compared to existing approaches, the method proposed in this paper has several advantages. First, it uses intuitive rules to determine the weight of nodes and links, which closely reflect real-world dynamics. Second, it has a reliable mathematical basis for correlation analysis. Third, it takes into account previously unused parameters, such as time series of the dynamics of publications, to group entities according to their development trends over time. Finally, the method is objective and relatively easy to implement.

2.5. Network Clustering, Visualization

At the fifth stage, clustering of the selected network is carried out and objects - centers of clusters are found according to the modularity algorithm, as well as visualization of the formed network using the Gephi system [11, 12] (Fig. 4).

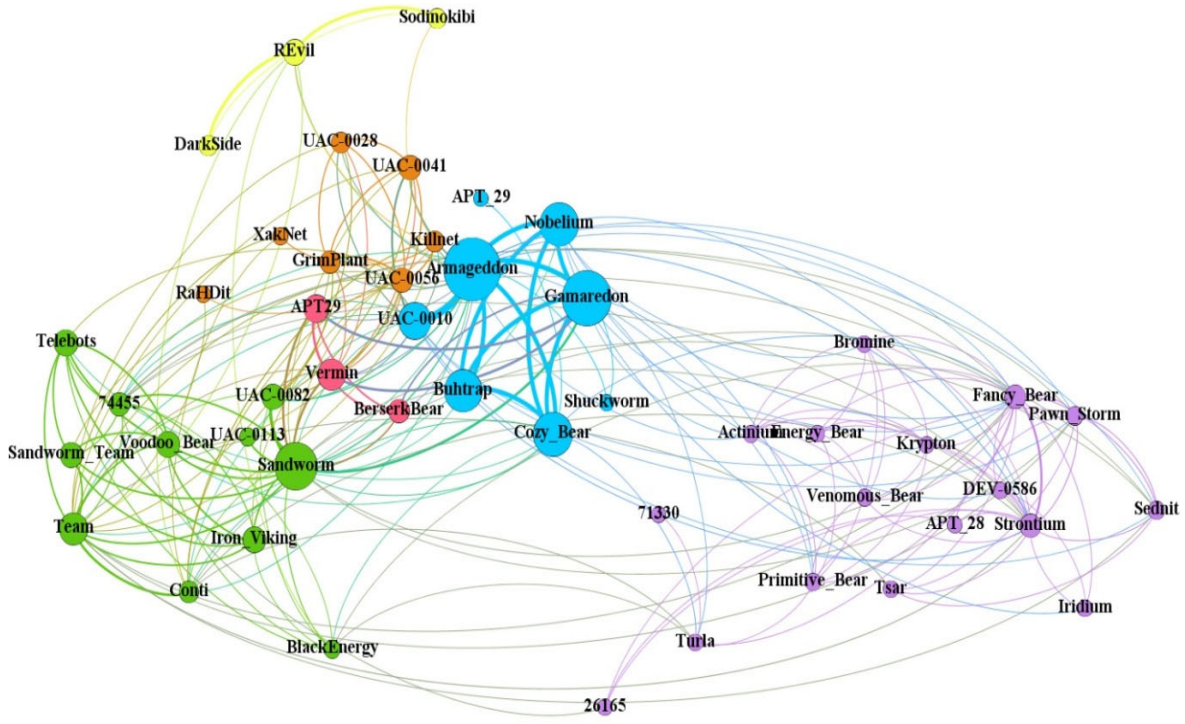

Figure 4: A fragment of the cybersecurity objects network

On a Fig. 4 entities related to the Main Directorate of the General Staff of the Armed Forces of the Russian Federation (GRU) military units 21165, 71330 and 74455, respectively, are marked in blue and green; basilica and orange - subjects of cybersecurity related to the Federal Security Service of the Russian Federation (FSB); purple - cybersecurity entities related to the Foreign Intelligence Serviceof the Russian Federation (SVR).

Gephi [12, 13] is currently the most popular program for visualization and analysis of networks and graphs ("network graphs"). Gephi provides fast layout, efficient filtering, and interactive data exploration, and, besides, it is one of the best options for visualizing large-scale networks. The main option for exporting graph data from an external file is to load the initial network data in CSV format, in which the elements are separated by semicolons. For the analysis of large and dense networks (arranging graph nodes) with the Gephi system, efficient layout modules such as Yifan-Hu, Force-directed are supplied. In particular, the Yifan-Hu algorithm is an ideal option for application after other, faster and coarser algorithms. Most of the methods suggested by Gephi can be performed within a reasonable time; a combination of, for example, OpenOrd and Yifan-Hu gives the highest quality visuals.

Table 1.
Fragment of the list of cybersecurity objects

Cybersecurity object	Node degree	Modularity class (cluster)
REvil	7	0
Sodinokibi	3	0
DarkSide	2	0
Armageddon	26	1
UAC-0010	16	1
Gamaredon	13	1
Cozy_Bear	12	1
Nobelium	10	1
Buhtrap	5	1
Shuckworm	2	1
APT_29	1	1
Sandworm	25	2
Conti	12	2
UAC-0082	11	2
BlackEnergy	10	2
Telebots	9	2
Voodoo_Bear	9	2
Iron_Viking	8	2
Sandworm_Team	7	2
UAC-0113	2	2
Fancy_Bear	23	3
Strontium	21	3
Pawn_Storm	11	3
Primitive_Bear	10	3
Sednit	9	3
Actinium	9	3
Bromine	8	3
Turla	7	3
Energy_Bear	7	3

Table 1 shows a fragment of the list of cybersecurity objects in the example under consideration with an indication of the modularity class (cluster number).

For further clustering within the framework of the Gephi system, the modularity of individual nodes – named entities is calculated, on the basis of which groups in networks (clusters) are identified.

Modularity was introduced to measure the degree of division of the network into modules (clusters, cliques). This parameter is calculated as the difference between the fraction of edges within a cluster in the considered network and the expected fraction of edges within a cluster in a network in which the vertices have the same degree as in the original one, but the edges are randomly distributed.

The concept of adjacency matrix is used to calculate modularity.

The adjacency matrix A consists of elements A_{vw}. whose values are equal to 0, if the node *v* is not connected to the node *w*, and the weight of the connection between *v* and *w*, if these nodes are connected to each other.

The modularity of the network can be expressed by the formula:

$$Q = \frac{1}{2m}\sum_{v,w}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\ \delta(c_v, c_w), \tag{2}$$

where A_{vw} is the element of the adjacency matrix A, m is the number of edges in the graph, k_v, k_w is the degree of nodes v and w respectively, δ – is Kronecker's delta (indicates whether nodes v and w are in the same module).

The results of the network analysis in the given example indicate the affiliation of the considered hacker groups to the special services of the Russian Federation, namely: the GRU, the SVR and the FSB. Currently, some of the most well known Russian-affiliated hacker groups include Fancy Bear, Cozy Bear, Turla, Sandworm, and Berserk Bear.

Based on the selected information arrays, named entities from the cybersecurity field related to different time periods are extracted.

If we consider a set of documents related to a certain topic as:

$$D = \{d_1, d_2, \ldots, d_n\},$$

where the indexes of the documents $i = 1,\ldots,n$ correspond to time (in particular, days).

Let us denote the set of named entities $F = \{f_1, f_2, \ldots, f_m\}$, where the numbers of the entities and their indices are $i = 1,\ldots,m$.

We denote the entity extraction function for day i as:

$$Ex(d_i) = \{f_1^i, f_2^i, \ldots, f_m^i\},$$

where f_j^i is the frequency of mention of entity *j* on day *i*.

2.6. Dynamics Visualization

In order to visualize the dynamics of the appearance of a set of nominal entities $\{f_1^i, f_2^i, \ldots, f_m^i\}$, within *n* days, a special form of visual display of entities in a section, a phraseology diagram (Ph-Di)is offered. The cells of this diagram are filled with numerical values corresponding to the f_j^i – frequency of appearance of named entities in relation to the dates of their appearance. That is, the columns of this table correspond to dates, while the rows correspond to named entities, which can be used as a kind of meaningful information flow filter.

In fact, the diagram is a two-dimensional projection of a set of time series of the dynamics of the relevant information flows, similar to those shown in Fig. 3.

The proposed Ph-Di diagram is presented as a table, with cells colored in varying shades according to the number of publications on the selected object per day. In this chart, a lighter shade corresponds

to a higher value.The offered schemes for a relatively small number of lines - named entities allow you to visually distinguish groups of similar objects by date and intensity of publication without additional processing.

When constructing a diagram, rows may be rearranged (regrouping of named entities). For further clustering, it is suggested to form a relations network of named entities (connections based on the correlation of time series) and to highlight groups that are most interconnected and distant from each other (identify cliques).

Later on, the dynamics of mentions of these objects is studied; a form of visual representation of the information flow in the section of objects and dates is offered, which is a rectangular table, the cells of which are filled with numerical values corresponding to the frequency of appearance of the objects names in the information flow in the dates section [13]. The considered approach can be used to solve the problems of analysis and visualization of the objects distribution for any selected information arrays in terms of issues that are of interest to the researchers and have a significant time frame.

Fig. 5. provides a Ph-Di diagram for concepts relevant to the subject area of cybersecurity. In the diagram, the vertical dimension corresponds to the subjects of cybersecurity, and the horizontal dimension corresponds to the dates of publications about them. The color of the cells (dot) corresponds to the numerical values of messages per day relative to the corresponding cybersecurity subjects: light shades correspond to larger values, dark shades correspond to smaller ones. Horizontal light risks in the diagram correspond to the periods of activity of the corresponding subject in social networks.

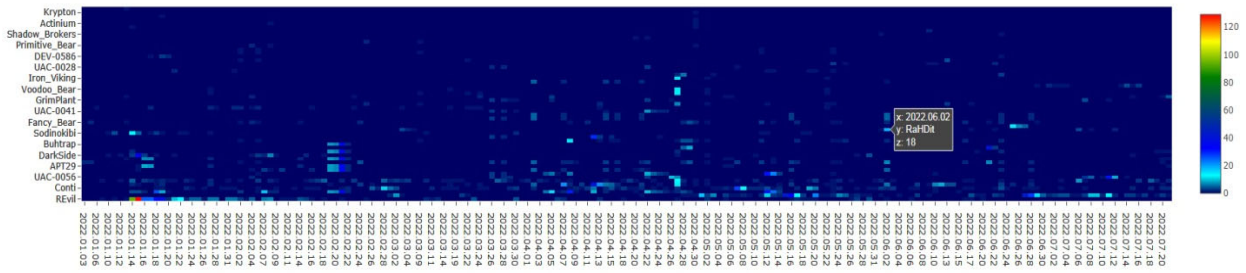

Figure 5: A diagram corresponding to the activity of cybersecurity objects

In practice, the form is implemented as an HTML file using the language. In this diagram, bright horizontal lines (high frequency of individual named entities during a certain period) can clearly tell the user about trends and high activity of individual objects in the information field of the Internet.

3. Conclusions

To sum up, we can state that a method of identifying the named entities of cybersecurity objects from documents, as well as analyzing the relationships and dynamics of objects in the subject area, is suggested. This method takes into account the hidden knowledge contributed by the expert network environment. It is based on the application of documents from the Internet in non-Latin encoding, despite the fact that most cybersecurity objects, such as hacker groups, names of analytical centers, etc., are mostly written in Latin encoding. Taking this fact into account significantly simplifies the task of extracting named entities and speeds up the solution of the problem.

Thus, in order to implement the proposed method: 1) a set of initial requests to existing information and search systems is created; 2) developed software (software) for extracting the necessary fragments from the selected documents; 3) object extraction software based on a heuristic model was developed; 4) the software for the formation of correlation networks of the interconnection of objects, their visualization, and cluster analysis was adapted; 5) Ph-Di visualization software was developed.

The results of content monitoring of the Internet resources and the conducted cluster analysis indicate the affiliation of the considered hacker groups to the special services of the Russian

Federation, namely: the GRU, the SVR and the FSB. Currently, the most famous criminal cyber groups of Russian origin are Fancy Bear, Cozy Bear, Turla, Sandworm, and Berserk Bear.Cluster analysis and visualization of the resulting network of cybersecurity objects and the use of Ph-Di diagrams allow us to visually observe the state and dynamics of the conceptual base development of the cybersecurity subject area.

As a result of the research, it was shown that the use of the Ph-Di visualization tool allows decomposing the original time series by the composition and features of objects, identifying the activity of publications corresponding to certain concepts, determining the links between objects, details of the dynamics of the emergence of new objects in the information streams.This methodology can be based on data obtained from content monitoring systems, which are commonly used for various analytical purposes. The goal is to identify and group entities based on their relationships and dynamics, even if explicit links between them are not present.

The considered approach can be used to analyze and visualize the distribution of objects for any selected arrays of information over a significant period of time based on the interests of the study.

4. References

[1] Tabatabaei, F., Wells, D. (2016). OSINT in the Context of Cyber-Security. In: Akhgar, B., Bayerl, P., Sampson, F. (eds) Open Source Intelligence Investigation. Advanced Sciences and Technologies for Security Applications. Springer, Cham. DOI: 10.1007/978-3-319-47671-1_14

[2] ATP 2-22.9. Army Techniques Publication No. 2-22.9 (FMI 2-22.9). Open-Source Intelligence. Headquarters Department of the Army Washington, DC, 10 July 2012.

[3] Yong-WoonHwang,Im-Yeong Lee, Hwankuk Kim, Hyejung Lee, and Donghyun Kim. Current Status and Security Trend of OSINT. Wireless Communications and Mobile Computing, vol. 2022, Article ID 1290129, 14 pages, 2022. https://doi.org/10.1155/2022/1290129

[4] Komil B. Vora, Avani R. Vasant, Saurabh Shah. (2022). Custom Named Entity Recognition for Gujrati Text Using Spacy. Mathematical Statistician and Engineering Applications, 71(3), 1483–1495. DOI:10.17762/msea.v71i3.502

[5] Sharma, A., Amrita, Chakraborty, S., Kumar, S. (2022). Named Entity Recognition in Natural Language Processing: A Systematic Review. In: Gupta, D., Khanna, A., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Second Doctoral Symposium on Computational Intelligence . Advances in Intelligent Systems and Computing, vol 1374. Springer, Singapore. https://doi.org/10.1007/978-981-16-3346-1_66.

[6] spaCy, URL: https://spacy.io/models

[7] Hugging Face, URL:https://huggingface.co/models?library=flair

[8] P. Ramesh Babu. Measuring Research in RSS Feed Literature: A Scientometric Study. In Measuring and Implementing Altmetrics in Library and Information Science Research. Alliance Broadcast Pvt. Ltd, India. – 13 p. 2020.DOI: 10.4018/978-1-7998-1309-5.ch008.

[9] Lande, D., Dmytrenko, O. Creating Directed Weighted Network of Terms Based on Analysis of Text Corpora. 2020 IEEE 2nd International Conference on System Analysis and Intelligent Computing, SAIC 2020, 2020, 9239182. DOI: 10.1109/SAIC51296.2020.9239182

[10] Social media and depression symptoms: A network perspective.By Aalbers, George,McNally, Richard J.,Heeren, Alexandre,de Wit, Sanne,Fried, Eiko I.Journal of Experimental Psychology: General, Vol 148(8), Aug 2019, 1454-1462. DOI: 10.1037/xge0000528

[11] Cherven K. Mastering Gephi Network Visualization. – Packt Publishing, 2015. – 378 p. ISBN 78-1-78398-734-4.

[12] Gephi, URL: https://gephi.org/

[13] Michael Zgurovsky, Dmitry Lande, KostiantynYefremov, OlehDmytrenko, AndriyBoldak, ArtemSoboliev.Extracting and Identifying Relationships of Key Phrases in Information Flows. Published in: 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC) 04-07 October 2022. DOI: 10.1109/SAIC57818.2022.9923019.