Using Ontologies in Cybersecurity Field

Tiberiu Marian GEORGESCU, Ion SMEUREANU
Bucharest University of Economic Studies, Romania
tiberiugeorgescu@ase.ro, smeurean@ase.ro

This paper is an exploratory research which aims to improve the cybersecurity field by means of semantic web technologies. The authors present a framework which uses Semantic Web technologies to automatically extract and analyse text in natural language available online. The system provides results that are further analysed by cybersecurity experts to detect black hat hackers' activities. The authors examine several characteristics of how hacking communities communicate and collaborate online and how much information can be obtained by analysing different types of internet text communication channels. Having online sources as input data, the model proposed extracts and analyses natural language that relates with cybersecurity field, with the aid of ontologies. The main objective is to generate information about possible black hat hacking actions, which later can be analysed punctually by experts. This paper describes the data flow of the framework and it proposes technological solutions so that the model can be applied. In their future work, the authors plan to implement the framework described as a system software application.

Keywords: *Cybersecurity, Ontologies, Semantic Web*

1 Introduction

The evolution and expansion of the internet facilitated by great developments in fields such as Big Data, Artificial Intelligence and Machine Learning led to a great change in the virtual environment. The internet transitioned from an environment designed for humans to one where both humans and machines exist and interact. The Semantic Web was first introduced by Tim Berners-Lee et al back in 2001, Berners-Lee being no one else, but the creator of World Wide Web (abbreviated WWW). [1] Since then, the Semantic Web technologies became wide-spread with applications in various fields. The transition from machine readable information to machine understandable information is possible by expressing the information in languages such as RDF and OWL. [2]

In this article the authors discuss how Semantic Web technologies can be used in Cybersecurity field. Cybersecurity is arguably a very complex and extensive domain, whose activities can be classified in two: those undertaken to design an optimal system, with as few vulnerabilities as possible and those that are taken as a result of the problems that appear after the system is operational. While the first type of activities have a relatively common approach to improve security for programming developers, the second is a cat and mouse game, where as soon as a black hat hacker manages to find (and exploit) a type of vulnerability, the system experts work to solve it. In contrast to the white hat hackers, the black hat hackers access and perform actions on a computer system illegally, without the owner's permission, in order to gain personal advantages. One of the objectives of this article is to recognize and discuss actions that can be done between the two types of activities described above, with the help of Semantic Web technologies.

The authors propose a framework based on Semantic Web technologies which aims to extract and analyse text in natural (human) language available online and provide results that can improve Cybersecurity. As Abbasi et al point out, "there is a lack of research that explores automated identification and characterization of expert hackers within online communities" [3].

Section 2 presents the main semantic web standards which are considered for the model proposed. Section 3 discusses the borders in which Semantic Web technologies can be

DOI: 10.12948/issn14531305/21.3.2017.01

used to improve Cybersecurity. The authors describe the types of results expected, based on different types of online sources. They also analyse the types of input data, which consists in any online source about Cybersecurity which may link with black hat hacking. Section 4 highlights the main solutions for web data extraction, illustrates the main differences between scrapers and crawlers and compares the main characteristics of crawlers. Section 5 presents a framework which detects potential Cybersecurity threats based on Semantic Web technologies, as well as the data flow of the model. The 6th section display the authors' conclusion and future work.

2 Semantic Web Standards

Semantic Web is an extension of World Wide Web, where unstructured data is interpreted by machines through ontologies. Borrowed from philosophy, in IT, ontologies are considered explicit, formal definitions of the entities of reality, based on classes, relations and individuals. Essentially, ontologies are tools that provide to the machines the means of understanding natural language. If machines can properly interpret hacker community's discussions then it is likely that cybersecurity field can be improved.

For the model described below, the authors expect to develop ontologies by using the following standards: XML (Extensible Markup Language), RDF (Resource Description Framework), RDFS (Resource Description Framework Schema), OWL (Web Ontology Language) and SPARQL (SPARQL Protocol and RDF Query Language). Figure 1 illustrates the main concepts and abstractions as well as the semantic web specifications and solutions.

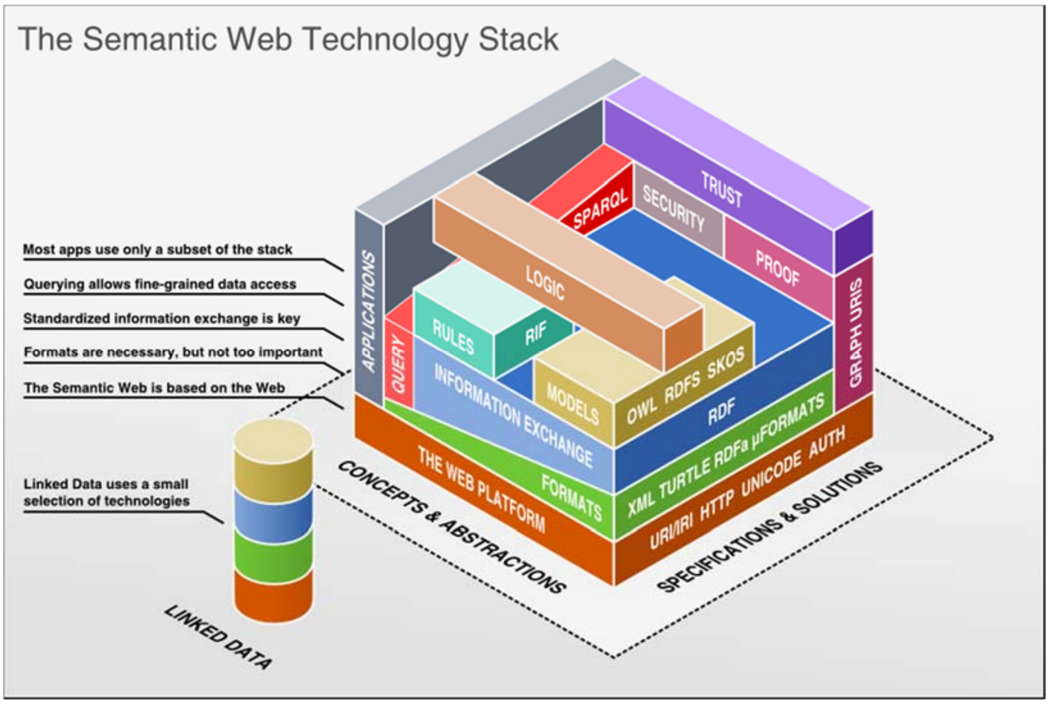
The Semantic Web Technology Stack
Most apps use only a subset of the stack
Querying allows fine-grained data access
Standardized information exchange is key
Formats are necessary, but not too important
The Semantic Web is based on the Web
Linked Data uses a small selection of technologies
LINKED DATA
APPLICATIONS
QUERY
SPARQL
SECURITY
TRUST
PROOF
LOGIC
RULES
RIF
INFORMATION EXCHANGE
MODELS
OWL RDFS SKOS
RDF
GRAPH URIS
FORMATS
XML TURTLE RDFa μFORMATS
THE WEB PLATFORM
URI/IRI HTTP UNICODE AUTH
CONCEPTS & ABSTRACTIONS
SPECIFICATIONS & SOLUTIONS

Fig. 1. The Semantic Web Technology Stack [4]

XML is a markup language widely used for encoding documents so that the data is both human and machine readable. XML is easy and accessible, the data structures represented in XML can be easily edited, while being represented in a manner independent of the application. In addition, they are extensible and the correctness of the data structures is checked by validation enquiries. RDFS and OWL are semantic instruments that represent ontologies. They define classes (concepts) and attributes of URIs (Uniform Resource Identifiers) and the relations between them.

Ontologies are essential tools that generate intelligent semantic web. Ontologies provide a common set of rules, terms and vocabularies that asses how various terms work together. “The purpose of web-based ontology is to provide richer integration and interoperability of data among descriptive communities”. [5]

An URI is a unique symbol of a physical or abstract resource. Any object with clear identity and relevance in the context of the chosen application can be considered a resource. RDF is used to define instances of the ontology as well as relations between the instances by a list of statements. It transposes natural language in machine understandable language, representing any sentence by graph nodes and links between them.

RDFS and OWL define the possible connections between different URI’s by using abstract concepts such as domain, range and relations between classes. Inference rules are created based on these definitions. In Semantic Web, inference is an important tool used to discover new relationships between resources, by automatically analysing the data. [6] As an example we can consider the following statement: Lassie is a dog. A proper ontology has defined that any dog is a mammal (as a class hierarchy definition), thus the system can deduce that Lassie is a mammal. The software module capable of making deductions like the one exemplified above is called reasoner.

SPARQL is an interrogation language, inspired from SQL, which allows the users to extract information from RDF graphs. Besides data extraction and exploration, SPARQL allows operations such as transformation or constructing new RDF graphs from the existing ones.

3 Semantic Web and Cybersecurity

It is essential to identify the borders in where semantic web can improve cybersecurity and where it cannot. In order to do so it is needed to identify patterns of cyber-attacks and black hat hackers’ modus operandi. This can be done by checking the data analysing input available and analyse how it can be used in order to get reliable results (information). Based on input data it can be checked what kinds of inferences are expected for the machines to do.

Germination period

This article explores exclusively the patterns which may link to semantic web technologies. Hong-Mei Chen et al. identifies a “germination period” in an article published in 2017 which presents the promise of proactive approach. Germination period can be explained as ”the time lag between hacker communities discussing software flaw types and flaws actually being exploited”. [7] Its length depends on the difficulty of exploiting the vulnerabilities and the hackers’ interest in it. This is the time when proactive measures can be taken. Hong-Mei Chen et al consider that black hat hackers form “learning communities with unique ecological properties”. They also identify two main categories of data sources available online that can contain data which suits the model proposed in this paper: hacker communities and public security databases. Several studies such as [7] [8] show that hacker communities need information and (continuously) share information among themselves in order to be effective. The collaboration between black hat hackers is based on subtle and indirect ways of work and discussing, presented in [3].

3.1 Expected depth levels

For every type of data source the authors analyse what are the expected types of information. It is required to filter out the noise from the potentially useful data. Therefore, the authors defined 5 categories of expected results, which are described below. In section 3.3 the correlations between these categories and the types of input data described in section 3.2 are emphasized.

a) Intentions/ Targets

This category refers to the intentions and/or the targets of black hat hackers communities The information regarded as such warns the system experts which further take actions to prevent potential attacks. Gathering information about the trends in hacking communities can warn system experts so they can be prepared for future attacks.

b) Attackers' main objectives

The second type of expected results refers to the more concrete purposes of black hat hackers. As an example, a target can be Windows 10 and an objective to try to find vulnerabilities when Microsoft Hyper-V is installed on some Windows 10 versions. As in other fields, experienced hackers develop some kind of educated intuition. Before discovering a vulnerability, they anticipate the possibility of its existence and focus on finding it.

c) Vulnerabilities

As groups, black hat hackers are, among others, learning communities. They share information about vulnerabilities; they discuss and collaborate in order to find out the latest state of most used software systems. Vulnerabilities usually refer to a weakness in a computer system, which generates losses if exploited. [9]

Depending on the stage of a computer system, there are three main categories of vulnerabilities: (1) design, (2) implementation and (3) configuration vulnerabilities. The first refers to the conceptual errors occurring in the first phase of the life of a product and is usually not possible to be removed in the implementation phase. The second type of vulnerability arises at the stage of implementation of the project. The last occurs from incorrect/ineffective system configuration.

d) Black hat hackers collaborations

Exploiting different types of vulnerabilities is often a very complex task which requires collaborations between black hat hackers.

One of the most popular types of cyberattacks is DDOS (Distributed Denial of Service) which consists in making a computer system temporary ineffective due to too many requests. This not only can cause big loses to companies, but also it is usually the first step of an elaborate plan to exploit another vulnerability. In order to be successful on major computer systems, DDOS often requires lots of computation force, thus collaborations between black hat hackers is necessary.

It is very difficult to gather information about black hat hackers' collaborations, because of their ambiguous and indirect ways of collaborating through private means and encrypted language.

e) Potential operating mode

This category refers both to the vulnerabilities and the mechanisms applied in order to exploit them. One key purpose of the solution presented in this paper is to find information about the newest vulnerabilities, especially the ones which are a *zero-day exploit.* When a new method to exploit a vulnerability is discovered, it is considered to be a *zero-day exploit*. The systems which have this vulnerability, as well as antivirus programs which protect them are not able to manage such problems. This is the reason why, as soon as programmers identify a *zero-day exploit,* they create a patch through which the vulnerable application is updated. Zero-day exploits appear on a regular basis, this is why applications are frequently updated.

3.2 Sources for data input

The authors discuss the possible results based on different types of sources available online: forums, chats, blogs and cybersecurity dedicated websites. Black hat communities use "specialty lexicons" classified by Abbasi et al in "general hacker dictionary, technical jargon dictionary and black market dictionary" [3] Several articles such as [10], [3] argue that there is a strong connection between ”vulnerabilities disseminated in hacker communities" [3] and attacks in real life. More than that, studies such as [11], [12] suggest that sharing stolen information and malware is an established habit in black hat hacker' communities.

- Forums

The authors consider forums as one of the main sources of information sharing for hackers (for all three types of hackers: white, black and grey hat). In [3] there are discussed the main reasons that motivate hackers to contribute and to collaborate on forums. Team work, learning and increasing their reputation are the three most common reasons identified by Abbasi et al.

They also analyse a typical hacker forum and manage to classify the actors in 4 main clusters, as shown in table 1 below. The most important actors that can generate valuable input data for cybersecurity ontological interpretation are black market activists. Founding members and technical enthusiasts can produce interesting technical data. Despite the fact that the main purpose of average users is to learn, they can be useful as they tend to share online the new information they acquire.

Forums could generate information for all the depth level classified on 3.2. Black hat hackers may leak their intentions, objectives, discuss new vulnerabilities and potential operating mode.

Table 1. Key hacker groups identified on forums [3]

Groups	Main Interests	Involvement	Description
Black market activists (1%)	Black market business	Very Low	Users with high probability to mention black market keywords
Founding members (1%)	Reputation	High	Founding or old members
Technical enthusiasts (12%)	Learning and team work	High	Technical skilled users
Average users (86%)	Learning	Low to medium	Few technical knowledge

- Private Chats

Private chats represent the online environment where black hat hacker may share with each other information that could lead to all depth levels of results described in section 3.1. However, gathering private chats as input data is very challenging - mainly due to privacy policies and encrypted channels

- Social Networks

The authors have low expectation that social network (S.N.) discussions can generate valuable data source. They identified two main types of data that Social Networks can provide.

The first one refers to the possible connections between black hat hackers. If actors are connected on a social network platform, then there is a higher probability that they collaborate with each other.

The second category refers to, the group identified on forum as "average users", which roughly represents about 86% percent of the total hackers [3] they are mostly learning enthusiasts who willingly share interesting information that they encounter on various hacking channels. This information is usually less valuable for the cybersecurity's purpose, but it is more easily gathered and it can be used as indicator about the activities and trends of the other groups of hackers.

- Blogs

Blogs are generally rich in cybersecurity data, but they provide relatively few information about black hat hacker in progress or future activities. Nevertheless, blogs should be taken into account as a source of learning for all hacking communities, so they can provide information that may help to estimate black hat communities' targets, objectives and computer systems vulnerabilities.

- Cybersecurity dedicated websites

Cybersecurity dedicated websites are useful for hackers both for gaining knowledge and for getting tools. Similarly to the case of blogs, the activity ran on this type of environment can be suitable for understanding the general image about hacking communities' new trends, intentions, objectives, vulnerabilities.

3.3 Correlation between sources for input data and results' expected depth levels

By analysing various types of data, as previously classified, it is expected to obtain specific type of information. Table 2 shows the correlation between input data, discussed in section 3.2 and the possible results, presented in section 3.1.

Table 2. Correlation between sources for input data and **results' expected depth levels**

Depth levels for results / **Sources for input data**	Intention/ Targets	Attackers' main objectives	Vulnerabilities	Black hat hackers' collaborations	Potential operating mode
Forums	🟥✓	🟥✓	🟥✓	🟩✓	🟨✓
Private Chats	🟥✓	🟥✓	🟥✓	🟥✓	🟥✓
Social Networks	🟨✓	🟩✓	🟨✓	🟨✓	×
Blogs	🟩✓	🟩✓	🟩✓	×	×
Cybersecurity dedicated websites	🟩✓	🟩✓	🟩✓	×	×

Legend

✓ = positive correlation | × = no correlation

red colour = high likelihood | yellow colour = medium likelihood | green = low likelihood

Given the previously presented arguments, the authors consider forums, private chats and social networks as main potential data input source. Taking into account the difficulties in gathering data, the main focus will be on forums and social networks and only after on private chats.

4 Web Data Extraction

The World Wide Web (abbreviated WWW) is a massive collection of web pages where new information is continuously added. In this context, search and retrieval of relevant web resources from such a collection can only be effective by automating processes and through the usage of intelligent agents. Search engines use such software agents to index the Internet, providing users with search abilities based on various criteria. These programs are called web crawlers or simply crawlers.

A web crawler is a software program that searches and extracts data from web pages, navigating from URL to URL, according to predefined algorithms. These types of programs also bear the name of bot, robot, web robot, spider, etc.

In terms of crawling data, web crawlers can be classified into two categories: traditional crawlers and focused (or topic) crawlers. Traditional crawlers aim to identify and index web pages regardless of their specificity. They do not have the ability to distinguish between relevant and partial relevant web pages. Because of this, traditional crawlers extract a large amount of data and often a big part of it proves to be irrelevant to users. [13]

On the other hand, topic crawlers are agents that collect web pages that satisfy certain specific properties. They offer the possibility of downloading relevant web documents for a predefined domain, providing the most up-to-date resources (web pages) relevant to the needs of users, with minimum consumption of resources such as storage, time and network bandwidth. [14]

Web crawlers are often confused with web scrapers due to their similar functionality. Web scrapers are intelligent agents, but they show some differences from crawlers, as illustrated in Table 3.

In computer science, a parser is a software program that receives input data as sequential instructions, interactive commands, tag labels, or other defined interface, and separates them into parts (e.g., nouns, verbs and their attributes or options). Based on these, it builds a data structure, abstract syntax tree, or other hierarchical structure therefore providing a structural representation of the input. These are further managed and analysed by other programs, such as other components in a compiler. [15] In the present paper, it is desired to include data extracted by the parser into ontologies, following the object-property-object structure described in the previous section.

A web scraper can be defined as a software program that extracts and combines web

DOI: 10.12948/issn14531305/21.3.2017.01

content in a systematic way. In such a process, a software agent, also known as a web robot, imitates web browsing interactions between web servers and humans in an automated way. Step by step, the robot accesses as many websites as needed, analyses their content to find and extract data of interest, and organize them appropriately. [16]

Table 3. The main differences between a crawler and a scraper

	Crawler	**Scraper**
Data source	WWW	Various sources, including the WWW
Types of smart agents required	Crawler	Crawler and parser
Deduplication	It is a mandatory component	It is not an essential component
Submitting form with data	No	Yes
JavaScript code execution	No	Yes
Scalability	Used mainly for large scale	Used at any scale
Transforming data (form and format)	No	Yes
Saving data into database	No	Yes

A focused crawler, designed to retrieve text-based pages in a given domain, can use a predefined ontology. Ontology defines or specifies concepts, relationships, and other distinctions that are self-explanatory for modeling a domain. In order to describe semantic relations between different terms or concepts, ontology provides an effective solution. Applications such as eLearning need to specify relationships between concepts for building the knowledge base. [17]

There are two categories of focused crawlers: classical and learning-oriented. Classic focused crawlers are only based on predefined sets. Unlike them, learning-oriented subject crawlers can automatically develop new rules by integrating the results obtained in their collection.

Ontologies are used by crawlers as tools for understanding the text. Through them, a crawler can become a classic topic crawler. If the ontologies used are dynamically enriched, as the robot identifies new classes or instances, then it becomes a learning object-oriented crawler.

Ontologies form the basis of determining the semantic relationship between different concepts and are used to calculate the empirical semantic distance between each pair of concepts. Through domain ontologies, the crawler identifies terms linked on conceptual level, structured terms in the form of concept maps. Bedi et al note that, as a rule, technical terms do not have synonyms, antonyms, hyponims or hypernims (concepts on which lexical databases such as Wordnet [18]) They are based on related concepts, sub-concepts, super-concepts, etc. [17]

Semantic crawlers are classic topic crawlers that determine the relevance of the web page by using the knowledge base. However, they can also be extended to learning-oriented crawlers, having the disadvantage of increased processing time.

5 Framework description

In this section, the authors propose a model that can improve cybersecurity by using semantic web technologies.

In section 5.1 the authors outline the

requirements for the Cybersecurity solution. Section 5.2 describes how data will be found and gathered from the online sources. The authors intend to use semantic web technologies for data mining as described in section 5.3. Section 5.4 presents how data will be stored and processed. Section 5.5 illustrates the main results that the application should provide. On 5.6 the general image of the solution can be observed.

5.1. Requirements for the Cybersecurity solution from the user perspective

In order to properly design a system, it is important to start from the functionalities that the system will provide. Table 4 illustrates the main functionalities provided by the solution proposed from user's perspective.

Table 4. Main functionalities provided by the solution proposed from the user's perspective

Functionality	Description
Access to cybersecurity data	The users can access cybersecurity data organized by an index
Information grouped by different sets of categories	The data is labelled by different sets of clusters
Automatic alerts/notifications	Any relevant information discovered by the solution automatically shown as a notification
Access to the URLs from which data is extracted	The user can easily find the source of information, the analysed data and the source of data.
SPARQL interrogations	The user can update and modify ontologies by using SPARQL interrogations
Transform and update data	Data stored can be transformed and updated
Statistical analysis reports	Statistical analysis reports are possible

5.2 Data scraping

A web scraper is recommended for gathering data, based on HTML parsing and semantic annotation recognition. The scraper extracts data based on established instructions and on ontologies, thus diminishing the data volume. The web pages that are scraped may include metadata, annotation, semantic markups. A filter based on ontology domains and ranges is applied to adequately interpret and incorporate the data. The web scraper stores gathered data as JSON files, because they can be easily stored on a Hadoop system where data is extracted and introduced in ontologies. Dastidar et al identify the main web data extraction methods: "HTTP Programming, HTML Parsers, DOM (Document Object Model) and SAX (the Simple API for XML) parsers, Open Source web scraping libraries, semantic annotation recognizing, NLP (Natural Language Processing) recognition of keywords." [19]

5.3 Cybersecurity ontology

High quality cybersecurity ontology is an essential component for the framework proposed. The authors intend to combine existing cybersecurity ontologies and to improve them. Syed et al describe such ontologies which they conclude that they "provide a common understanding of cybersecurity domain and unifies most commonly used cyber-security standards". [2]
Besides that, the cybersecurity ontology required for the framework proposed in this paper should also incorporate specialty lexicons, such as technical jargon and black market dictionaries.

As mentioned before, the data needs to be filtered in order to decrease its volume. This can be done by using cybersecurity ontologies which act as first data filters and identify the potential online source.

5.4 Data storage and processing

The data gathered by the scraper will be stored in Hadoop. Ontologies will analyse the data and try to incorporate as many instances as possible. Data mining techniques are used to identify and analyse user's behaviour in the online environment. The main applications of data mining are: classification, regression, clustering, summarization, association. [20]

The data will be used for two types of processing:

- **Ontology transformation and reasoning**

The first one refers to inferring logical consequences based on a set of facts or axioms. Such an example can be checking for similar URI and relations between URIs or identifying if more URIs represent the same thing.

- **Clustering**

Different sets of clusters will be applied. Clustering can be based on types of actions, one of the main clusterization criteria is the type of results expected, defined in section 3.1.

5.5 Results

The system will provide notifications as output data. A notification is an RDF type sentence, followed by the text source that generated that specific data for ontology and the URLs from where that data was downloaded. The notifications have attached labels and can be categorised as previously described in section 3.1.

The notifications are checked and solved by an expert who manages them individually. The majority of them are expected to be invalid. In order for the system to be considered reliable a percentage of valid notifications has to be determined. In their future work the authors will establish this percentage and therefore create means of checking the reliability of the system.

According to the degree of danger, the notifications can be classified as follows:

- *High risk* – labelled with red, this type of result will have the highest priority and will be displayed automatically on top; these alerts should be treated as soon as possible by the expert;
- *Medium risk* – labelled with yellow, are the data sources that present a medium degree of danger;
- *Not Applicable* – labelled with blue, are the sources that are suitable as input data, but the program is not able to properly analyse them.

The input data identified as harmless is labeled with green and appears in a special section, where it can be used in the future. Whereas the experts identify invalid notifications, they can delete it permanently or mark it with green label and keep the data in the system.

5.6 Proposed framework architecture

In Figure 2, the framework architecture shows the overview of the model and the data flow previously described.

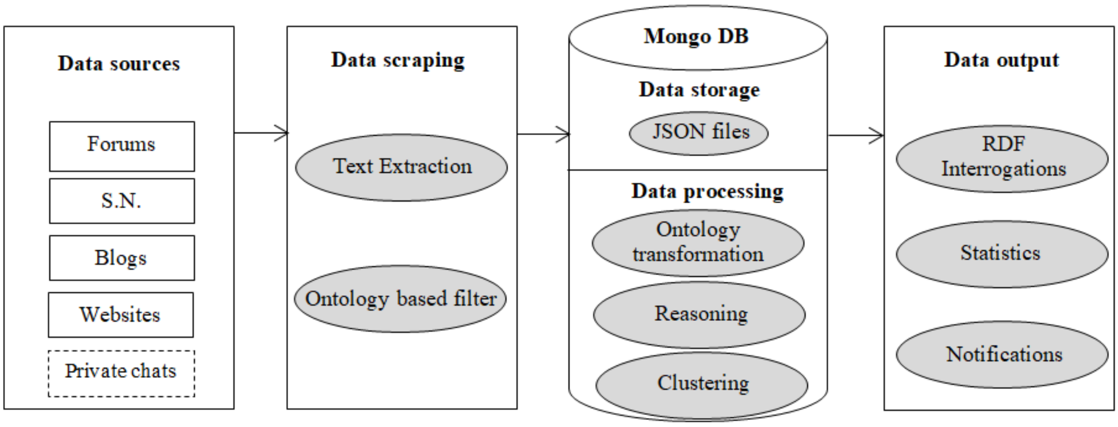

Fig. 2. Framework architecture

The framework addresses five types of online data sources, as described in section 3.2. The data extraction methods should be fitted accordingly to the type of source. Thus, before extracting data, the web scraper identifies the source type. The ontology based filter is an important tool for the solution presented by this paper. It allows the scraper to identify possible sources, not only by keywords, but through context understanding.

Data extracted is first stored as JSON files. These types of files can be easily integrated on both web scrapers and Hadoop systems. Moreover, it is possible to extract data from JSON files and incorporate it in cybersecurity ontologies. The main data processing functions of the solution are ontology transformation, reasoning and clustering. Based on these functions, the authors expect the application will be able to identify valuable results.

The most important/urgent results are displayed as notifications/alerts as soon they as they appear. An expert will use the solution and manually analyse which ones are valuable and which are noise. The expert can also perform different functions through the solution like SPARQL Interrogations or check statistics based on gathered data.

7 Conclusion and future work

The framework described in this paper can be a solution to improve cybersecurity bwith the aid of semantic web technologies. One of the main problems in cybersecurity field is that most of the actions are taken as a response to attackers' operations only after the attack had occurred. The framework can be considered a proactive solution against black hat hackers' activities.

The paper also discusses the main types of online sources that can serve as data for the framework, as well as the kind of information that can be obtained after analyzing and interpreting the discussions of black hat hackers' communities in the online environment

This is the second paper of the authors that analyses how semantic web technologies can be used in cybersecurity. In their future work, the authors plan to build a software system based on the framework described. They also continue to investigate how semantic web technologies can improve cybersecurity field.

Acknowledgment

Parts of this article have been previously published in the Proceedings of the 16th International Conference on Informatics in Economy, 2017.

References

[1] Berners-Lee, Tim, J. Hendler and O. Lassila, "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," Scientific American, vol. 284, no. 5, pp. 1-5, 2001.

[2] Z. Syed, A. Padia, M. L. Mathews, T. Finin and A. Joshi, "UCO: A unified cybersecurity ontology," in Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security, 2016.

[3] A. Abbasi, W. B. Li, H. S. V. A. and H. Chen, "Descriptive Analytics: Examining Expert Hackers in Web Forums," in Joint Conference in Intelligence Security Informatics, 2014.

[4] B. Nowack, "BNODE," [Online]. Available: http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake. [Accessed 28 4 2017].

[5] L. Băjenaru, A.-M. Borozan and I. Smeureanu, "Using Ontologies for the E-learning System in Healthcare Human Resources Management," Revista Informatică Economică, vol. 19, no. 2, pp. 15-24, 2015.

[6] The World Wide Web Consortium, "W3C," [Online]. Available: https://www.w3.org/standards/semanticweb/inference. [Accessed 10 4 2017].

[7] H. M. Chen, R. Kazman, I. Monarch and P. Wang, "Can Cybersecurity Be Proactive? A Big Data Approach and Challenges," in Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.

[8] V. Benjamin, W. Li, T. Holt and H. Chen, "Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops," in Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on, 2015.

[9] S. L. Pfleeger and C. P. Pfleeger, Analyzing Computer Security: A Threat/ Vulnerability/ Countermeasure Approach, New Jersy: Prentice Hall Professional,

2012.

[10] Q. H. Wang, W. T. Yue and K. L. Hui, "Do Hacker Forums Contribute to Security Attacks?," in Workshop on E-Business. Springer Berlin Heidelberg, 2011.

[11] "Exploring stolen data markets online: products and market forces," Criminal Justice Studies, vol. 23, no. 1, pp. 33-50, 2010.

[12] T. J. Holt, D. Strumsky, O. Smirnova and M. Kilger, "Examining the social networks of malware writers and hackers," International Journal of Cyber Criminology, vol. 6, no. 1, 2012.

[13] K. M. A. Khan and D. K. Sharma, "Self-adaptive ontology-based focused crawling: A literature survey," in Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 5th International Conference on (pp. 595-601). IEEE, pp. 595-601, 2016.

[14] A. Thukral, V. Mendiratta, B. Abhishek, H. Banati and P. Bedi, "FCHC: A Social Semantic Focused Crawler," in Advances in Computing and Communications, Kochi, India, 2011.

[15]"http://searchmicroservices.techtarget.com/definition/parser," [Online].

[16] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato and F. Fdez-Riverola, "Web scraping technologies in an API world," Briefings in bioinformatics, vol. 15(5), pp. 788-797, 2014.

[17] B. Punam, A. Thukral and H. Banati, "Focused crawling of tagged web resources using ontology," Computers & Electrical Engineering, pp. 613-628, 2013.

[18] Princeton University, "Wordnet," [Online]. Available: https://wordnet.princeton.edu/. [Accessed 26 05 2017].

[19] B. G. Dastidar, D. Banerjee and S. Sengupta, "An Intelligent Survey of Personalized Information Retrieval using Web Scraper," International Journal of Education and Management Engineering, vol. 6, no. 5, pp. 24-31, 2016.

[20] Fatima and J. I. Khan, "Classification Of Data Mining Techniques & Tools: A Suvery," International Journal of Innovative Research and Advanced Studies (IJIRAS), vol. 3, no. 13, 2016.

Tiberiu-Marian GEORGESCU has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2012. In 2015 he has graduated the Informatics Systems for the Management of Economic Resources Master program. Currently he pursues a PhD research in Economic Informatics at the Bucharest University of Economic Studies, under the guidance of professor Ion SMEUREANU, PhD. He is working as a Teaching Assistant in the Department of Economic Informatics and Cybernetics. His main interests in the Informatics field are Cybersecurity and Semantic Web.

Ion SMEUREANU has graduated the Faculty of Economic Cybernetics in 1980, as promotion leader. He holds a PhD diploma in "Economic Cybernetics" from 1992 and has a remarkable didactic activity since 1984, when he joined the staff of Bucharest University of Economic Studies. Currently, he is a full Professor of Economic Informatics within the Department of Economic Informatics and Prorector of Bucharest University of Economic Studies. Also, he was the Dean of the Faculty of Cybernetics, Statistics and Economic Informatics. He is the author of more than 16 books and an impressive number of articles. He was also project coordinator or member in many national and international research projects. He was awarded the Nicolae Georgescu-Roegen diploma, the General Romanian Economist Association Excellence Diploma and many others.

DOI: 10.12948/issn14531305/21.3.2017.01