UDC 004.5

OSINT as a part of cyber defense system

Dmytro V. Lande[1], Ellina V. Shnurko-Tabakova[2]

[1]*Institute for Information Recording of NAS of Ukraine*
[2]*Index Systems ltd.*

Abstract

The paper presents the results of research on the development of fundamental and applied principles for analyzing information flows in global computer networks while conducting open source intelligence (OSINT). The relevance of this task, in particular, concerning the provision of cyber security, the parameters of the modern information space, the existing theoretical and technological solutions are substantiated. The description of methodological and instrumental means of analysis and modeling of information flows, distributed content monitoring of global networks, the creation of multilingual full-text databases, analysis of the dynamics of thematic information flows with the use of nonlinear analysis, automatic formation of models of subject areas in the field of cyber security are presented.

Keywords: Open source intelligence (OSINT), global computer networks, cyber system.

Introduction

A significant amount of information resources in global networks contains various expert assessments, some of which are related to the implementation of cybernetic information threats.

Expert assessments contained in open source documents can be analyzed, synthesized, created as a basis for further decision-making. They differ from traditional expert assessments both in volumes and in terms of objectivity. In addition, networks may contain information related to the organization of unlawful activities, in particular, cybercrime, cyber terrorism.

On this basis, taking into account information from the web feeds is of great importance for solving problems in the field of ensuring cyber security.

Today, the so-called "OpenS INTelligence" (OSINT) is one of the most important cyber security tools. OSINT is one of the intelligence domains, including search, selection, and collection of intelligence information, available from publicly available sources, as well as analysis of this information.

OSINT is, usually, performed through monitoring, analysis, and research of information coming from the Internet. Materials, compiled based on information from open sources, support all intelligence methods and activities through accumulation of intelligence knowledge, its analysis and dissemination.

According to the CIA analyst Sherman Kent (1947), politicians get up to 80 percent of information, needed for decision-making in times of peace, from open sources. Later lieutenant-general Samuel Wilson, head of Intelligence department of US Department of defense in 1976—1977, noted that 90 percent of intelligence data came from open sources, and only 10 – from agents.

Issues Facing OSINT in the Field of Cyber Security

According to [1], OSINT is also one of the ways of intelligence that significantly contributes to planning of military actions, and provides all the necessary information for these actions. It is also noted that:

1) OSINT is one of the methods of intelligence through gathering of information from open sources, its analysis, preparation, and timely submission of the final product to higher management for solution of certain intelligence problems.
2) OSINT is an intelligence method, developed based on the collection and analysis of publicly available information, and not subject to direct supervision by US government. OSINT is a result of systematized collection, processing, and analysis of the necessary publicly available information.

When researching cyber security issues, OSINT allows to receive answers to the following questions:

– Who is the initiator of a cyber attack?
– What are his motives?
– How is it organized?
– What tools are used?
– What resources are involved in the attack?
– What events accompany attacks?
– What are the consequences of an attack?

Benefits of OSINT

International community is using more and more information from open sources to solve a wide spectrum of problems. Particularly, the role of OSINT during implementation of information operations is defined by the set of aspects, including efficiency of information flow, volume, clarity, ease of subsequent usage, cost of obtaining, etc.

The following factors influence the process of planning and preparation of OSINT measures:

- Effective information support. Most of the necessary reference materials on information operation objects are gathered from open sources. This base is built through collection of information from the media. Accumulation of data from open sources is the key function of OSINT.
- Relevance. Availability, depth, and scale of publicly available information allow us to find the necessary information without engaging specialized human and technical intelligence means.
- Simplification of data collection processes. OSINT provides the necessary information eliminating the need for engaging redundant technical and human intelligence methods.
- Depth of data analysis. Being a part of intelligence process, OSINT allows managers to perform in-depth analysis of publicly available information in order to make respective decisions.
- Efficiency. Sharp reduction of time of access to information on the Internet. Reduction of the number of man-hours, spent on search for information, people, and their interrelations based on open sources. Quick obtaining of valuable relevant information. Abrupt situation changes during crises are most thoroughly reflected by current news, so (as we know for sure), the downfall of the Berlin wall was witnessed in both Washington and the CIA headquarters in Langley, not through intelligence service reports, but through TV screens, broadcasting CNN reports right from the scene.
- Volume. Opportunity for mass monitoring of certain information sources, intended for search of the needed content, people, and events. Experience shows that proficiently collected information fragments from open sources, when taken as a whole, can prove equivalent to or even more significant than professional intelligence reports.
- Quality. In comparison with reports of special agents, information from open sources turns out to be more preferable, at least because it is unbiased and not mixed with lies.
- Clarity. So, while in the cases when OSINT is used, trustworthiness of open sources can be both clear and unclear, in the case of secretly obtained data, their credibility is always doubtful.
- Usability. Any secrets are supposed to be protected by barriers of "classifications", clearances, restricted access etc. As for OSINT data, it can be easily communicated to any interested organizational bodies. There is an opportunity to conduct complex research based on data from the Internet.
- Cost. Cost of obtaining data through OSINT is minimum; it is defined only by the price of the service used.

Information from Network Sources

Today, the development of special methods and, means of search and analytical generalization of network information is particularly relevant. However, there are still no qualitative solutions to the problem of rapid analytical processing of information, searching for necessary factual data, identifying development trends in subject domains, and forecasting. Problems of dimension and dynamics of multilingual information resources in global networks require basic research in the areas of mathematics (graph theory, complex networks), pattern recognition (classification, cluster analysis, neural networks), computer linguistics, digital signal processing, nonlinear analysis, and the like.

Currently, the volume of information resources placed in global networks exceeds hundreds of trillions of documents. According to the well-known Internet service Netcraft (http://netcraft.com), only the number of sites on the network already in 2014 exceeded a billion and currently stands at 1.7 billion.

In 2014, the approximate amount of data on the Internet is already 10^{24} bytes, one Yottabyte. Facebook social network alone generates more than 4 petabytes of data per day. According to CISCO, the volume of Internet traffic in 2016 reached one Zettabyte (10^{21} bytes).

The statistics of the most popular social networks are as follows:

- Facebook (2016): 17.9 billion active users per month (monthly active users - MAU), 1 billion MAU from mobile devices.
- Chinese social network Sina Weibo - 392 million MAU.
- Twitter microblogging network - 319 million MAU.
- Linkedin professional contacts network - 106 million MAU.

Technological concepts

- When collecting and analyzing open data from the Internet, there are problems of processing very large volumes, the need to search and navigate dynamic information flows. A huge number of multilingual dynamic information resources, the dominance of information noise makes it difficult to find the necessary information, operational analysis, and hence the use of OSINT in information and analytical work.
- Most of the above problems are current issues of semantic processing of very large dynamic text arrays of information.
- Currently, the following technological concepts are used to solve these problems, such as Big Data, Complex Networks, Cloud Computing, Data/Text Mining.
- All over the world, the tasks of effective analytical processing of information from global networks, the rapid extraction of necessary factual data, the identification of trends in individual subject areas, the recognition of meaningful anomalies, forecasting, and the like are still unsolved. Most of these tasks are actual problems of semantic processing of very large dynamic data arrays. Even partial attempts at a practical solution of these problems today determine the success of projects such as the search engines Google, Yandex, Baidu, Keyhole-

based social network monitoring systems (SMM), Brandwatch, CyberAlert, YouControl, analytical systems such as Palantir, Centrifuge, *i*2, etc.
- An ontological approach is increasingly being used to build subject domain models, in particular, cybersecurity.
- For building domain models, in the framework of the OSINT work, the following theoretically justified and created tools:
  - distributed content monitoring of global networks;
  - the formation of multilingual full-text databases;
  - analysis of the dynamics of thematic information flows;
  - identification of terminological bases of subject areas;
  - the formation of networks of the interrelation of concepts that are extracted from text arrays and the analysis of such networks;
  - identification of implicit connections of objects.

Network of Distributed Information Proxy Servers

Monitoring systems for information flows on the Internet should not be a source of overloading of targeted web resources. In addition, they must ensure the reliable operation of OSINT systems, despite the possible blocking of network fragments or individual IP addresses, and the conclusion of individual scanning servers.

As an approach to solving these problems, it is proposed to build a system of related information proxy servers [2]. At the same time, the following requirements are imposed on the functioning of such a network of information proxy servers:
- the dynamic news component of web space should be taken into account as the most critical from the point of view of providing access;
- many websites for caching are selected by experts in accordance with the contribution of these sources in the information space;
- information in the proxy server should be stored in a universal format that provides for an unambiguous syntactic interpretation (for example, XML);
- the data in the information repository should be updated according to a schedule, which corresponds to the dynamics of their updating on websites.

A proxy server, on the one hand, is designed to reliably serve users of corporate networks, and on the other hand, it can communicate with similar external proxy servers.

A System of Multilingual Full-Text Databases

The system of multilingual full-text databases should be created on the basis of the Big Data concept should include the following technological tools:

1) the collection and initial processing of information, the formation of multilingual incoming information flows, given in national codings, for which the formation of macro information of information resources corresponding to different languages should be provided;
2) identification of keywords by statistical algorithms in information materials presented in different languages;
3) thematic rubric of documents;
4) creation, rotation of databases and provision of the formation of internal index data sets;
5) creation and maintenance of databases that correspond to different languages, including the creation of separate search indexes of databases covering words spoken in different languages;
6) searching in databases on the requests given in different languages;
7) the formation of analytical reports, including information portraits and strings of target stories.

Analysis of the Thematic Information Flows Dynamics

Thematic information flows can be put in accordance with the time series (publication intensity per unit time) [3], for the analysis of which the formal methods are more and more often applied: statistical, fractal, Fourier or wavelet analysis.

The analysis of these numerical series provides the possibility of identifying trends, cycles, anomalies, further forecasting the development of information plots related to selected topics, and a comparison of numerical series corresponding to various information plots.

In fig. 1 shows how the infoStream content monitoring system (http://infostream.ua) tracks publications related to the spread of the Petya computer virus in mid-2017. For this, the query "Virus & Petya" was entered through the web interface of the system.

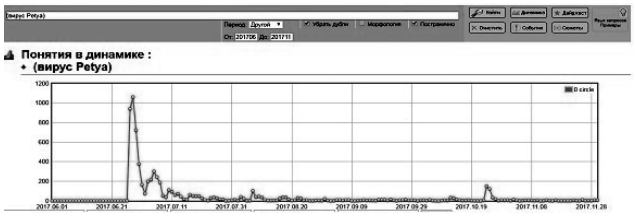

Fig. 1. Diagram of the dynamic of the content-monitoring system by the concept "Virus Petya"

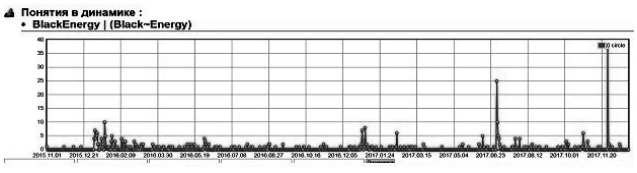

Fig. 2. Diagram of the dynamic of the content-monitoring system by the concept BlackEnergy

In fig. 2 shows the dynamics of mentions in the messages of the content monitoring system about the BlackEnergy virus from November 2015 to November 2017.

Information Operations Recognition

As a methodological basis for identifying information operations, research on the dynamics of information flows has been proposed [4]. On the basis of system studies of such a multidimensional phenomenon as information operations, a generalized diagram is considered corresponding to the stages of the life cycle of information operations (fig. 3).

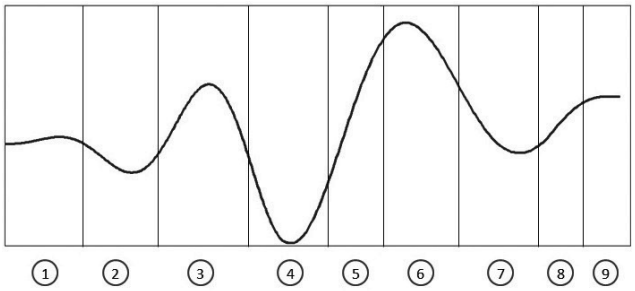

Fig. 3. Generalized diagram, illustrating all the stages of information operation lifecycle: 1 – background; 2 – calm; 3 – "preparatory shelling"; 4 – calm; 5 – attack/growth trigger; 6 – overestimated expectations peak; 7 – loss of illusions; 8 – public realization; 9 – productivity/background

Application of Wavelet Analysis

One of the effective means of identifying patterns in various informational scales in thematic information flows is wavelet analysis. It is determined that the dynamics of information operations most accurately reflect such well-known wavelets as the "Mexican hat" and Morlet.

The technology of using wavelets [5, 6] allows to identify single and irregular "bursts", sharp changes in the values of quantitative indicators in different periods of time, in particular, the volume of thematic publications on the Internet. At the same time, moments of occurrence of cycles can be detected, as well as moments when chaotic oscillations occur over periods of regular dynamics.

Periodic changes that occur for the values of the wavelet transform coefficients on some set of frequencies look like a chain of "hills" that have vertices located at points (along the time axis) at which these changes reach maximum values. Another important indicator is the trend of the time series regardless of periodic fluctuations. The presence of a trend may not be obvious by simply considering the time series, for example, if the trend is combined with periodic fluctuations. The trend is displayed on the scalers as a smooth change in brightness along the time axis simultaneously on all scales. If the trend is growing, the brightness will increase, if decreasing – will decrease. Also an important factor that must be taken into account when analyzing time series is local features, that is, sudden, abrupt changes in the characteristics of the original series are possible.

The corresponding wavelet spectrogram (fig. 4) shows all the characteristic features of the original series: the scale and intensity of periodic changes, the direction and value of trends, the presence, location and duration of local features.

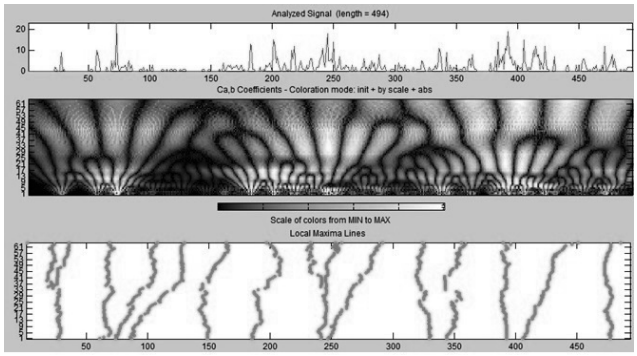

Fig. 4. Dynamics of publications on target topics and wavelet-scalegram (Morlet wavelet) of information stream

Forecasting of Events

For the study of time series, corresponding to arrays of messages in thematic information flows, along with traditional statistical approaches, the theory of fractals is increasingly used today, the traditional scope of which is fractal geometry, image processing, and so on. At the same time, time series generated by thematic information flows also have fractal properties and can be considered as stochastic fractals. This approach extends the field of application of the theory of fractals to information flows, the dynamics of which are described by means of the theory of random processes.

On the other hand, the theory of fractals allows you to obtain important characteristics of information flows without going into a detailed analysis of their internal structure and relationships. The most important characteristic of the series with chaotic behaviour is the fractal dimension, which in many cases can be calculated using the so-called R/S -analysis [7]. More precisely, the fractal dimension is not calculated, but the Hurst index (H), which is associated with it in a simple relation S. R/S -analysis is based on the analysis of the spread of values R of the studied series and the mean square deviation S. Numerical values H characterize different types of correlation dynamics (persistence). When $H = 0.5$ an uncorrelated behavior is observed, the values $0.5 < H < 1$ correspond to the level of autocorrelation of the series.

Automatic Formation of Concepts Interconnections Networks and Subject Domain Models

Using automatic methods of extracting factual data used in systems for integrating Internet resources, so-called "information portraits" are formed, covering lists of reference words, person names, toponyms, like companies, etc., contained in documents relevant to a certain thematic request. These lists can be aggregated, as a re-

sult of which it is possible to determine the interrelations of individual concepts.

Today, the subject domain model, in particular, is understood as a specially formed network of concepts, ontology. As a terminological basis for the formation of terminological ontologies, the basic concepts from this subject area — elements of the so-called "information portraits" — are automatically revealed. After automatic detection of connections between individual concepts, networks are formed that correspond to the subject areas [8] and are displayed by means of graph analysis and visualization software (fig. 5).

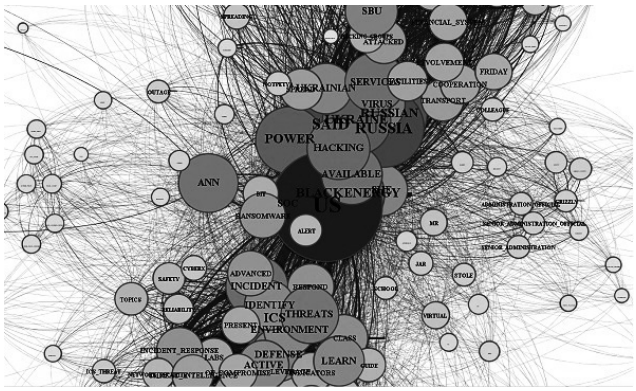

Fig. 5. Visualization of a fragment of the subject domain model corresponding information stream on subject BlackEnergy

Scenario Analysis Based on the Use of Cognitive Maps

Based on the analysis of cognitive maps, information support scenarios are formed, confronting destructive influences, which are associated with certain concepts of the subject area. On the basis of domain models, the formation of cognitive maps is provided [9], the vertices of which correspond to certain concepts (concepts), and edges (links) - cause-and-effect (causal) connections between concepts. After selecting target concepts in cognitive maps, subgraphs are selected that are most closely associated with target objects.

In the formation of cognitive maps solved several meaningful tasks, among which are:

1) Detection of nodes-concepts.
2) Detection of semantic (causal) links between nodes.
3) Ranking concepts, identifying the main ones.
4) Ranking of connections.
5) Detection of concepts affecting target nodes objects.
6) Detection of subgraphs, most closely related to target objects.
7) Visualization and editing of cognitive maps.
8) Formation of scenarios.
9) Prediction of situation development as a result of implementation of scenarios.

International Experience

Open source intelligence enhances the efficiency of the whole US intelligence community, from national to tactical levels. Below we list some organizations in the US that collect, purchase, utilize, analyze, and disseminate information from open sources:

- Defense Open Source Council (DOSC);
- Intelligence and Security Command (INSCOM);
- Department of the Army Intelligence Information Service (DA IIS);
- Director of National Intelligence of Open Source Center (DNI OSC);
- Open Source Academy;
- Advanced Systems Department (ASD);
- FBI;
- Federal Research Department (FRD), Congress Library.

Units of US armed forces that participate in OSINT activities are listed below:

- Unified Combatant Command;
- Defense Intelligence Agency;
- National Geospatial-Intelligence Agency;
- US Army Foreign Military Studies Office;
- EUCOM JAC Molesworth;
- Foreign Media Monitoring in Support of Information Operations, U.S. Strategic Command.

In addition to broad OSINT application in the USA, let us provide examples of the technology usage in other countries.

German external intelligence service, Federal intelligence service, also utilizes the advantages of Open Source Intelligence in the units of Abteilung Gesamtlage/FIZ and Unterstützende Fachdienste (GU).

In Australia the primary expert body for open sources is the Office of National Assessments that is one of intelligence structures.

In the UK there is the BBC Monitoring information service, focusing on collection of open-access information with the efforts of journalists. The task of analysis of data, collected by the BBC, is placed upon the subscribers, including employees of British secret services.

Conclusion

At present, software and technology solutions for OSINT are being created and implemented around the world. These solutions are used in automation of information and cyber security measures, increase the efficiency of analysis of large volume of information, enable identification of natural and simulated information flows, identify the links of participants in information processes and involve information resources. Recognition of information operations allows detecting and organizing counteraction, tracking information and cybernetic security of systems, vulnerability of critical infrastructures.

References

[1] "Army techniques publication no. 2-22.9." Headquarters Department of the Army Washington, DC, 7 2012. (FMI 2-22.9).

[2] A. Dodonov and D. Lande, "Organization of information proxy servers network," *Data Recording, Storage & Processing*, no. 3, pp. 24–32, 2006.

[3] J. Kleinberg, *Temporal Dynamics of On-Line Information Streams*. Springer, 2005.

[4] A. Dodonov, D. Lande, and V. Putyatin, *Computer networks and analytical research*. IIR of NAS of Ukraine, 2014.

[5] N. Astafieva, "Wavelet analysis: bases of the theory and examples of application," *Achievements of physical sciences*, no. 11, pp. 1145–1170, 1996.

[6] A. Davydov, "Wavelet analysis of social processes," *Sociological researches*, no. 11, pp. 97–103, 2003.

[7] J. Feder, *Fractals*. Plenum, 1988.

[8] D. Lande, I. Balagura, and V. Andrushchenko, "The detection of actual research topics using co-word networks," *Open Semantic Technologies for Intelligent Systems (OSTIS-2018): Proceedings of the international scientific and technical conference*, 2018.

[9] R. Axelrod, *Structure of decision: The cognitive maps of political elites*. Princeton University Press, 1976.