Correlating Cyber Incident Information to Establish Situational Awareness in Critical Infrastructures

Giuseppe Settanni, Yegor Shovgenya, Florian Skopik, Roman Graf, Markus Wurzenberger, Roman Fiedler
Digital Safety and Security Department - AIT Austrian Institute of Technology
A-1220 Vienna, Austria
firstname.lastname@ait.ac.at

***Abstract*—Protecting Critical Infrastructures (CIs) against contemporary cyber attacks has become a crucial as well as complex task. Modern attack campaigns, such as Advanced Persistent Threats (APTs), leverage weaknesses in the organization's business processes and exploit vulnerabilities of several systems to hit their target. Although their life-cycle can last for months, these campaigns typically go undetected until they achieve their goal. They usually aim at performing data exfiltration, cause service disruptions and can also undermine the safety of humans. Novel detection techniques and incident handling approaches are therefore required, to effectively protect CI's networks and timely react to this type of threats. Correlating large amounts of data, collected from a multitude of relevant sources, is necessary and sometimes required by national authorities to establish cyber situational awareness, and allow to promptly adopt suitable countermeasures in case of an attack. In this paper we propose three novel methods for security information correlation designed to discover relevant insights and support the establishment of cyber situational awareness.**

***Index Terms*—information correlation, cyber incidents handling, cyber situational awareness, critical infrastructure protection,**

I. Introduction

Cyber threats affecting Critical Infrastructures (CIs) have become widespread as well as more damaging and disruptive. Advanced Persistent Threats (APTs) leverage the complexity and the interconnectedness of CI networks, and exploit vulnerabilities of diverse systems aiming at hitting a specific target [1]. Traditional host-based detection techniques are therefore not effective anymore for protecting CIs against such threats. Collaborative approaches based on information sharing and data correlation are required in order to thoroughly comprehend the security status of a CI, and timely react to counter revealed threats [2].

Organization's Security Operation Centers (SOCs), responsible for the protection of CIs, begin therefore to cooperate with one another, with Computer Emergency Response Teams (CERTs), and with national authorities, by exchanging relevant security information. While handling security incidents, SOCs analyze large amounts of data, attempt to derive meaningful relations among them, and eventually obtain possible solutions to mitigate the reported incidents.

Moreover, as set forth by the provisions of the recently published European directive on security of Network and Information Systems (NIS) [3], CIs deployed in the EU Member States are required to report critical NIS incidents to the national competent authorities. These authorities are hence responsible for the collection, aggregation and correlation of such information, in order to establish so called national cyber Situational Awareness (SA) [4].

Advanced data processing techniques for analyzing diverse data collected from multiple sources are of fundamental importance. Information fusion and correlation approaches are frequently used to support such operation [5]. Correlating incident reports and threat information means finding similarities in the text that they comprise and in the meaning they convey. Advanced text analysis techniques based on Vector Space Models (VSM) are widely adopted in order to model text, and extract meaning from it [6]. Deriving the similarities between VSM-represented documents is a task that can be carried out with different methods.

In this paper we present and evaluate three different VSM-based information correlation methods which adopt the *Cosine Similarity* as a metric to compare security information.

The remainder of the paper is structured as follows. In Section II we review the state of the art in the scope of text correlation techniques. In Section III we introduce three novel methods for correlating cyber security information and deduce relevant similarities among them. Section IV outlines the planned implementation activities and the foreseen evaluation process. We conclude the paper in Section V.

II. Related Work

Correlating cyber incident reports to derive similarities and meaningful relations is tightly connected to natural language processing and semantics. Many approaches for such text analysis are based on *Vector Space Models* (VSMs). The idea of the VSM is to represent each document in a collection as a point, or a vector, in a multidimensional space. Points that are close together in this space are very similar, while points that are far apart are less similar or entirely different. An extensive survey exploring existing VSMs and their applications in semantics has been published by Turney and Pantel [6].

VSMs automatically extract knowledge from a given corpus, and therefore require less effort than other approaches such as hand-coded knowledge bases and ontologies. VSMs are effective in tasks that involve measuring the similarity of meaning between words, phrases, and documents. Among

the several VSMs designed for addressing different semantics problems, the model that aims at measuring the similarity between two documents (or between a query and a document), is called *Term-Document Matrix* [7].

In information retrieval, the *bag of words hypothesis* is that we can estimate the relevance of documents to other documents (or to a query) by representing the documents (and the query) as *bags of words*[1]. That is, the frequencies of words in a document tend to indicate the relevance of the document to another document (or to a query). The bag of words hypothesis is the basis for applying VSM to information retrieval and correlation [8].

Term-document matrices are nowadays also adopted in diverse applications including document clustering [7], document classification [9], document segmentation [10], question answering [11], and call routing [12]. The most popular implementation of a term-document matrix, powering many search engines, is Lucene[2], an open source text search engine library supported by the Apache Foundation [13].

Main alternatives to VSM for measuring document similarities are probabilistic models such as [14] and [15]. The idea is to measure the similarity between documents by creating a probabilistic language model of the given documents according to the language model. However, with the progress in information retrieval, the distinction between the VSM approach and the probabilistic approach has become blurred.

In this paper we present three custom methods, based on *term-document VSM*, designed to correlate cyber incidents reports and threat information, to provide insights on the security situation of complex computer networks, and hence support cyber incident handling tasks carried out by security operation teams.

III. Document Correlation Methods

Human-readable IT security information comes usually in form of semi-structured text documents such as incident reports, vulnerability alerts, advisories, bulletins, etc. Analyzing such documents means extracting significant information they comprise and identifying potential existing interrelations among them, in order to comprehend their impact and outline possible mitigation strategies. To support such analysis operations we designed three custom *term-document VSM* correlation approaches (in the following referred to as *linking methods*): the *artifact-based*, the *word-based*, and the *dictionary-based* linking methods. We present these three methods in this section.

A. Defining Documents Similarity

Following the general VSM approach described in Section II, we represent each document as a multidimensional vector of features. Let

$$\mathcal{F} = \{f_1, f_2, ..., f_n\} \tag{1}$$

be the set of all n unique features, and

$$\mathcal{D} = \{d_1, d_2, ..., d_m\} \tag{2}$$

be the set of m documents.

Each document d in $\mathcal{D}$ is therefore represented by its feature vector

$$\mathbf{v_r} = (v_{1d}, v_{2d}, ..., v_{nd}) \tag{3}$$

Given the feature vectors $\mathbf{v_x}$ and $\mathbf{v_y}$ of two given documents d_x and d_y, we can calculate the similarity between d_x and d_y by determining their *cosine similarity* $s(d_x, d_y)$, i.e. the *cosine* of the angle between their respective feature vectors:

$$s(d_x, d_y) = \frac{\mathbf{v_x} \cdot \mathbf{v_y}}{\|\mathbf{v_x}\|\|\mathbf{v_y}\|} \tag{4}$$

where $\|\mathbf{v_x}\|$ is the norm of the feature vector $\mathbf{v_x}$:

$$\|\mathbf{v_x}\| = \sqrt{v_{1x}^2 + v_{2x}^2 + ... + v_{nx}^2} \tag{5}$$

There exist numerous measures of vector distance such as Hellinger, Bhattacharya, and Kullback-Leibler. A study by Bulliinaria and Levy [16] compared the aforementioned measures with cosine similarity and identified cosine as the best measure.

The three proposed linking methods are different from each other in two aspects:

1) the definition of the elements in the feature vectors,
2) the selection of features.

While in the *dictionary-based* method we adopt binary frequencies to populate the feature vectors, in the *artifact-based* and in the *word-based* methods we use *term frequency (TF)* and *inverse document frequency (IDF)* metrics [8], calculated as follows.

Let z be the total number of unique features occurring in the document d. The normalized *term frequency (TF)* of the feature f in d is then:

$$TF_{f,d} = \frac{F_{f,d}}{\sum_{i=0}^{z} F_{f_i,d}} \tag{6}$$

where $F_{f_i,d}$ is the raw frequency of the feature f_i in d.

Let $\mathcal{D}$ be the total set of documents, and $\mathcal{D}_f$ the set of documents where feature f occurs at least once. The *inverse document frequency* of the feature f is then:

$$IDF_f = \ln \frac{|\mathcal{D}|}{|\mathcal{D}_f|} \tag{7}$$

where $|\mathcal{D}|$ is the number of documents in the set $\mathcal{D}$, and $|\mathcal{D}_f|$ is the number of documents in the set $\mathcal{D}_f$.

The way we select features in each method is discussed in the following three subsections.

[1] In mathematics, a *bag* of words is like a set, except that duplicates are allowed.

[2] Apache Lucene: http://lucene.apache.org/

B. Artifact-based Linking

In our previous work [17] we made the assumption that a security-relevant text document can be characterized by words, or word combinations, that represent known entities (*artifacts*) relevant for the ICT security domain: concepts such as "encryption" or "cross-site request forgery", product names and versions, company names etc.

For a single occurrence of an artifact a in a document d it is sufficient that any word set associated with a fully appears in d. The raw frequency $F_{a,d}$ of a in d is the total number of such occurrences within this document.

The feature vector $\mathbf{v_d}$ of the document d will then consist of every known artifact's TF·IDF values in context of d:

$$\mathbf{v_d} = (TF_{a_1,d} \cdot IDF_{a_1}, ..., TF_{a_n,d} \cdot IDF_{a_n}) \tag{8}$$

where n is the total number of existing artifacts, TF is calculated as in Equation 6, and IDF is calculated as in Equation 7.

The *artifact-based* linking method is a *supervised* method because it leverages the "intelligence" included in the artifact set, i.e. by selecting the artifact set we can specify what the most relevant concepts are. However the correlation capabilities depend on the broadness of the adopted artifact set.

C. Word-based Linking

In contrast to the approach described above, in the *word-based* linking method we adopt as features the documents' own words. We first split every document's text into single words, and we filter out those included in a common English stop-word list[3]. For every unique word we compute the TF (as in Equation 6) and the IDF (as in Equation 7).

Words with an IDF below a certain empirically defined threshold are then ignored because considered too frequent. This allows to exclude words like *'vulnerability'*, *'threat'* or *'attack'*, that are commonly used in security context and have lower entropy due to their high occurrence rate.

TF·IDF values of the remaining n high-IDF words will determine the feature vector $\mathbf{v_d}$ of a document d:

$$\mathbf{v_d} = (TF_{w_1,d} \cdot IDF_{w_1}, ..., TF_{w_n,d} \cdot IDF_{w_n}) \tag{9}$$

Contrarily to the *artifact-based* method, by using only the documents' own words for building feature vectors, the *word-based* method does not require any predefined dictionary, which makes this approach *unsupervised* and therefore context-independent.

D. Dictionary-based Linking

The third method we propose also involves words as features; however, rather than extracting them from the documents in the dataset, we employ here an empirically determined dictionary including ICT-security-pertinent words.

A further difference to the previous two methods is that the feature vector $\mathbf{v_d}$ of a document d is not composed of the words' TF·IDF values, but instead of their binary frequencies: each element of the vector can be either 1 (if the word is present in a document) or 0 (otherwise):

$$\mathbf{v_d} = (b_{f_1}, b_{f_2}, ..., b_{f_n}) \tag{10}$$

$$b_{f_i} = \begin{cases} 1, & \text{if } f_i \in \text{d} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where n is the size of the dictionary, currently set to 1000 as mentioned above.

The *dictionary-based* approach combines the traits of the *artifact-based* and the *word-based* methods: features are single words, but only those included in a predefined domain-specific dictionary are considered.

IV. Planned Implementation and Evaluation

The correlation methods presented in this paper have been developed as proof of concepts. Future work include their implementation into an operational analysis engine, called CAESAIR (*Collaborative Analysis Engine for Situational Awareness and Incident Response*) presented in our previous work [17].

As depicted in Figure 1, CAESAIR imports security-data from diverse input sources and in several standard formats (such as STIX[4], IODEF[5] and JSON). When the analyst selects a document, the system extracts its relevant features (artifacts or words depending on the enabled correlation method) and maps them to the documents feature vector; it then performs the document linking, examining all the other documents present in the knowledge base. Through its graphical interface, CAESAIR displays the rated list of the derived most relevant documents, sorted according to their similarity to the selected one.

Moreover, it is planned to extensively evaluate the three correlation methods in terms of accuracy and efficiency. In order to assess the precision of the methods, we intend to initially generate a dataset of semi-synthetic security-related documents; this dataset needs to be specifically designed to allow to observe how accurately documents related to one another are identified by the correlation method under test, and how the related documents are scored and ranked respectively. In this evaluation it is of interest to observe if documents that are not related to a given one obtain a (significantly) low score, and are therefore considered as non relevant to it. To test the efficiency of the methods, correlation of documents has to be performed considering several datasets of different size; by doing this the trend of average correlation time can be observed, and the scalability of the method can be tested.

These first assessments will allow to opportunely customize and adjust the different methods before deploying them in an operational environment, and employ them with real security-related data.

[3] http://www.ranks.nl/stopwords

[4] https://stixproject.github.io/

[5] https://www.ietf.org/rfc/rfc5070.txt

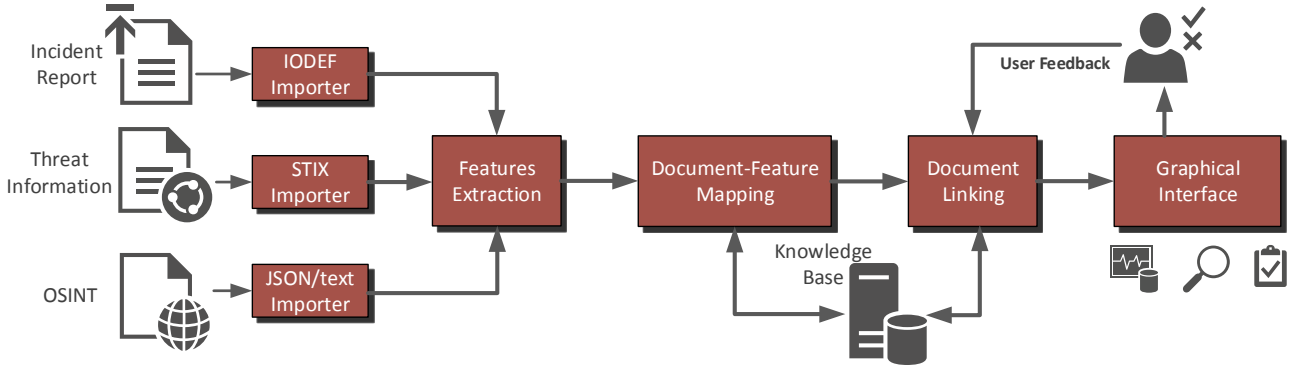

Fig. 1. CAESAIR process diagram.

Finally, it is planned to integrate CAESAIR analysis system in the context of a European Control System Security Incident Analysis Network (ECOSSIAN)[18]. This framework aims at providing critical infrastructures across Europe with the necessary methods and technologies to: i) timely detect cyber threats affecting their networks, ii) process and securely share IT- and ICS-related incident among CIs, and with national authorities, and iii) obtain early warnings and strategic support to promptly mitigate revealed threats.

V. Conclusion and Outlook

A. Conclusion

Correlating natural language documents to identify their similarities is essential in the process of cyber incident handling. It allows to rate the available information according to their relevance and discover their interrelations. In this paper we proposed three novel term-document VSM methods for correlating IT security information. Each method performs document correlation using a differently defined feature vector; depending on the computational power required to calculate the similarities, one method can be suitable for a specific use case rather than another.

The results obtained from a preliminary evaluation show that the most accurate methods require also more resources; on the other hand, the fastest methods resulted to be less accurate. Their adoption can however depend on the application: in circumstances where a quicker but less precise correlation is desired the *dictionary-based* method suits best; conversely, when the accuracy is fundamental, but the time requirements are not stringent, the *word-based* approach is the most valid choice. In cases where a trade-off between precision and speed is necessary, the *artifact-based* method provides suitable results.

Acknowledgements

This work was partly funded by the European Union FP7 project ECOSSIAN (607577).

References

[1] C. Tankard, "Advanced persistent threats and how to monitor and deter them." *Network Security*, vol. 2011, no. 8, pp. 16–19, 2011.

[2] F. Skopik, G. Settanni, and R. Fiedler, "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing," *Computers & Security*, vol. 60, pp. 154–176, Jul. 2016.

[3] European Commission, "The Directive on Security of Network and Information Systems (NIS Directive)," 2016.

[4] U. Franke and J. Brynielsson, "Cyber situational awareness–a systematic review of the literature," *Computers & Security*, vol. 46, pp. 18–31, 2014.

[5] S. J. Yang, A. Stotz, J. Holsopple, M. Sudit, and M. Kuhl, "High level information fusion for tracking and projection of multistage cyber attacks," *Information Fusion*, vol. 10, no. 1, pp. 107–121, 2009.

[6] P. D. Turney, P. Pantel *et al.*, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, no. 1, pp. 141–188, 2010.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[8] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[9] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[10] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, Mar. 1997. [Online]. Available: http://dl.acm.org/citation.cfm?id=972684.972687

[11] H. T. Dang, D. Kelly, and J. J. Lin, "Overview of the trec 2007 question answering track." in *TREC*, vol. 7, 2007, p. 63.

[12] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Comput. Linguist.*, vol. 25, no. 3, pp. 361–388, Sep. 1999.

[13] E. Hatcher and O. Gospodnetic, *Lucene in Action*. Manning Publications, Dec. 2004. [Online]. Available: http://www.worldcat.org/isbn/1932394281

[14] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[15] X. Liu and W. B. Croft, "Statistical language modeling for information retrieval," DTIC Document, Tech. Rep., 2005.

[16] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior research methods*, vol. 39, no. 3, pp. 510–526, 2007.

[17] G. Settanni, F. Skopik, Y. Shovgenya, and R. Fiedler, "A collaborative analysis system for cross-organization cyber incident handling," in *Proceedings of the 2nd International Conference on Information Systems Security and Privacy*, 2016, pp. 105–116.

[18] H. Kaufmann, R. Hutter, F. Skopik, and M. Mantere, "A structural design for a pan-european early warning system for critical infrastructures"," in *Elektrotechnik und Informationstechnik*. Springer, 2014.