Cyberthreat risk identification based on constructing entity-event ontologies from publicly available texts

Michael K. Ridley, *MAI, Russian Federation, Moscow*
mr@kalabi.ru

Michael K. Ridley

Abstract. Aim. *Out of the currently used methods of ensuring cyber security the most productive ones are traffic analysis, malware detection, denial of unauthorized access to internal networks, incident analysis and other methods of corporate perimeter protection. The efficiency of such methods however depends on the timeliness and quality of threat data. The Aim of the paper is to study the ways of improving the cyber threat awareness and capabilities to analyze texts in open sources for the purpose of cyberattack prediction, identification and monitoring of new threats, detection of zero-day vulnerabilities before they are made public and leaks are discovered.* **Methods.** *Publicly available knowledge on cyber security is acquired through continuous collection of data from the Internet (including fragments of its non-indexed part and specialized sources) and other public data networks (including a large number of specialized resources and sites in the TOR network). The collected texts in various languages are analyzed using methods of natural language processing for the purpose of extracting entities and events that are then grouped into canonical entities and events, and all of that information is used for continuous updating of a subject-matter event-entity ontology. It includes general forms of entities and events required for the context and specialized forms of events and entities for purposes of cyber security (technical identifiers, attack vectors, attack surfaces, hashes, identifiers, etc.) Such ontology can function as a knowledge base and be used for structured queries by cyber security analysts.* **Results.** *The proposed method and the system based upon it can be used for analyzing computer security information, monitoring, detection of zero-day vulnerabilities before they are made public and leaks are discovered. The information retrieved by the system can be used as highly informative features in statistical models. The latter served as the basis for a classifier that defines the risk of exploits for a specific vulnerability, as well as an IP address scoring system that can be used for automatic blocking. Additionally, a method was developed for risk-based ranking of events and entities associated with cyber threats that allows identifying – within the abundance of available information – the entities and events that require special attention, as well as taking timely and appropriate preventive measures.* **Conclusion**. *The proposed method is of direct practical value as regards the problems of analytics, risk-based ranking and monitoring of cyber threats, and can be used for the analysis of large volumes of text-based information and creation of informative features for improving the quality of machine learning models used in computer security.*

Keywords: *cybersecurity, railway infrastructure security, knowledge extraction, semantic web, ontology, natural language processing.*

For citation: *Ridley M.K. Cyberthreat risk identification based on constructing entity-event ontologies from publicly available texts. Dependability 2020;3: 53-60. https://doi.org/10.21683/1729-2646-2020-20-3-53-60*

Received on: *23.06.2020* / **Revised on:** *18.07.2020* / **For printing:** *21.09.2020*

Introduction

Over the last decade, cybercrime made a quantum leap and became a highly competitive market. In 2016, its direct damage to the global economy amounted to 3 billion dollars, while in 2020 this figure was as high as 6 billion dollars. The sum is growing along with the rate of digitization: the higher the number of automated systems, the more there are ways of disrupting the activities of a business. For instance, during six weeks of its operation, a little-known German project HoneyTrain that simulated railway infrastructure management systems was exposed to 2.3 million attacks [1].

Attacks against railway infrastructure are often aimed at client services. For instance, in May 2018, a DDoS attack (a distributed attack for the purpose of causing a denial of service) prevented the passengers of the Danish State Railways (DSB) from buying tickets both online and from fixed terminals. Attacks against control systems happen more rarely, but they are more hazardous. For instance, in October 2017, the Swedish transportation system temporarily lost the ability to monitor the location of the railway rolling stock and geoinformation services. Additionally, attacks against SCADA (supervisory control and data acquisition) systems happen as well. Among the examples are the Stuxnet infection of a uranium-manufacturing facility in Iran or attack against a nuclear power plant in Kansas, US.

JSC RZD is making active efforts in terms of cyber security of railway infrastructure. For instance, 2016 saw the beginning of a joint project by JSC RZD, Positive Technologies and JSC NIIAS[1] that examined the EBILock 950 interlocking system that, through object controllers, controls such trackside devices as level crossings, track circuits and point machines.

Importantly, JSC RZD employs versatile technology. The sixteen railways operated by JSC RZD use different types of equipment and protocols. At the top level, about 100 automated management systems are used, while at the bottom (local) level tens of thousands computer-based traffic management systems of almost 70 different types are in operation [2].

Currently, the focus is on ensuring protection and repelling known attacks. For instance, the above project resulted in the deployment of the Positive Technologies Industrial Security Incident Manager. That was of course the right thing to do, as in 2018 alone the number of components of process control systems available online grew 1.5 times, as well as the number of vulnerabilities that can be used remotely without privileged access.

Nevertheless, perimeter defense and traffic analysis are not sufficient. The system must insure real-time notification of new vulnerabilities, cyberattacks occurring worldwide (including planned ones), hacktivist activities, attacks against other systems of similar class, etc. Such monitoring is on the list of important recommendations for railway infrastructure facilities [3]. For example, it allows reacting to a vulnerability identified in a used software solution a week before its official publication.

As it is known, open sources often report information on vulnerabilities and exploits before they are featured in the common databases, the CVE (Common Vulnerabilities and Exposures) and NVD (National Vulnerability Database), whereby the time gap may be as long as several months [4]. Such information appears in open source software issue trackers, on Twitter, subject-matter blogs, Q&A services for software designers like StackOverflow, e-mails, hacker forums and trading sites in anonymous networks. For the purpose of efficient monitoring, computer security analysts need methods of automatic retrieval of information from texts in public networks, including some unindexed segments of the Internet and anonymous networks like TOR. Thus, the monitoring will allow not only identifying new threats and cyber security risks, but analyzing and scoring threats in a more complete and systematic way.

Monitoring such source is a way of automatically tracking zero-day vulnerabilities that are especially hazardous and often remain unnoticed in network traffic analysis. Such vulnerabilities include those, for which there are no developed or published protection mechanisms, which allows intruders to freely use them until the fix is published, as well as interferes with the protective features' efforts to detect the attempts to exploit them. That is also an efficient tool for real-time search for information leaks similar to the one that occurred in June 2019, when hundreds of thousands of documents from JSC RZD's corporate resources were stolen[2]. As intruders often look for ways to monetize the stolen assets, they place announcement in special sites, where the leak will be discovered several minutes after the publication.

Construction of entity-event ontologies for applied knowledge bases

One of the methods of conceptualization of text-based information is the construction of an ontology based on the described facts. Such ontologies include domain concepts, their relations and their attributes. They are extracted in accordance with the meta-ontology (top-level ontology that describes a specific ontology) using computer linguistics tools, Hearst pattern-type and regular expression matching rules, statistical models.

Automatic ontology construction is most often used for creating universal ontologies that are based on linguistic categories of the type hyponym/hyperonym and meronym/holonym, IS-A, INSTANCE-OF and other relationships. That is relevant in the context of many problems related to artificial intelligence, but not the problems associated with the representation and collection of applied knowledge.

For the purpose of knowledge acquisition, it is more practical to use self-extending entity-event ontologies. The author has previously developed an analytics system that extracts texts from sources, analyses them, builds an ontol-

[1] Source: https://bit.ly/2YkAQ4N

[2] Source: https://www.kommersant.ru/doc/4252728

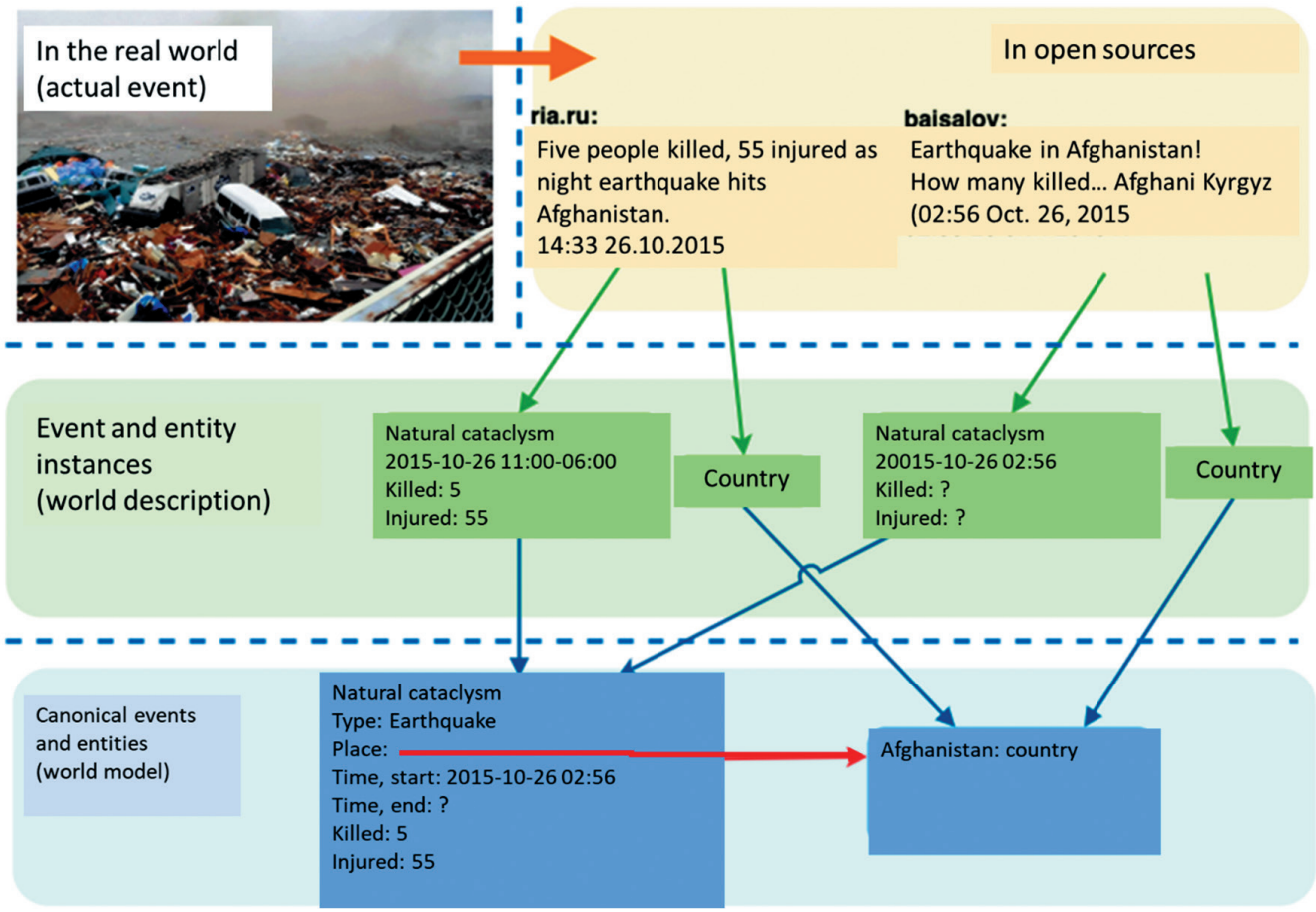

Fig. 1. An example of description of real phenomena through canonical events and their mentions

ogy and provides it to the user as a knowledge base [5]. The system was developed for news and political applications, as well as for commercial aviation [6]. One of the aims of this paper is to adapt the system for the purpose of analyzing and monitoring cyber threats.

The approach involving entity-event ontologies implies that the world is modeled by means of separating documents from their contents, i.e. the canonical entities and events that correspond to real people, technologies, companies, meetings, business transactions, attacks and political events. Each canonical entity and each canonical event may be associated with a number of instances that relate the mentions of canonical objects in texts with the time, place and other contextual information, as illustrated in Fig. 1. Such ontology allows for structured queries to the knowledge base. One can find out the technical identifiers associated with an attack, obtain the list of sources that referred to an attack, get to know the types of systems that can be attacked using a specific exploit, who and where is selling a specific set of exploits and other matters that come down to attribute-based filtering.

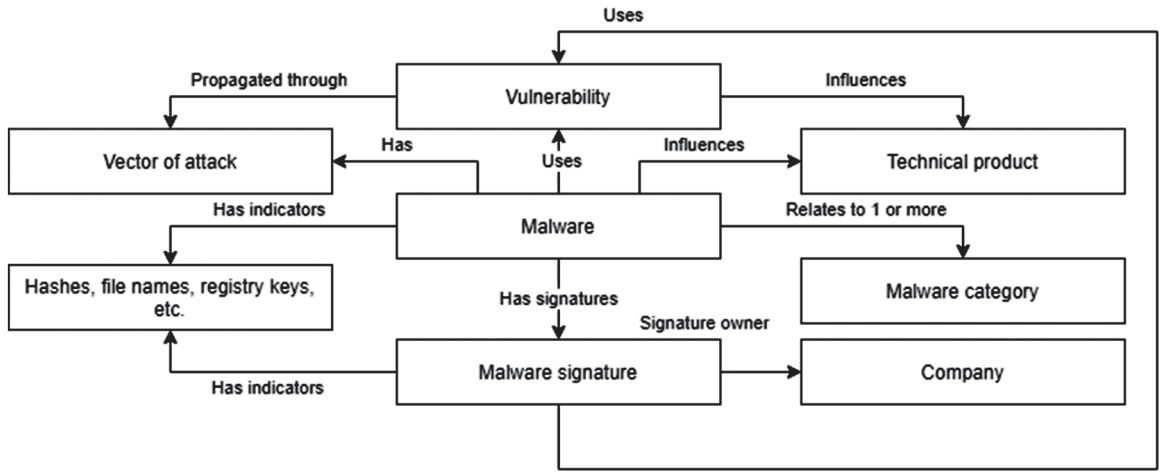

Fig. 2. A fragment of a meta-ontology of cyber security entities

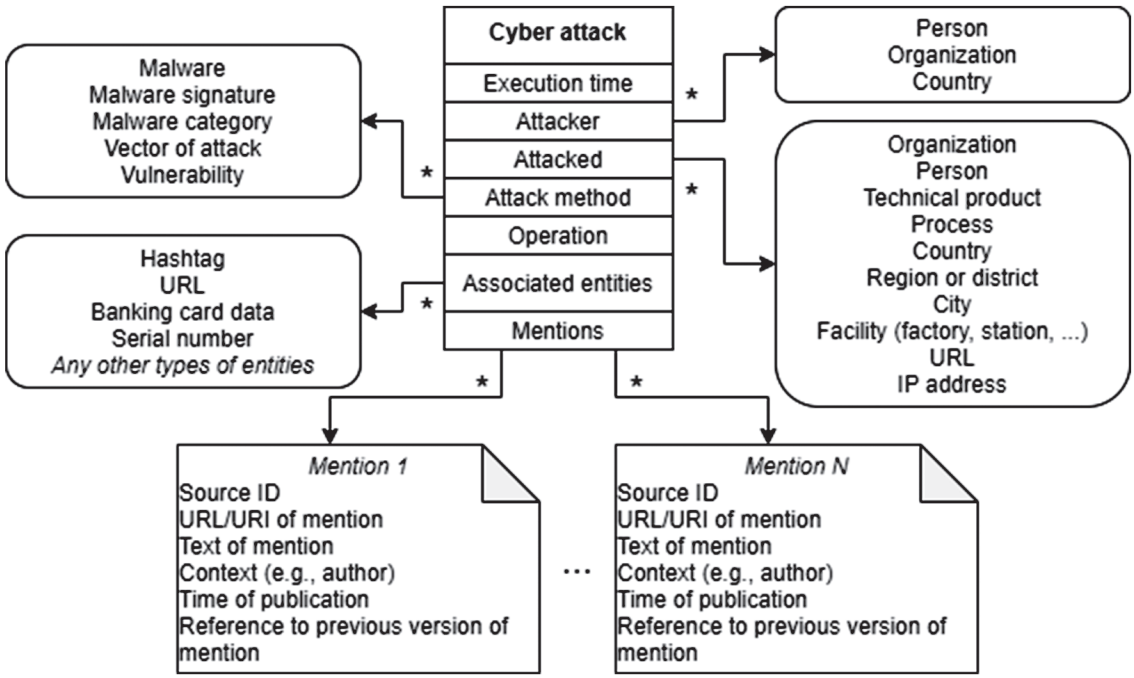

Fig. 3. Event scheme of type "Cyber attack"

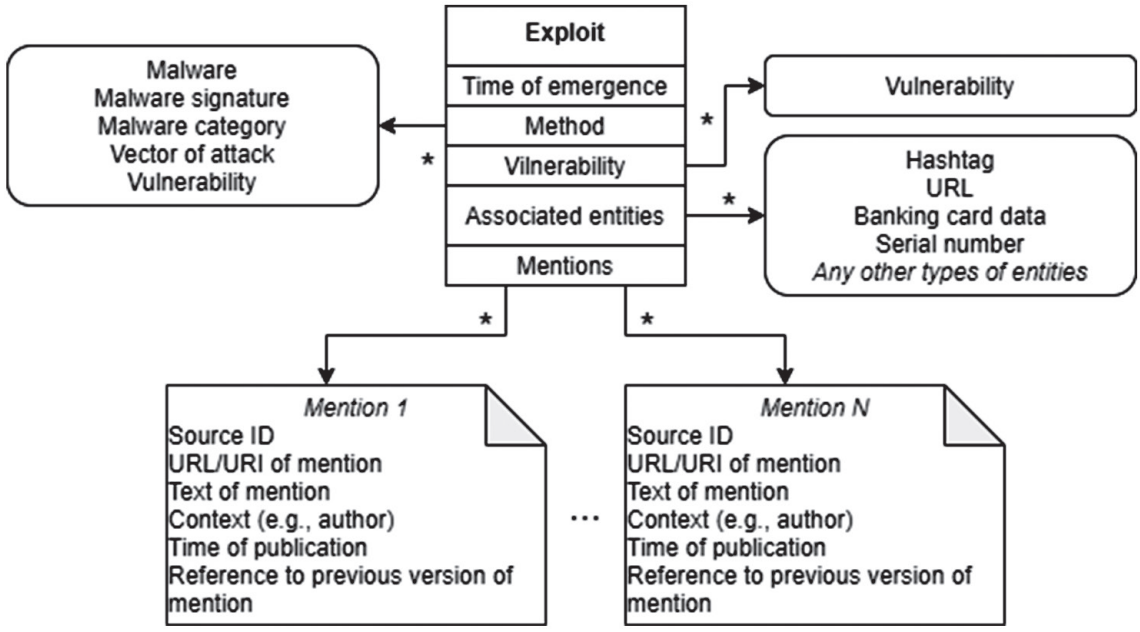

Fig. 4. Event scheme of type "Exploit"

The meta-ontology that supports the continuous updating of the entity-event ontology based on text references was completed to suit the needs of cyber security. The cyber security ontology for critical infrastructure was used as the foundation. It was revised to be made compatible with events and entities [7]. In terms of entities, the meta-ontology is illustrated in Fig. 2, in terms of events, it is shown in Fig. 3 and 4.

Implementation of text analysis for construction of entity-event ontologies

Out of various natural language texts it is required to extract information on entities (individuals, organizations, card numbers, mailing addresses, hashes, IP addresses, software signatures, file names, etc.) and events (company mergers, political protests, cyberattacks, bankruptcies, etc.).

The system uses five ready-made natural language processing tools (Table 1) and a set of own tools for exotic entities (hashes, serial numbers, vulnerability codes, code fragments, etc.). The developed extraction tools use regular expression rules or the conditional random field (CRF) method, whose specificity consists in the absence of necessity to model the probabilistic dependencies between the observable variables and the problem of marker shift unlike in a maximum-entropy Markov model.

Table 1. Employed ready-made computational linguistics tools

StanfordNLP	Tomita parser	OpenCalais	OpenNLP	Rosette EX
Supported languages				
English, German, Spanish, Chinese	Any, defined by dictionaries and grammars	English	European languages	55 languages (including Russian, Arabic and Chinese)
Primary interfaces				
API, JAVA and Python libraries, web interface	Console-based application, API	API, web interface	JAVA library	API, web interface
The best developed branch of functionality (used in the system)				
Entity extraction using statistical algorithms and neural networks	Generation of grammar and entity extraction using dictionaries	Entity and event extraction based on the news ontology	Entity extraction using statistical algorithms and neural networks	Entity, fact extraction, coreference resolution
Output methods recommended by the developer				
JSON, XML, CoNLL, graphic	Output format is defined by the grammar	RDF, XML, graphic	XML	RDF, XML, graphic

The extracted entities and facts are compared and resolved according to the ontology in order to specify their meaning and resolve the coreference. Ontologies of structured relationships between entities are used for managing the process of filtering and as the gazetteer for improving the extraction process.

Having acquired a set of filtered entities, the module responsible for proposition extraction associates the entities with the events mentioned in the document. The events are assigned fragments of text out of the document that provide the best possible short description, while also ensuring summarization. For each fragment, the sentiment is analyzed.

Next, the facts are submitted to temporal analysis. Cultural and regional categories are extracted from a document in order to take into consideration the hemisphere, first day of the week and date format. The events may be either past or anticipated. The future events are either planned (election

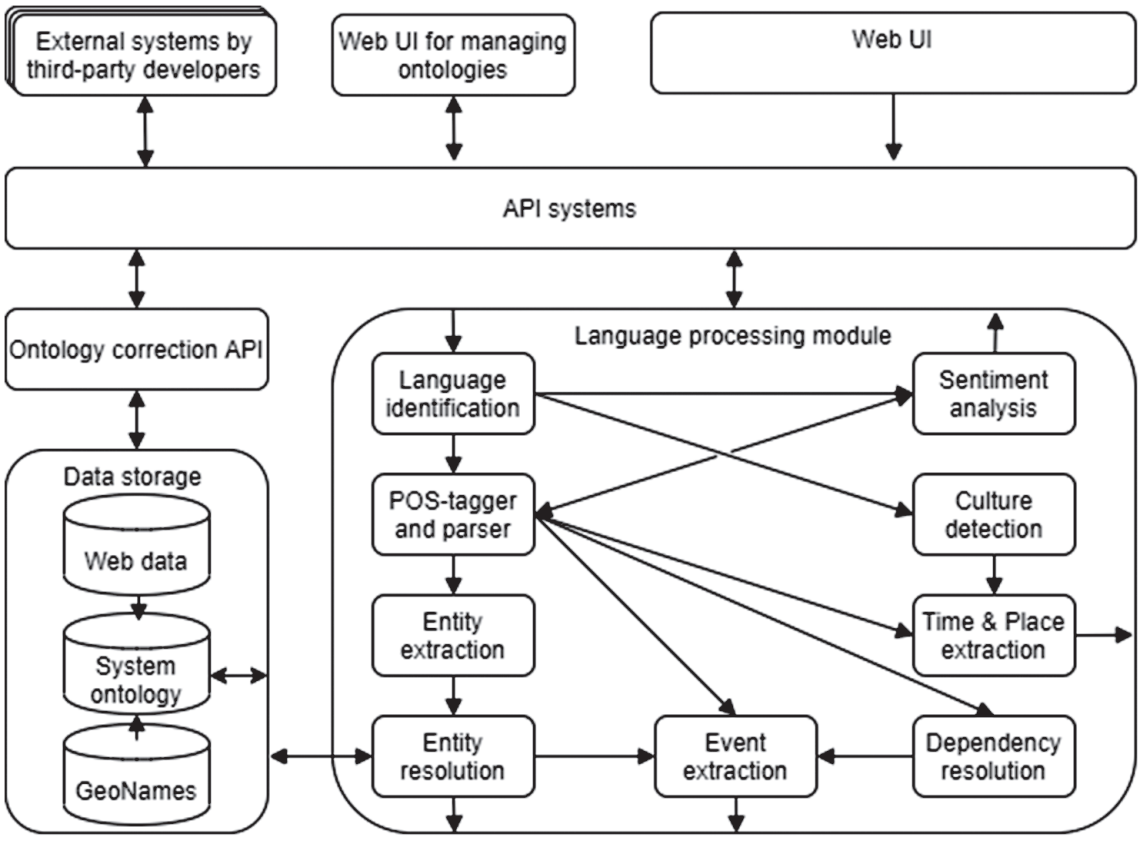

Fig. 5. Top-level architecture of the language data analysis subsystem

days), or speculative (assumption of when a rally should be called). For instance, that enables anticipating actions of hacktivists like Anonymous and release of patches by software manufacturers.

The events are marked with the type, time interval, involved entities, their roles (which attributes of an event they are assigned to), sentiment and source. The layout of the architecture of the language data processing subsystem is shown in Fig. 5. It is an independent system integrated with the main system that is examined in the following section of the paper.

Architecture and implementation of the ontology-based data collection and analysis system

Figure 6 shows the top-level system architecture.

The collection subsystem is responsible for the acquisition of text-based data from the Internet. As the input, it receives a list of predefined and configured resources and regularly retrieves data from them. The system outputs a flow of unstructured texts with the indication of the time of collection, context and source.

The input of the *language data processing subsystem* consists of cleaned texts from the collection subsystem. The subsystem outputs a list of snippets (fragments of text) marked-out in XML with the entities, events, time and place stamps highlighted. This is the only language-dependent place within the system. Providing support for a new language requires the development of a new module within the subsystem, while the system as a whole operates with either "raw" texts, or language-independent facts.

The input of the *storage subsystem* consists of marked-up snippets that it processes, upon which the extracted facts are stored. It also receives requests from external systems. The fact storage is accessed through an application program interface (API) in the JSON format in accordance with the REST API principles. The subsystem outputs ontology slices (sets of facts, events and entities, as well as their metrics of importance type) corresponding to the API request.

The fact storage is viewed and modified by the modules of the *data integration subsystem* that enrich and refine the data. This subsystem attenuates improbable events and amplifies the high-profile events, simultaneously enriching them with additional structure and forming canonical events and entities.

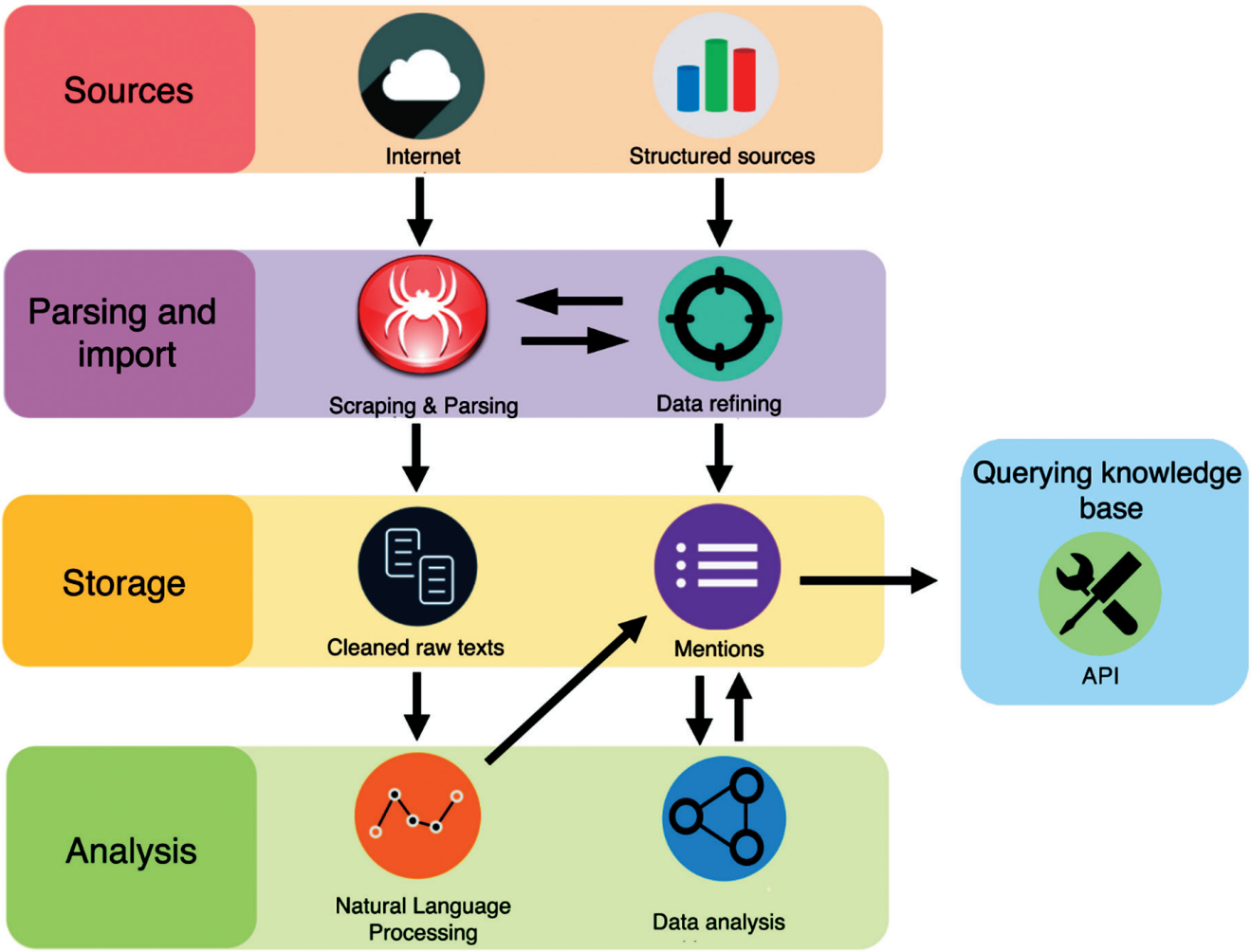

Fig. 6. Top-level operating diagram of the system

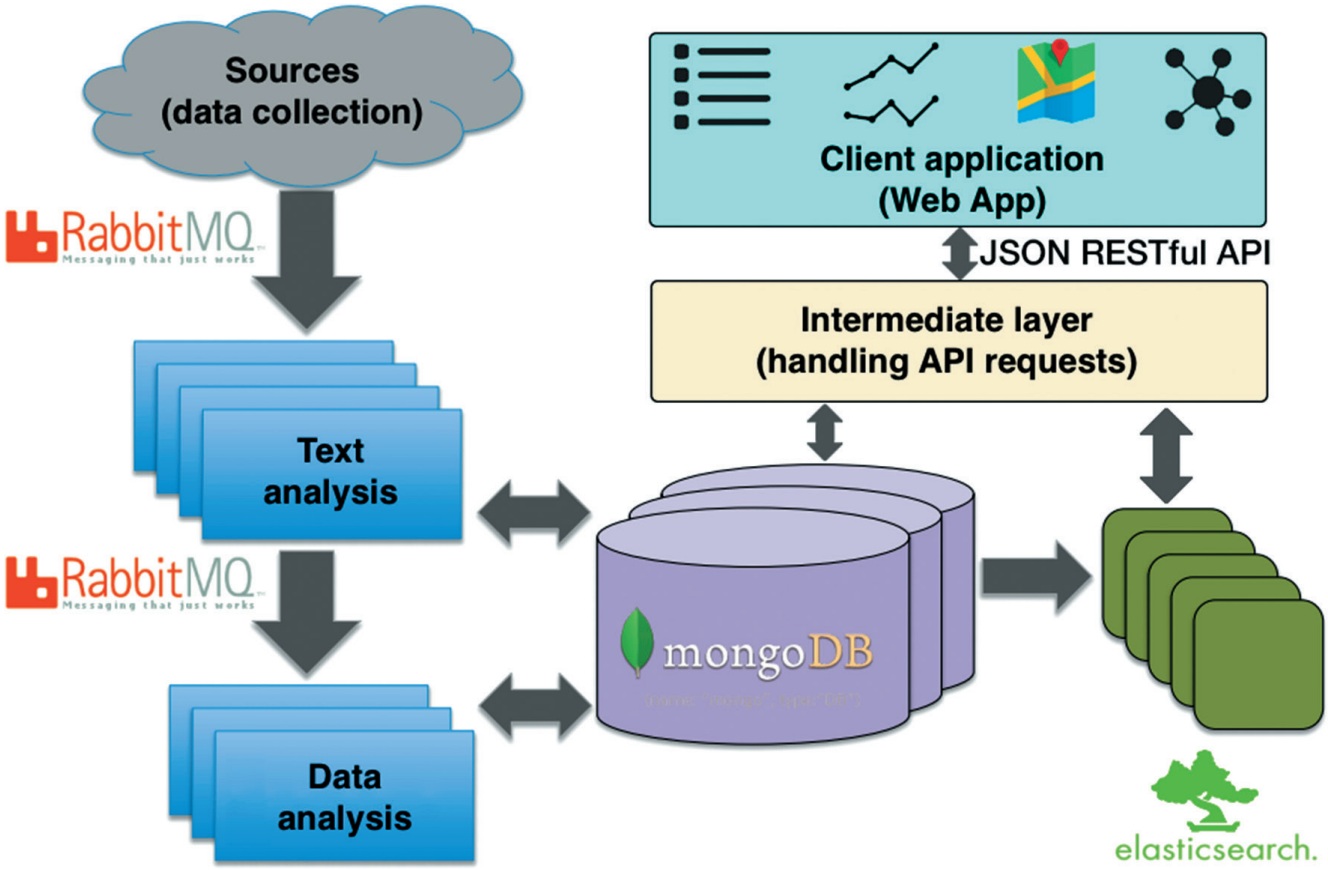

Fig. 7. Technical architecture of the system in terms of data flow

The system is distributed and has a microservice architecture (Fig. 7). The components communicate using RabbitMQ message queues generating about 4000 messages a second. The knowledge base is stored in a MongoDB NoSQL storage (9 shards with read and write roles), while for the snippets, the ElasticSearch full-text search system is used (7 shards). Metadata for a hundred documents take 1 Mb on average.

Results of the method's application as part of applied problems

Monitoring and cyber security analytics, prediction.

The generated entity-event ontologies are directly used for predicting hacktivist attacks, detecting zero-day vulnerabilities and searching for leaks. Another direct use is the analysis of computer security information: what entities and through what are associated with a set of exploits, what methods a specific group uses, what industries are currently threatened, etc.

Early detection of zero-day vulnerabilities.

It has been identified that 77% of vulnerabilities out of the list of CVE (Common Vulnerabilities and Exposures) in Linux were known before they were made public as zero-day vulnerabilities, while the average delay between the first mention and the date of official publishing is 19 days. Additionally, all the vulnerabilities featured in CVE could be found on Twitter [8].

Creation of highly informative features for machine learning.

The obtained and continuously updated knowledge base of cyber security facts can be used for creating highly-informative features in machine learning models as shown in the following problem.

Definition of the risk of an exploit for a vulnerability.

Using supervised training based on support vector machines, a classifier has been obtained that, for each specific vulnerability, predicts whether an exploit will be created with the accuracy of 0.79 and completeness of 0.80. A balanced learning sample included 7000 examples of vulnerabilities. The classifier predicts the risk of a specific vulnerability being exploited and suits the purposes of prioritization of activities aimed at developing countermeasures, emergency isolation of vulnerable systems in case of high risk, etc.

IP address scoring

IP address scoring allows cyber security analysts making decisions regarding further analysis, and, in high alert situations, automatically blocking IP addresses in order to complicate access by the attackers. Importantly, the cybercrime market is commoditized and provides botnets and specialized infrastructure for attacks for rent. Therefore, in practice, blocking high-risk IP addresses is quite efficient, especially against DoS attacks aimed at causing the denial of service by systems.

Risk-based ranking of cyber security events and entities.

The risk of an entity or event is calculated based on the reference dynamics, presence of significant targets and diversity of a language in the mentions. All features

are calculated using the moving average for the estimated entity, as the mentions often occur in "spikes" (day-night, week day-week end, etc.). The level of criticality for entities is based on the number of references to events of cyberattacks/exploits occurring today or within the next month and including the entity. The overall scope of references of a certain entity does not affect the level of criticality: small spikes and anomalies have a more significant effect, as in cyber security not the known status, but deviations from it are what matters.

The linguistic diversity in the references is evaluated based on the repetitiveness of the descriptive vocabulary. Descriptions in different languages are deemed to be different. Such metrics allow distinguishing events that cause real discussions (sign that the event directly affects someone's interests) and allows avoiding over-evaluation, which occurs in many social networking monitoring systems due to repetitions.

Conclusion

The paper showed that the use of entity-event ontologies in cyber security is of practical value. Additionally, it can be a significant component or an auxiliary mechanism as part of other methods.

Another finding is the high informativeness of features based on retrieved information when it is used in machine learning models, which allows improving the quality of such models used in cyber security for the purpose of identifying anomalies, extrapolation and prediction, classification and clusterization, search for patterns and associations.

The theoretical result consists in the feasibility to preprocess corpora that enables the use of classical quantitative and categorical methods with regard to texts by means of information acquisition.

Further research should cover the applicability of the method for the analysis of logs (including within borders), correspondence (search for leaks), social networks monitoring, analysis of cyber security documents.

References

[1] Kühner H., Seider D. Security Engineering für den Schienenverkehr. *Eisenbahn Ingenieur Kompendium* 2018:245-264.

[2] Makarov B.A. Topicality of cybersecurity on railway transport. *Railway Equipment Journal* 2015;3(31):10-15.

[3] Kiseliova E.M. [Railways as an object of cyber security]. www.eduherald.ru; 2018 [accessed 15.06.2020]. Available at: http://www.eduherald.ru/ru/article/view?id=19179. (in Russ.)

[4] McNeil N., Bridges R.A., Iannacone M.D., Czejdo B., Perez N., Goodall J.R. Pace: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts. 12th International Conference on Machine Learning and Applications 2013;2:60-65.

[5] Kuzmina N.M., Ridley M.K. About automatic construction in information systems of civil aviation ontology of the subject field on the corps of texts. *Scientific Bulletin of The State Scientific Research Institute of Civil Aviation* 2018;21:122-131. (in Russ.)

[6] Kuzmina N.M., Ridley M.K. Architecture of ontology construction and semantic search system. *Scientific Bulletin of the State Scientific Research Institute of Civil Aviation* 2019;28:103-113. (in Russ.)

[7] Bergner S., Lechner U. Cybersecurity ontology for critical infrastructures. Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management 2017;2:80-85.

[8] Trabelsi S., Plate H., Abida A., Aoun M., Zouaoui A., Missaoui C., Gharbi S., Ayari A. Mining social networks for software vulnerabilities monitoring. 7th International Conference on New Technologies, Mobility and Security (NTMS) 2015:1-7. DOI:10.1109/NTMS.2015.7266506.

About the authors

Michael K. Ridley, post-graduate student, Moscow Aviation Institute (National Research University), Russian Federation, Moscow, e-mail: mr@kalabi.ru

The author's contribution

The author has analyzed the subject area, suggested a method of acquisition of information from open sources using entity-event ontologies. An analytics system previously developed by the author for the purpose of acquisition and storage of entity-event ontologies was improved and adapted to the needs of cyber security. Five application problems were solved, i.e. monitoring and analytics in cyber security, early detection of zero-day vulnerabilities, identification of the risk of a vulnerability's exploit, IP address scoring, risk-based event and entity ranking in cyber security).

Conflict of interests

The author declares the absence of a conflict of interests.