INFORMATION TECHNOLOGY

DOI 10.20535/2411-1031.2021.9.1.247256
UDC 004[942+056.5]

DMYTRO LANDE,
OLEKSANDR PUCHKOV,
IHOR SUBACH

AGGREGATION OF INFORMATION FROM DIVERSE NETWORKS AS THE BASIS FOR TRAINING CYBER SECURITY SPECIALISTS ON PROCESSING ULTRA LARGE DATA SETS

The basic principles of training cybersecurity specialists on processing large data sets to solve complex unstructured tasks in the course of their functional responsibilities based on the achievements of Data Science in the field of cybersecurity, by acquiring the necessary competencies and practical application of the latest information technologies based on methods of aggregation of large amounts of data are substantiated and presented. The most common latest technologies and tools in the field of cybersecurity, the list of which allows getting a fairly holistic view of what is used today by specialists in the field of Data Science, are considered. The tools you need to have to solve complex problems using big data are analyzed. The subject of the study is the fundamental provisions of the concept of “big data”; appropriate data models; architectural concepts of creating information systems for “big data”; big data analytics, as well as the practical application of big data processing results. The theoretical basis of the training, which includes two sections: “Big Data: theoretical principles”, and “Technological applications for big data”, which, in turn, are logically divided into ten, is considered. As a material and technical basis for the acquisition of practical skills by students, a model based on the system “CyberAggregator” was created and described, which operates and is constantly improved in accordance with the expansion of the list of tasks assigned to it. The CyberAggregator system consists of three main parts: a server for collecting and primary processing of information; an information retrieval server (search engine); an interface server from which the service is provided to users and other systems via the API. The system is based on technological components such as the Elasticsearch information retrieval system, the Kibana utility, the Neo4j database graph management system, JavaScript-based results visualization tools (D3.js) and network information scanning modules. The system provides the implementation of such functions as the formation of databases from certain information resources; maintaining full-text databases of information; detection of duplicates similar in content to information messages; full-text search; analysis of text messages, determination of tonality, formation of analytical reports; integration with the geographic information system; data analysis and visualization; research of thematic information flows dynamics; forecasting events based on the analysis of the publications dynamics, etc. The suggested approach allows students to acquire the necessary competencies needed to process effectively large amounts of data from social networks, create systems for monitoring network information on cybersecurity, selection of relevant information from social networks, search engine implementation, analytical research, forecasting.

Keywords: big data, social networks, cybersecurity, information retrieval systems, data aggregation, data science, information technology.

Problem statement. Nowadays, the concept of “Big Data” is playing an increasingly important role in the field of cybersecurity [1]. Of course, the amount of data to be considered in the field of cybersecurity is constantly growing, as well as the amount of information noise, sometimes destructive. Specialists involved in the processing, aggregation of large amounts of data, solving problems caused by their growth, dynamics and variability are currently called “data scientists” (Data Scientists), appropriately, science – Data Science.

Big data is a term that refers to data sets that are so voluminous and complex that it makes it impossible to use existing traditional database management tools and applications to process them. The problem is gathering, purification, storage, retrieval, access, transfer, analysis, and visualization of such sets as a single entity, rather than local fragments. As defining characteristics for big data "three V" are determined: volume (Volume, in terms of the physical volume size); velocity (Velocity that means in this context speed of growth and necessity of high-speed processing and reception of results); variety (Variety, in terms of possibilities of simultaneous processing of different types of structured and semi-structured data).

Based on this, it is important, at this time, to train cybersecurity professionals using the achievements of Data Science, and those who have deep knowledge and practical skills in working with Big Data technologies. It should be noted that the success of this issue directly depends on the material and technical base of training specialists, which allows obtaining the necessary theoretical information and skills to work with relevant information technologies.

Analysis of recent research and publications. The term Big Data first appeared in an editorial by Clifford Lynch, editor of the journal Nature, on September 3, 2008, which devoted an entire special issue of the journal to "what big data sets can mean for modern science". There are currently various publications related to the role of Data Science and the necessity to train professionals in this field. This is, first of all, the work of Bill Franks [1], Davy Silen and Arno Meisman [3], Dodonova O.G. [4] (section on big data analytics). At the same time, there is still no comprehensive approach related to the real practical basis, which is built on modern free software for solving cybersecurity problems.

The formation of the Big Data direction is associated with the development of social networks, despite the fact that large amounts of data are also inherent in such industries as telecommunications, energy, transport, etc. One of the first information technologies in the field of security and defense-related to this area is OSINT (Open Source Intelligence). In [2] it is substantiated that OSINT is an integral part of cybersecurity.

According to [1], [3] we can identify the following basic functional operations on big data:

- aggregation (consolidation) of data;
- classification, clustering;
- machine-learning;
- visualization.

The aim of this paper is to substantiate and present the basic principles of training cybersecurity professionals using technologies for processing ultra-large data sets based on the achievements of Data Science in cybersecurity, by studying the theoretical foundations of this science in close combination with the practical application of new aggregation information technologies of big data.

The main material research. Today to train cybersecurity professionals with competencies related to the processing of ultra-large data sets, the basic, most common technologies and tools in the field of cybersecurity are suggested. The following list does not cover all of the tested technologies, but it does provide a fairly holistic view of what "Data Science" professionals use nowadays and the tools they need to have to solve complex big data problems.

Thus, the subject of the study is the fundamental provisions of the concept of "big data"; appropriate data models; architectural concepts of creating information systems for "big data"; big data analytics, as well as the issue of the practical application of big data processing results.

This approach considers:

- reasons for a new direction of big data and the problems and opportunities associated with the emergence of big data;
- possibilities of technologies of ultra-large data sets analysis for solving problems of enterprises, organizations, or business, as well as possibilities of application of scientific methods, including methods of data mining, to big data;
- features of architectural decisions at creation and development of systems of processing large data sets, and also a choice of technology of big data storage and processing, usage of modern high-performance systems of big data storage and processing;

- basic technologies and tools for working with big data: Hadoop, HDFS, MapReduce, Elastic Stack, Elasticsearch, Kibana, Neo4j;
- software components required to work in distributed information systems for processing large data.

The main directions of solving the problem. The suggested approach to training cybersecurity professionals with competencies to work with Big Data includes two sections: “Big Data: Theoretical Principles” and “Technological Applications for Big Data”, as well as ten topics within these sections, which we will discuss in more details.

The first topic is devoted to the introduction to Big Data, conceptual provisions, issues of aggregation (conceptual), clustering and classification, machine learning, visualization. Within this topic, the definitions and terminology of big data, the role of big data in technology, science, economics, and public life are considered. The characteristics of big data, such as volume, speed, and diversity, are studied. Possible sources of big data (data of social networks; personal data; sensory data; data of monitoring systems; data of transactions; administrative data) are also reflected.

Accordingly, the second topic is devoted to Data Science – modern data science. The basic concepts, areas of application, issues of machine learning in the very large data sets processing are studied. Elements of information technologies, which include the implementation of machine learning algorithms for big data, are also considered. At present, the so-called NoSQL [5] database management systems (DBMS) are used for data management. Features of development of information systems based on NoSQL-solutions on examples of DBMS MongoDB, CouchDB, and Redis are taken into account.

The third and the fourth topics are devoted to methods of classification and cluster analysis of big data. The main definitions of classification and cluster analysis as basic methods of big data mining are presented. The relationship between classification and clustering is examined. In this case, the classification is machine learning with a teacher (Supervised Machine Learning), cluster analysis is machine learning without a teacher (Unsupervised Machine Learning). The mathematical formalization of the classification processes and cluster analysis as optimization problems are provided. Classification algorithms such as the k-nearest neighbor method, linear classifier, DNF method, reference vector method (SVM), and cluster analysis: k-means method, hierarchical aggregation (HAC), matrix latent semantic indexing (LSI) are studied.

The fifth topic deals with such fundamental concepts of Data Science as machine learning (ML) and neural networks, as a means of intellectual analysis of large data sets, methods of machine learning.

The sixth topic is devoted to the concept of complex networks (Complex Networks), which are considered as a special kind of big data. The basic issues of the complex networks concept, separate parameters, and complex networks properties, among which the distribution of degrees of complex networks nodes, clustering, modularity are studied.

The next four topics are the second section of the training, which is devoted to big data technology platforms, including Apache Hadoop technology [6], which is designed to organize the distributed processing of large amounts of data; MapReduce is a technology for distributed parallel processing of large data sets using a large number of computational clusters. The main means of aggregation of large amounts of unstructured data, their search, processing, and visualization, in the framework of the suggested approach, is the ecosystem of components Elastic Stack [7], [8], used for data retrieval and processing. The main components of this stack are analyzed in detail, specifically, Kibana [8], Logstash, Beats, X-Pack, and Elasticsearch. Elasticsearch is an information retrieval system, the core of Elastic Stack, which allows you to process unstructured data, information retrieval, data analysis, provides support for custom libraries and REST API; easy management and scaling. Kibana is a window in Elastic Stack, a visualization tool that implements such types of data display from Elasticsearch as histograms, maps, line graphs, time series.

The tenth topic is devoted to the means of large networks analysis by graphical DBMS. The possibilities of two main systems – software for analysis and visualization of graphs Gephi [11], [12] and graph database management system Neo4j [13] are considered. The features of the Gephi program include a detailed user interface, graph layout capabilities, filtering, data research, visualization, and support for graphical data formats. Graph DBMS Neo4j provides storage and processing network data of large volumes, contains declarative language of inquiries to Cypher graphs.

Practical skills in the application of all the above mentioned information technologies are practiced on the basis of the developed layout ("CyberAggregator" system), which operates as a part of the situational training center for cybersecurity, is constantly evolving and improving.

Model. Practical implementation. During the practical classes, such tasks as creating intelligent information retrieval systems based on Elastic stack technologies, filling them with data collected from web pages and social networks, aggregating this data, creating tools for analytical processing of this data, trend detection, forecasting, etc. are solved. Automated formation of models of subject areas and their visualization is also foreseen.

Like most similar systems for aggregating information from social networks, the system "CyberAggregator" [9], [10] consists of three main parts (servers) (Fig. 1). It is a server for collecting and primary information processing, an information retrieval server (search engine), and an interface server from which the service is provided to users and other systems via the API.

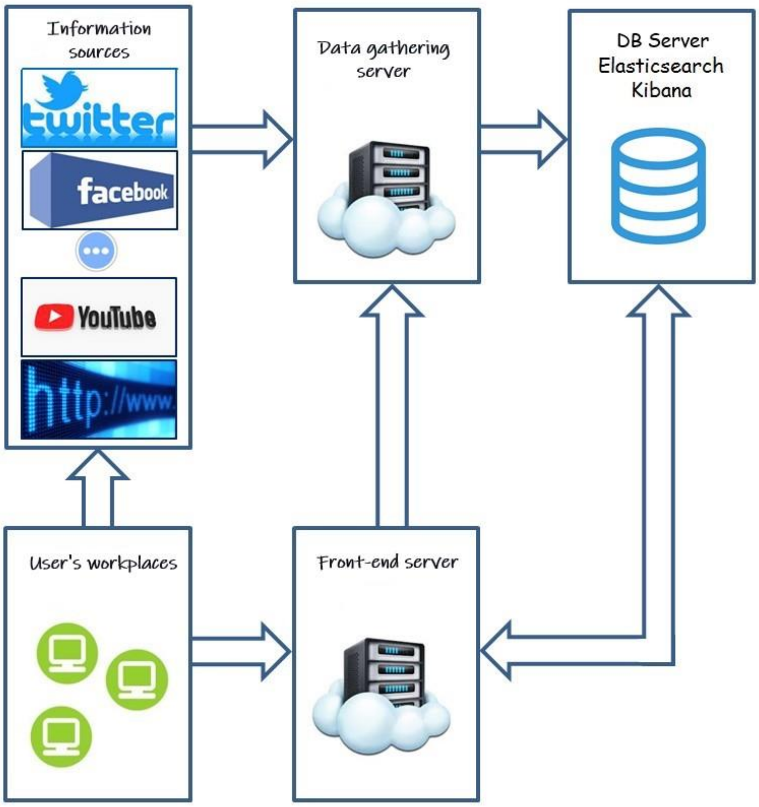


Figure 1 – Scheme of information flows in the "CyberAggregator" system

The basis of the hardware platform of systems for the analysis of big data from social networks are such servers:

- information proxy server (a leased virtual server that provides anonymous information collection, located in an external data center. With the development of the system, there may be several such servers. This server, on the one hand, is designed to provide reliable services to users of corporate networks, and on the other hand, it can provide data exchange with similar external proxies);

- data gathering server (a server for collecting data from Internet resources. It can extract data from scripts defined by the administrator directly from Internet resources or through information proxy servers);
- analytics server (server performs analytical information processing and information retrieval. This server supports historical information databases. Analytical information processing includes extraction of concepts; geoinformation support; determining the tone of messages; information formation; analysis of message dynamics; forecasting; analysis of information sources arrays, etc.);
- interface server (a web server from which end users can access RSS aggregators through web browsers, or system resources through the API applications).

The operation of information aggregation systems from social networks includes the following stages (Fig. 2):

- search for messages from social networks that are relevant to the general broad topic – the formation of information flow from thematic messages;
- determining the language of individual messages that are downloaded from social networks;
- extracts from information messages, such concepts as keywords, personalities, companies, geographical names, etc.;
- tone analysis of individual messages;
- data formatting, converting into standard formats (XML, JSON);
- download the received stream to full-text databases.

The interconnection scheme of the components of the "CyberAggregator" system consists of three main parts – system software, system kernel, and user programs.

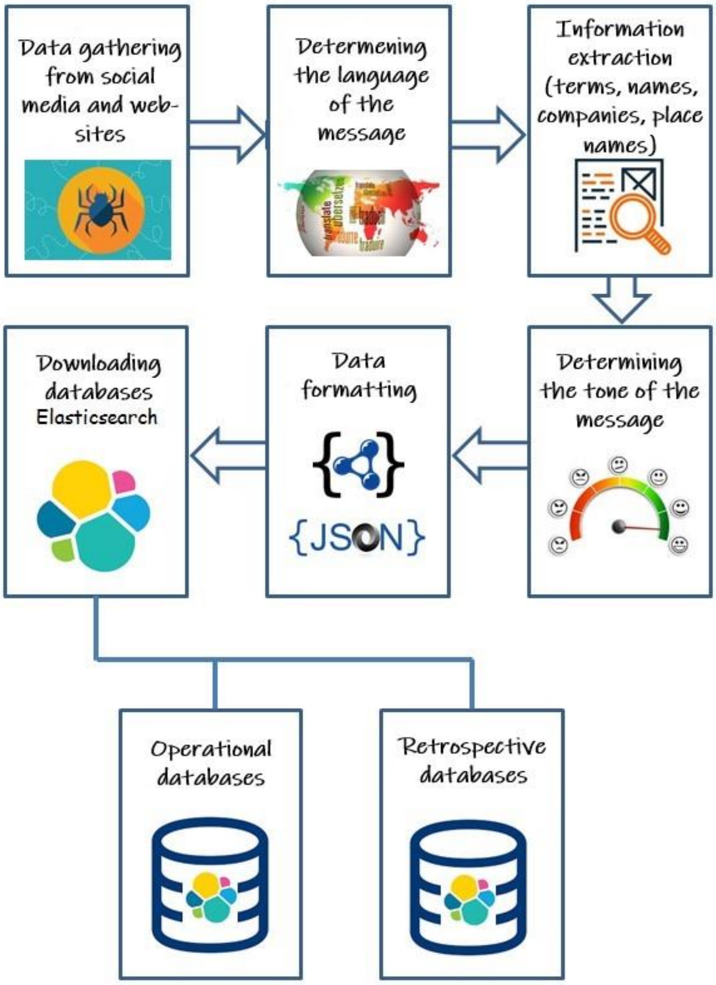

Data gathering from social media and web-sites
Determening the language of the message
Information extraction (terms, names, companies, place names)
Determining the tone of the message
Data formatting
{JSON}
Downloading databases Elasticsearch
Operational databases
Retrospective databases

Figure 2 – Stages of information processing in the "CyberAggregator" system

The system infrastructure includes:

- hardware (servers, telecommunications equipment);
- operating system (Linux);
- programming languages and relevant libraries (Shell, JavaScript, Python, Perl, PHP);
- webserver (Apache, Nginx).
- the core of the system, which in turn includes tools that implement:
  - data collection from social networks;
  - creation and maintenance of databases;
  - full-text search (Elasticsearch system, supplemented by special means of data conversion in RSS format);
  - analytics and forecast based on the study of networks, statistics/dynamics of thematic information flows (Kibana, Gephi, Matlab).

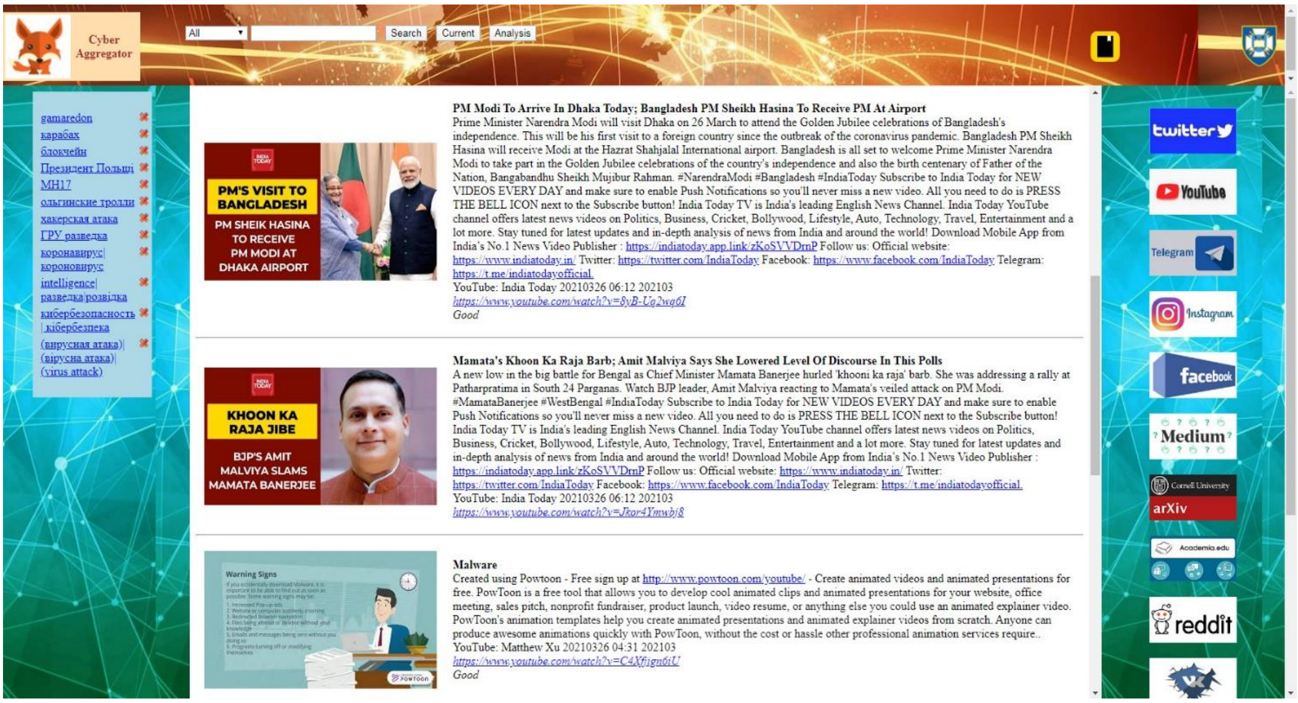
In addition, user programs are provided:

- web browsers;
- RSS aggregators (e.g. FeedDemon 3.5, Feedreader 3.14, RSS Guard 3.4.1), which provide access to CyberAggregator databases and personalization options (maintaining personal databases).

The peculiarities of the considered model are simultaneous use of methods and tools of information retrieval, data analysis, and aggregation of information flows.

User interface

The CyberAggregator system provides the user with a web interface from which information search and analysis functions are available to him (Fig. 3).



gamaredon
карабах
блокчейн
Президент Польщі
MH17
ольгинские тролли
хакерская атака
ГРУ разведка
коронавирус|короновирус
intelligence|разведка|розвідка
кибербезопасность| кібербезпека
(вирусная атака)|(вірусна атака)|(virus attack)

PM Modi To Arrive In Dhaka Today; Bangladesh PM Sheikh Hasina To Receive PM At Airport
Prime Minister Narendra Modi will visit Dhaka on 26 March to attend the Golden Jubilee celebrations of Bangladesh's independence. This will be his first visit to a foreign country since the outbreak of the coronavirus pandemic. Bangladesh PM Sheikh Hasina will receive Modi at the Hazrat Shahjalal International airport. Bangladesh is all set to welcome Prime Minister Narendra Modi to take part in the Golden Jubilee celebrations of the country's independence and also the birth centenary of Father of the Nation, Bangabandhu Sheikh Mujibur Rahman. #NarendraModi #Bangladesh #IndiaToday Subscribe to India Today for NEW VIDEOS EVERY DAY and make sure to enable Push Notifications so you'll never miss a new video. All you need to do is PRESS THE BELL ICON next to the Subscribe button! India Today TV is India's leading English News Channel. India Today YouTube channel offers latest news videos on Politics, Business, Cricket, Bollywood, Lifestyle, Auto, Technology, Travel, Entertainment and a lot more. Stay tuned for latest updates and in-depth analysis of news from India and around the world! Download Mobile App from India's No.1 News Video Publisher : https://indiatoday.app.link/zKoSVVDrnP Follow us: Official website: https://www.indiatoday.in/ Twitter: https://twitter.com/IndiaToday Facebook: https://www.facebook.com/IndiaToday Telegram: https://t.me/indiatodayofficial.
YouTube: India Today 20210326 06:12 202103
https://www.youtube.com/watch?v=8yB-Uq2wq6I
Good

Mamata's Khoon Ka Raja Barb; Amit Malviya Says She Lowered Level Of Discourse In This Polls
A new low in the big battle for Bengal as Chief Minister Mamata Banerjee hurled 'khooni ka raja' barb. She was addressing a rally at Patharpratima in South 24 Parganas. Watch BJP leader, Amit Malviya reacting to Mamata's veiled attack on PM Modi. #MamataBanerjee #WestBengal #IndiaToday Subscribe to India Today for NEW VIDEOS EVERY DAY and make sure to enable Push Notifications so you'll never miss a new video. All you need to do is PRESS THE BELL ICON next to the Subscribe button! India Today TV is India's leading English News Channel. India Today YouTube channel offers latest news videos on Politics, Business, Cricket, Bollywood, Lifestyle, Auto, Technology, Travel, Entertainment and a lot more. Stay tuned for latest updates and in-depth analysis of news from India and around the world! Download Mobile App from India's No.1 News Video Publisher : https://indiatoday.app.link/zKoSVVDrnP Follow us: Official website: https://www.indiatoday.in/ Twitter: https://twitter.com/IndiaToday Facebook: https://www.facebook.com/IndiaToday Telegram: https://t.me/indiatodayofficial.
YouTube: India Today 20210326 06:12 202103
https://www.youtube.com/watch?v=Jkor4Ymwbj8

Malware
Created using Powtoon - Free sign up at http://www.powtoon.com/youtube/ - Create animated videos and animated presentations for free. PowToon is a free tool that allows you to develop cool animated clips and animated presentations for your website, office meeting, sales pitch, nonprofit fundraiser, product launch, video resume, or anything else you could use an animated explainer video. PowToon's animation templates help you create animated presentations and animated explainer videos from scratch. Anyone can produce awesome animations quickly with PowToon, without the cost or hassle other professional animation services require..
YouTube: Matthew Xu 20210326 04:31 202103
https://www.youtube.com/watch?v=C4Xfjign6iU
Good

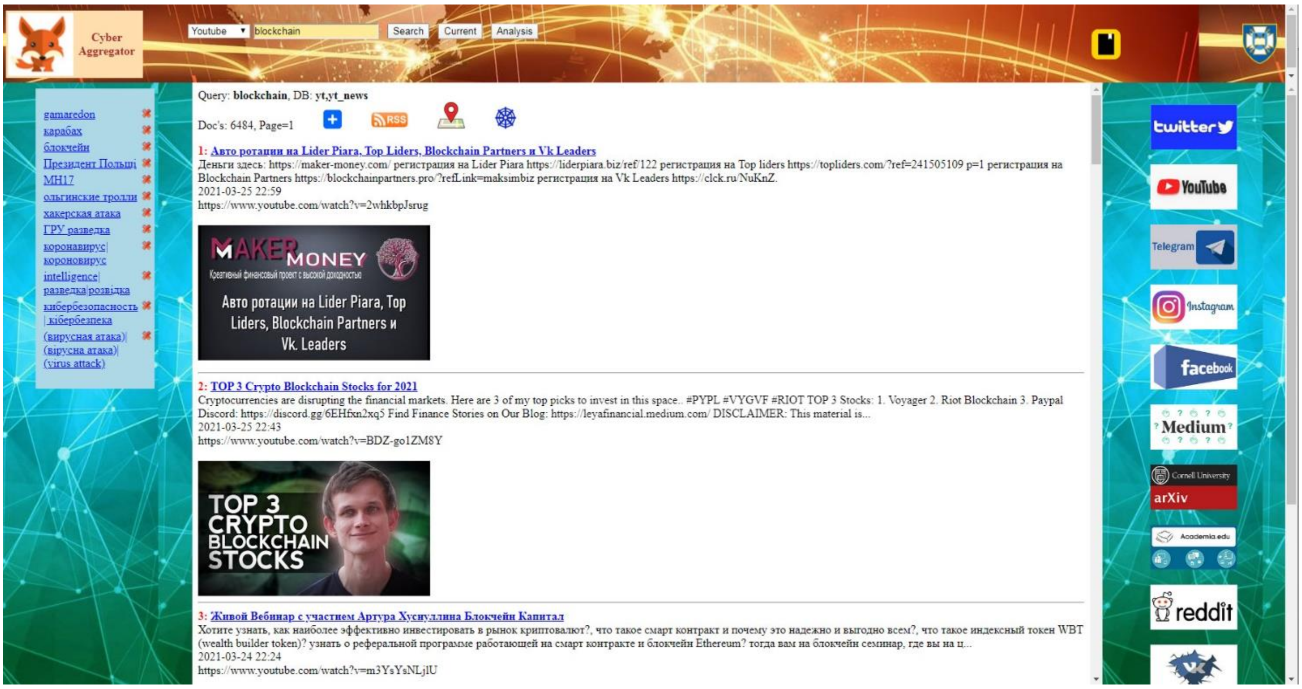
Figure 3 – CyberAggregator system interface

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

The main element of the interface is a digest of the most relevant messages. In a separate block (Requests) saved user requests are displayed. Statistical information on filling the database of the system from individual social networks is available in a special section (Statistics of sources).

As a result of the search on request (Fig. 4), the user is provided with a list of relevant message headers with hyperlinks to the full texts of these messages in the system, as well as to these messages on social networks.

If the request creates documents that meet its information needs [14], it can be saved for future use (Add request). You can further display the found messages in RSS format (with subsequent loading of these results in the so-called RSS aggregators regularly), as well as display search results with details on a map, which is scaled both automatically and through settings (Fig. 5).



Query: **blockchain**, DB: **yt,yt_news**

Doc's: 6484, Page=1

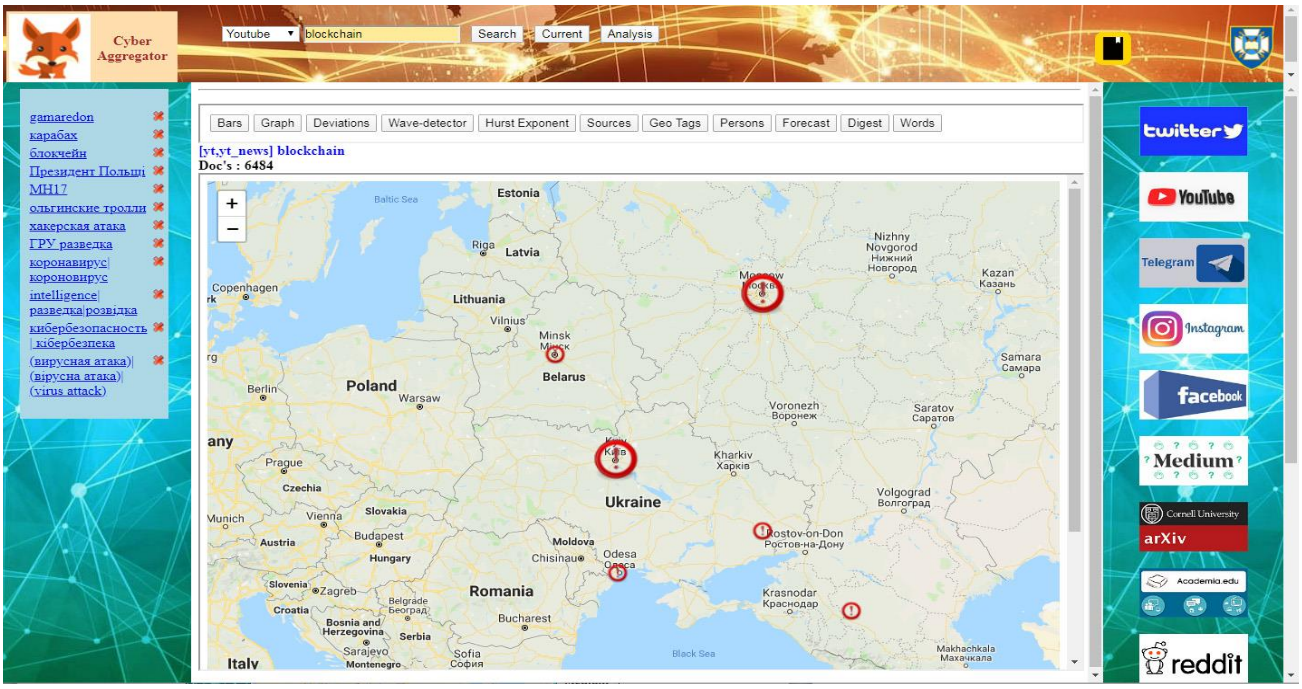
1: Авто ротации на Lider Piara, Top Liders, Blockchain Partners и Vk Leaders
Деньги здесь: https://maker-money.com/ регистрация на Lider Piara https://liderpiara.biz/ref/122 регистрация на Top liders https://topliders.com/?ref=241505109 p=1 регистрация на Blockchain Partners https://blockchainpartners.pro/?refLink=maksimbiz регистрация на Vk Leaders https://clck.ru/NuKnZ.
2021-03-25 22:59
https://www.youtube.com/watch?v=2whkbpJsrug

2: TOP 3 Crypto Blockchain Stocks for 2021
Cryptocurrencies are disrupting the financial markets. Here are 3 of my top picks to invest in this space.. #PYPL #VYGVF #RIOT TOP 3 Stocks: 1. Voyager 2. Riot Blockchain 3. Paypal Discord: https://discord.gg/6EHfxn2xq5 Find Finance Stories on Our Blog: https://leyafinancial.medium.com/ DISCLAIMER: This material is...
2021-03-25 22:43
https://www.youtube.com/watch?v=BDZ-go1ZM8Y

3: Живой Вебинар с участием Артура Хуснуллина Блокчейн Капитал
Хотите узнать, как наиболее эффективно инвестировать в рынок криптовалют?, что такое смарт контракт и почему это надежно и выгодно всем?, что такое индексный токен WBT (wealth builder token)? узнать о реферальной программе работающей на смарт контракте и блокчейн Ethereum? тогда вам на блокчейн семинар, где вы на ц...
2021-03-24 22:24
https://www.youtube.com/watch?v=m3YsYsNLjlU

Figure 4 – Fragment of the user interface in search mode
(search results for "Blockchain")

In Analytical mode (Analysis), the user receives a number of tools, the first of which is a graph (Graph), which corresponds to the time series of the number of corresponding message requests per day.



[yt,yt_news] blockchain
Doc's : 6484

Figure 5 – Fragment of the interface of the geographic information system

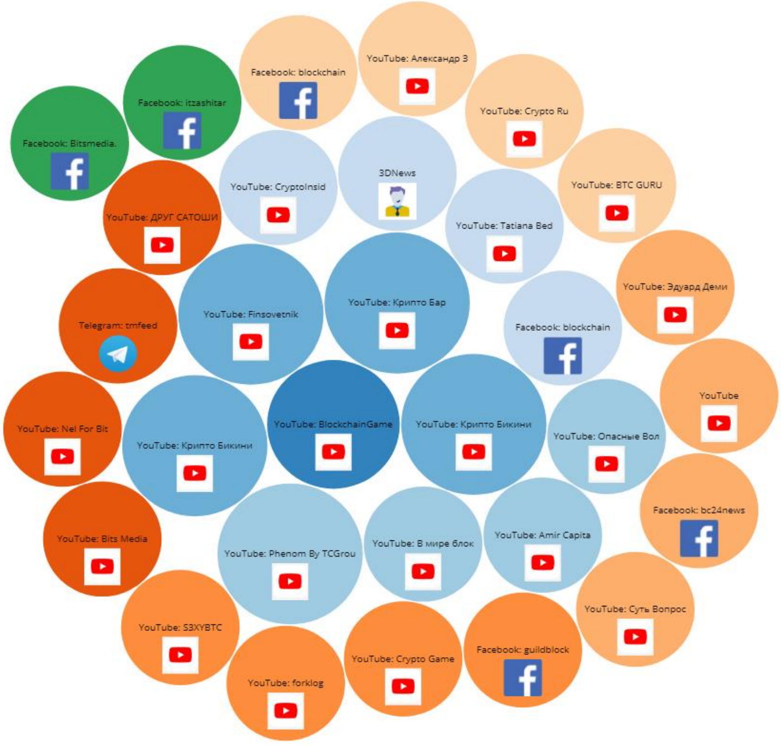


Figure 6 – Network of interconnected sources of information

The user is also allowed to view the main plots (Digest) on the topic and clusters, grouped by predefined keywords.

The system provides modes of forming concept networks that correspond to individual messages (people, brands), sources of information (Fig. 6). These modes allow you to evaluate the concept, to explore the relationship between them.

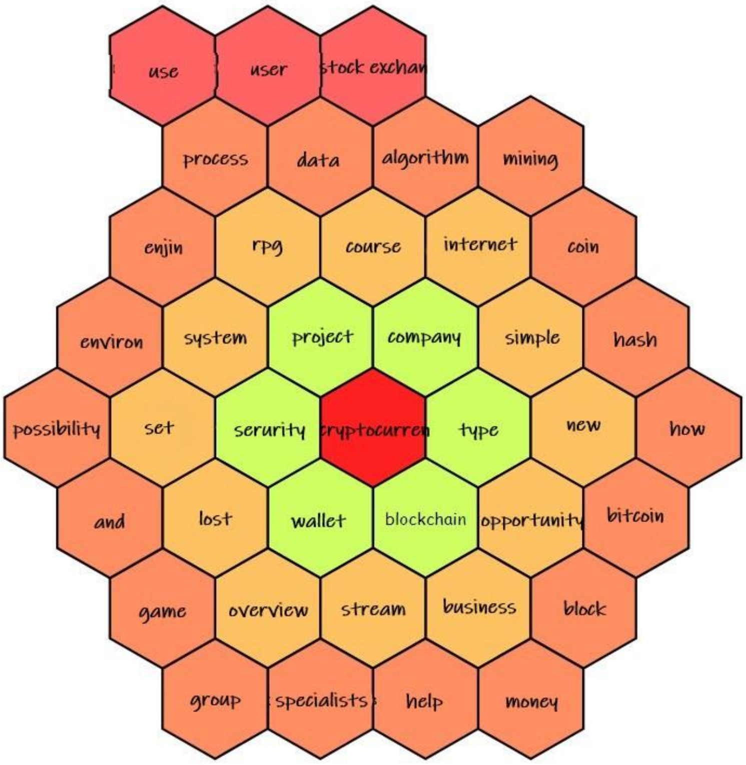


Figure 7 – Fragment of the primary network of terms from the array corresponding to the word from the query

The layout implements an approach to the visualization of thematic clusters, with the task of visualizing a network of terms for the response of the search engine in real-time – a classifier built on the network principle.

The model of dynamic classification of information can be considered as a “word game” [15]. It is the game principle that allowed designing a navigator, which as a result found its application in the real interface, implemented on the basis of using the Javascript library D3.js [16]. The algorithm for constructing a network of terms as a game that takes place on a plane marked by hexagonal cells includes the following steps:

Step 1: the term that corresponds to some concept, which is most common in the system response and corresponds to the initial request, is entered into the central cell (Fig. 7);

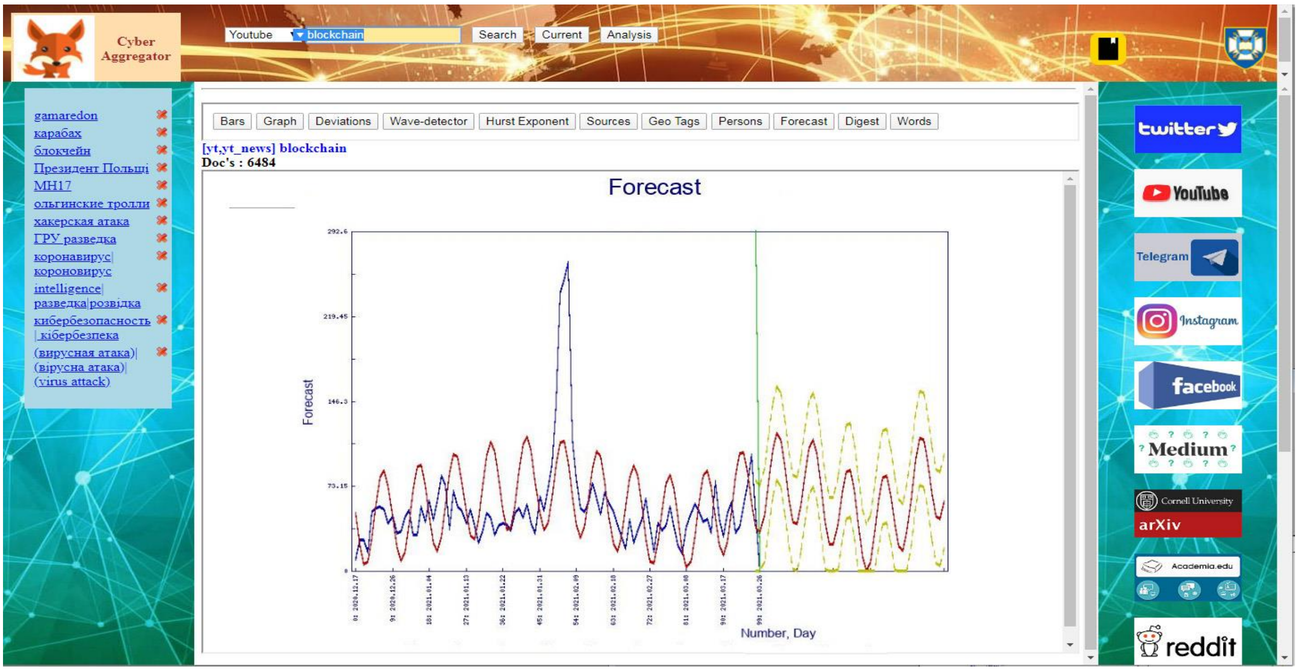
Step 2: Based on the analysis of the relevant query of the array of messages from social networks, six the most relevant terms related to the first term are selected. These terms are entered into the neighboring cells;

The third and the following steps: to the free cells around each of the filled one, the terms most related to the filled cells (up to six, obtained from the same array) are entered. In this case, if the terms have already been used, the neighboring cells remain empty.

The process stops when adding new terms becomes impossible.

In the networks shown in Fig. 7, each cell acts as a hyperlink. Activation of this hyperlink causes clarification of the original request and leads to the output of the relevant documents array.

The illustration of the cell game is quite justified for two reasons: on the one hand, the hexagons cover the plane tightly, and on the other hand, the number of investments in the classifier, not exceeding six, corresponds to the principles of ergonomics.



[yt,yt_news] blockchain
Doc's : 6484

Figure 8 – Sornet algorithm forecast line for the time series for the “Blockchain” query.

It is clear that the information content of the game model requires a very powerful information resource, which is created on the “CyberAggregator” system basis [4].

The “Analytics” mode provides the possibility of forecasting (Forecast) by the method offered by D. Sornette, which is based on the analysis of the regularity of market prices in commodity and stock markets before the crisis. In [17] it is noted that before the crisis, the value of the time series (price) is characterized by an increase in the power law, complicated by periodic fluctuations that converge to a critical point, where the probability of collapse reaches a maximum. The corresponding forecast model, which takes into account linear time-periodic fluctuations, is as follows:

$$F(t) = A + B(t_c - t)^m \left[1 + C\cos\left(\omega \log\left(\frac{t_c - t}{T} \right) + \varphi \right) \right].$$

In this model t_c – critical time (crisis time). The coefficients of the model A, B, ω, φ are determined using a selection procedure. Using the Sornet model (forecast key, Fig. 8), it is possible to obtain forecast values for the number of relevant online publications based on monitoring data.

Conclusions. The paper presents a new approach to the training of cybersecurity professionals in their acquisition of competencies for processing ultra-large data sets, the basis of which is the study of the theoretical foundations and technologies of Big Data. The main components of the approach under consideration are confirmed by implementations in the original layout and practical work performed within the framework of the approach under consideration. The layout, which is designed as a “CyberAggregator” system, was created using free software and software developed by the authors.

The use of information technologies for the creation of a social networks content monitoring system on cybersecurity for the selection of relevant information from social networks, the introduction of a search engine for their refinement by users, saving queries, conducting analytical research, and forecasting is suggested and substantiated.

Definitely, the use of Big Data technologies is possible for different types of data, not only for OSINT, based on which the main task of cybersecurity experts to acquire competencies for working with ultra-large data sets is solved. The presented model can be further deployed for use in information and analytical work by specialists in the security and defense sector.

REFERENCE

[1] B. Franks, and T. Davenport, *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Hoboken, USA: Wiley, 2012.

[2] D. Lande, and E. Shnurko-Tabakova, “OSINT as a part of cyber defense system”, *Theoretical and Applied Cybersecurity: scientific journal*, vol. 1, no. 1, pp. 103-108, 2019, doi: https://doi.org/10.20535/tacs.2664-29132019.1.169091.

[3] D. Cielen, and A. D. B. Maysmen, *Introducing Data Science Big Data, Machine Learning, and more using. Python tools.* New York, USA: Manning Publications, 2016.

[4] O. G. Dodonov, D. W. Lande, V. V. Pryshchepa, and V. G. Putyatin, *Competitive intelligence,* Kyiv, Ukraine: LLC “Engineering”, 2021.

[5] P. J. Sadapage, and M. Fowler, *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Boston, USA: Addison-Wesley Professional, 2013.

[6] T. White, *Hadoop: Detailed guide*, Newton, USA: O’Reilly Media, 2009.

[7] P. Shukla, and S. Kumar, *Elasticsearch, Kibana, Logstash and new generation search engines.* Moskow, Russia: Piter, 2019.

[8] B. Azarmi, *Learning Kibana 5.0. Exploit the visualization capabilities of Kibana and build powerful interactive dashboards*. Birmingham, England: Packt Publishing, 2017.

[9] D. Lande, I. Subach, and A. Puchkov, “System of analysis of big data from social media*”, Information & Security*, vol. 47, no. 1, pp. 44-61, 2020, doi: https://doi.org/10.11610/isij.4703.

[10] D. Lande, O. Puchkov, and I. Subach, “System for analysing of big data on cybersecurity issues from social media”, *Information Technology and Security*, vol. 8, iss. 1, pp. 4-18, 2020, doi: https://doi.org/10.20535/2411-1031.2020.8.1.217993.

[11] K. Cherven, *Mastering Gephi Network Visualization. Produce advanced network graphs in Gephi and gain valuable insights into your network datasets*. Birmingham, England: Packt Publishing, 2015.

[12] K. Cherven, *Network Graph Analysis and Visualization with Gephi Visualize and analyze your data swiftly using dynamic network graphs built with Gephi*. Birmingham, England: Packt Publishing, 2015.

[13] R. V. Bruggen, *Learning Neo4j. Run blazingly fast queries on complex graph datasets with the power of the Neo4j graph database*. Birmingham, England: Packt Publishing, 2014.

[14] B. M. Gerasimov, O. Y. Sergeev, I. Y. Subach, "Extraction of information phrases from primary electronic documents in information retrieval systems", *Control Systems and Machines,* no. 1, pp. 26-29, 2006.

[15] D. W. Lande, and A. N. Grigoriev, "Multilevel classifier-navigator according to the responses of the information retrieval system", in *Proc. Computational linguistics and intelligent technologies: proceedings of the international conference Dialogue,* Russia, 2006, pp. 329-331.

[16] S. Murray, *Interactive Data Visualization for the Web. An Introduction to Designing with D3*. Newton, USA: O'Reilly Media, 2017.

[17] D. Sornette, *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Oxford, England: University Press Scholarship, 2017.

The article was received 29.03.2021.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] B. Franks, and T. Davenport, *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Hoboken, USA: Wiley, 2012.

[2] D. Lande, and E. Shnurko-Tabakova, "OSINT as a part of cyber defense system", *Theoretical and Applied Cybersecurity: scientific journal,* vol. 1, no. 1, pp. 103-108, 2019, doi: https://doi.org/ 10.20535/tacs.2664-29132019.1.169091.

[3] D. Cielen, and A. D. B. Maysmen, *Introducing Data Science Big Data, Machine Learning, and more using. Python tools.* New York, USA: Manning Publications, 2016.

[4] А. Г. Додонов, Д. В. Ландэ, В. В. Прищепа, и В. Г. Путятин, "Компьютерная конкурентная разведка", Киев, Украина: ООО "Инжиниринг", 2021.

[5] P. J. Sadapage, and M. Fowler, *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Boston, USA: Addison-Wesley Professional, 2013.

[6] T. White, *Hadoop: Detailed guide*, Newton, USA: O'Reilly Media, 2009.

[7] P. Shukla, and S. Kumar, *Elasticsearch, Kibana, Logstash and new generation search engines.* Moskow, Russia: Piter, 2019.

[8] B. Azarmi, *Learning Kibana 5.0. Exploit the visualization capabilities of Kibana and build powerful interactive dashboards*. Birmingham, England: Packt Publishing, 2017.

[9] D. Lande, I. Subach, and A. Puchkov, "System of analysis of big data from social media*",* *Information & Security*, vol. 47, no. 1, pp. 44-61, 2020, doi: https://doi.org/10.11610/isij. 4703.

[10] Д. Ланде, О. Пучков, та І. Субач, "Система аналізу великих обсягів даних з питань кібербезпеки із соціальних медіа", *Information Technology and Security,* vol. 8, iss. 1, pp. 4-18, 2020, doi: https://doi.org/10.20535/2411-1031.2020.8.1.217993.

[11] K. Cherven, *Mastering Gephi Network Visualization. Produce advanced network graphs in Gephi and gain valuable insights into your network datasets*. Birmingham, England: Packt Publishing, 2015.

[12] K. Cherven, *Network Graph Analysis and Visualization with Gephi Visualize and analyze your data swiftly using dynamic network graphs built with Gephi*. Birmingham, England: Packt Publishing, 2015.

[13] R. V. Bruggen, *Learning Neo4j. Run blazingly fast queries on complex graph datasets with the power of the Neo4j graph database*. Birmingham, England: Packt Publishing, 2014.

[14] B. M. Gerasimov, O. Y. Sergeev, I. Y. Subach, "Extraction of information phrases from primary electronic documents in information retrieval systems", *Control Systems and Machines*, no. 1, pp. 26-29, 2006.

[15] Д. В. Ланде, и А. Н. Григорьев, "Многоуровневый классификатор-навигатор по откликам информационно-поисковой системы", на *Международной конференции*

Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции, Москва, 2006, С. 329-331.

[16] S. Murray, *Interactive Data Visualization for the Web. An Introduction to Designing with D3*. Newton, USA: O'Reilly Media, 2017.

[17] D. Sornette, *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Oxford, England: University Press Scholarship, 2017.

ДМИТРО ЛАНДЕ,
ОЛЕКСАНДР ПУЧКОВ,
ІГОР СУБАЧ

АГРЕГАЦІЯ ІНФОРМАЦІЇ З РІЗНОРІДНИХ МЕРЕЖ ЯК ОСНОВА ПІДГОТОВКИ ФАХІВЦІВ З КІБЕРБЕЗПЕКИ З ПИТАНЬ ОБРОБЛЕННЯ НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Обґрунтовано і представлено основні засади підготовки фахівців з кібербезпеки з питань оброблення надвеликих масивів даних для вирішення складних неструктурованих задач в ході виконання ними функціональних обов'язків на основі досягнень науки про дані у сфері кібербезпеки, шляхом набуття ними необхідних компетентностей, а також практичного застосування новітніх інформаційних технологій, що ґрунтуються на методах агрегації великих обсягів даних. Розглянуто найпоширеніші новітні технології та інструменти у сфері кібербезпеки, перелік яких дозволяє отримати досить цілісне уявлення про те, що використовують сьогодні фахівці в галузі Data Science. Проаналізовано інструменти, якими необхідно володіти, щоб вирішувати складні завдання з використанням великих даних. Предметом дослідження є фундаментальні положення про концепцію "великих даних"; відповідні моделі даних; архітектурні концепції створення інформаційних систем для "великих даних"; аналітика "великих даних", а також практичного застосування результатів обробки "великих даних". Розглянуто теоретичну основу підготовки, що включає два розділи: "Великі дані: теоретичні засади" і "Технологічні застосування для великих даних", які, у свою чергу, логічно розбиті на десять тем. У якості матеріально-технічної бази для набуття практичних навичок тих, хто навчаються, створено та описано макет на основі системи "КіберАгрегатор", який функціонує і постійно удосконалюється відповідно до розширення переліку завдань, які на нього покладаються. Система "КіберАгрегатор" складається з трьох основних частин: сервер для збору та первинної обробки інформації; сервер пошуку інформації (пошукова система); інтерфейсний сервер, з якого послуга надається користувачам та іншим системам через API. Система базується на таких технологічних компонентах, як інформаційно-пошукова система Elasticsearch, утіліти Kibana, графової системи керування базами даних Neo4j, засобів візуалізації результатів на основі JavaScript (D3.js) і модулів сканування мережевої інформації. Система забезпечує реалізацію таких функцій, як формування баз даних з визначених інформаційних ресурсів; ведення повнотекстових баз даних з інформації; виявлення дублікатів, схожих за змістом інформаційних повідомлень; повнотекстовий пошук; аналіз текстових повідомлень, визначення тональності, формування аналітичних звітів; інтеграцію з географічною інформаційною системою; аналіз та візуалізацію даних; дослідження динаміки тематичних інформаційних потоків; прогнозування розвитку подій на основі аналізу динаміки публікацій тощо. Запропонований підхід дозволяє тим, хто навчається отримати необхідні компетентності, які необхідні для ефективної обробки великих обсягів даних із соціальних мереж, створення систем моніторингу мережевої інформації з питань кібербезпеки, відбору релевантної інформації із соціальних мереж, впровадження пошукової системи, проведенні аналітичних досліджень, прогнозування.

Ключові слова: великі дані, соціальні мережі, кібербезпека, інформаційно-пошукові системи, агрегація даних, наука про дані, інформаційна технологія.

Lande Dmytro, doctor of technical science, professor, head at the specialized modeling tools department, Institute for information recording of National academy of science of Ukraine, Kyiv, Ukraine, ORCID 0000-0003-3945-1178, dwlande@gmail.com

Puchkov Oleksandr, candidate of philosophy science, professor, head, Institute of special communication and information protection National technical university of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine, ORCID 0000-0002-8585-1044, iszzi@iszzi.kpi.ua.

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, igor_subach@ukr.net.

Ланде Дмитро Володимирович, доктор технічних наук, професор, завідувач відділом спеціалізованих засобів моделювання, Інститут проблем реєстрації інформації Національної академії наук України, Київ, Україна.

Пучков Олександр Олександрович, кандидат філософських наук, професор, начальник, Інститут спеціального зв'язку та захисту інформації Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", 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.