Acquiring Cyber Threat Intelligence through Security Information Correlation

Giuseppe Settanni, Yegor Shovgenya, Florian Skopik, Roman Graf, Markus Wurzenberger, Roman Fiedler
Austrian Institute of Technology
Donau-City-Strasse 1, 1220 Vienna, Austria
firstname.lastname@ait.ac.at

***Abstract*—Cyber Physical Systems (CPS) operating in modern critical infrastructures (CIs) are increasingly being targeted by highly sophisticated cyber attacks. Threat actors have quickly learned of the value and potential impact of targeting CPS, and numerous tailored multi-stage cyber-physical attack campaigns, such as Advanced Persistent Threats (APTs), have been perpetrated in the last years. They aim at stealthily compromising systems' operations and cause severe impact on daily business operations such as shutdowns, equipment damage, reputation damage, financial loss, intellectual property theft, and health and safety risks. Protecting CIs against such threats has become as crucial as complicated. Novel distributed detection and reaction methodologies are necessary to effectively uncover these attacks, and timely mitigate their effects. Correlating large amounts of data, collected from a multitude of relevant sources, is fundamental for Security Operation Centers (SOCs) to establish cyber situational awareness, and allow to promptly adopt suitable countermeasures in case of attacks. In our previous work we introduced three methods for security information correlation. In this paper we define metrics and benchmarks to evaluate these correlation methods, we assess their accuracy, and we compare their performance. We finally demonstrate how the presented techniques, implemented within our cyber threat intelligence analysis engine called CAESAIR, can be applied to support incident handling tasks performed by SOCs.**

1. Introduction

News of targeted cyber attacks against critical infrastructures including power plants, energy grids, railway networks, and telecommunication systems, increasingly populate newscast and newspapers [1]. Multi-stage stealth campaigns, such as APTs, causing damages and disruption to large CIs occur indeed more and more often. Leveraging the complexity and interconnectedness of CI networks, they exploit vulnerabilities of diverse systems in the attempt of hitting a specific target [2]. Collaborative approaches based on information sharing and data correlation are therefore required to overcome the limits of traditional host-based detection methods, to thoroughly comprehend the security status of a CI, and to timely react and counter revealed threats [3].

Private organizations and public authorities began to cooperate in the effort of establishing more effective security measures for protecting their CIs. For example, CIs operating within the Member States of the European Union are now required to report critical network and information systems (NIS) incidents to the respective national competent authority [4]. These newly established authorities are responsible for the collection, aggregation and correlation of such information, for establishing so called national cyber Situational Awareness (SA) [5]. Organization's Security Operation Centers (SOCs) start hence exchanging relevant security information with one another, with Computer Emergency Response Teams (CERTs), and with national authorities.

Large amounts of data are analyzed while handling security incidents in order to derive meaningful relations among them, and eventually obtain possible solutions to mitigate the reported incidents. 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 [6]. Automated processing of security data is essential in the incident analysis process, however it is not sufficient to derive the root cause of an incident and provide a reasonable mitigation strategy. Human involvement in such tasks is in fact still fundamental to accurately interpret the analysis results and provide precise and tailored recommendations [7].

In our previous work [8] we introduced the concept of a cyber intelligence analysis system, called *CAESAIR* (Collaborative Analysis Engine for Situational Awareness and Incident Response), designed to provide analytical support for security experts carrying out cyber incident handling tasks on a national and international level [9]. We recently extended this approach by introducing three custom security information correlation techniques based on Vector Space Models (VSM) [10]. In this paper we evaluate the accuracy and we assess the performance of these correlation methods, and we demonstrate the applicability of our approach, integrated in the CAESAIR system, within a European Control System Security Incident Analysis Network (ECOSSIAN) [11].

The remainder of the paper is structured as follows. In Section 2 we review the state of the art in the scope of cyber incident handling and threat intelligence gathering approaches. In Section 3 we recall the previously introduced methods for correlating cyber security information.

In Section 4 we present the evaluation results and we discuss the accuracy and the performance of our correlation methods. In Section 5 we demonstrate the applicability of our approach, to ease the incident handling tasks of national and transnational security operation centers. We conclude the paper in Section 6.

2. Related Work

In our previous paper [10] we presented three information correlation methods, based on *term-document VSM*, designed to process cyber threat intelligence and to derive similarities and meaningful relations amongst security-relevant documents. In the following sections we demonstrate how, by adopting our methods to correlate cyber incident reports and threat information, CAESAIR[1] provides insights on the security situation of complex computer networks, hence supporting cyber incident handling tasks carried out by security operation teams.

To achieve the same objective other approaches have been proposed in recent years. Yang et al. [6] introduced a high level information fusion method for APTs, which processes alerts generated by Intrusion Detection Systems (IDSs) and fuses this data to address the tracking and projection of multistage cyber attacks. ENISA, the European Union Agency for Network and Information Security, in cooperation with a group of four European CERTs (Computer Emergency Response Teams), are currently working on IHAP, the Incident Handling Automation Project [12] aiming at improving the incident handling process by increasing automation. In IHAP, CERT teams use a unified *Data Harmonization Ontology* to enhance the actionable reporting and analysis of the collected information[2]. MISP[3], the Malware Information Sharing Platform, performs automatic data correlation finding relationships between attributes and indicators from malware, attacks campaigns or analysis. It incorporates an indicators database to store technical and non-technical information about malware samples, incidents, attackers and intelligence, and a sharing functionality to ease data exchange using different models of distribution.

Compared to the aforementioned tools, the advantage of our cyber threat intelligence solution, lies in the fact that CAESAIR does not only collect and aggregate incident and threat data, making it comfortably available to the analyst; rather, it also extensively correlates this data with large amounts of security information collected from several relevant sources, and provides the experts with a list of related information, greatly supporting them in the decision making process while handling cyber incidents. Moreover, it is seamlessly integrated in a pan-European critical infrastructure cyber incident analysis and response framework, as demonstrated in Section 5.

[1] http://caesair.ait.ac.at

[2] https://github.com/certtools/intelmq

[3] https://github.com/MISP/MISP

3. Document Correlation Methods

To identify and stop modern cyber attacks, organizations need to understand how attackers think, what they want, and how they work. It is hence essential to collect and analyze all the available information related to ongoing and previous attacks, and transform it into intelligence. Security information, such as incident reports, vulnerability alerts, advisories, bulletins etc., comes usually in form of semi-structured text documents. Acquiring cyber threat intelligence from such documents means extracting significant information they comprise, and identifying implicit interrelations among them, in order to comprehend their impact and outline possible mitigation strategies. To support such analysis operations we designed, and presented three custom *term-document VSM* correlation approaches [10] (in the following referred to as *linking methods*): the *artifact-based*, the *word-based*, and the *dictionary-based* linking methods. In this section we shortly recall these correlation methods.

Following the general VSM approach described in [13], we represent each document as a multidimensional vector of features. Each document d in the dataset $\mathcal{D}$ is therefore represented by its feature vector $\mathbf{v_r} = (v_{1d}, v_{2d}, ..., v_{nd})$. Being $\mathbf{v_x}$ and $\mathbf{v_y}$ the feature vectors of two given documents d_x and d_y, we calculate their correlation by determining their *cosine similarity* $s(d_x, d_y)$ [14].

The three proposed linking methods are different from each other in two aspects: i) the definition of the elements in their feature vectors, ii) 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)*[4] and *inverse document frequency (IDF)*[5] metrics [15]. The way we select features in each method is discussed in the following three subsections.

3.1. Artifact-based Linking

We assume here 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:

$$\mathbf{v_d} = (TF_{a_1,d} \cdot IDF_{a_1}, ..., TF_{a_n,d} \cdot IDF_{a_n}) \qquad (1)$$

[4] 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: $TF_{f,d} = F_{f,d} / \sum_{i=0}^{z} F_{f_i,d}$. Where $F_{f_i,d}$ is the raw frequency of the feature f_i in d.

[5] 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: $IDF_f = \ln(|\mathcal{D}|/|\mathcal{D}_f|)$. 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 $\mathcal{D}_f$.

3.2. Word-based Linking

In the *word-based* linking method we adopt as features the documents' own words. For every unique word we compute the TF and the IDF. Words with an IDF below a certain empirically defined threshold are ignored because considered too frequent. TF·IDF values of the remaining n high-IDF words will determine the feature vector $\mathbf{v_d}$ of a document d:

$$\mathbf{v_d} = (TF_{w_1,d} \cdot IDF_{w_1}, ..., TF_{w_n,d} \cdot IDF_{w_n}) \quad (2)$$

3.3. Dictionary-based Linking

In this method rather than extracting words from the documents in the dataset, we employ an empirically determined dictionary including ICT-security-pertinent words. The feature vector $\mathbf{v_d}$ of a document d is not composed of the words' TF·IDF values, but of their binary frequencies b_{f_i}: each element of the vector can be either 1 (if the word is present in a document) or 0 (otherwise):

$$\mathbf{v_d} = (b_{f_1}, b_{f_2}, ..., b_{f_n}) \quad (3)$$

4. Evaluation

This section describes how we assessed and compared the presented correlation methods; we introduce the dataset we generated and adopted in the evaluation phase, we define the metrics for measuring the linking accuracy, and we finally present the evaluation results and we discuss the methods' performance.

4.1. Dataset Generation

In order to evaluate and compare the proposed correlation methods we generated a realistic semi-synthetic dataset, so that the degree of similarity between each pair of documents is known a priori. Starting from 10 security bulletins[6], reporting about 10 diverse security events affecting 10 distinct information systems, we created 10 *Master Documents (MD)*. Each of these documents is a text file describing an event that is completely unrelated and syntactically diverse to every other MD. Indeed, we made sure that each document includes words and concepts that are not present in any other MD. Words such as *'vulnerability'*, *'threat'*, *'attack'*, etc., that are commonly used in security reports, are also present in several MDs. Such words will be considered as non-relevant, and therefore will be neglected in the *artifact-based* and *word-based* linking methods. In fact, due to their high occurrence rate (and hence low entropy), they will have a too low IDF. On the other hand, they will be considered as any other occurring word (as their binary frequency will be 1 anyway) in the *dictionary-based* method. To reflect the diversity of real IT security reports in the evaluation dataset, we made sure that the selected MDs have different lengths: some MDs comprise 2-3 lines of text, while others reach hundreds of lines. As clarified in the following subsection, the documents' length, however, does not influence the correlation results.

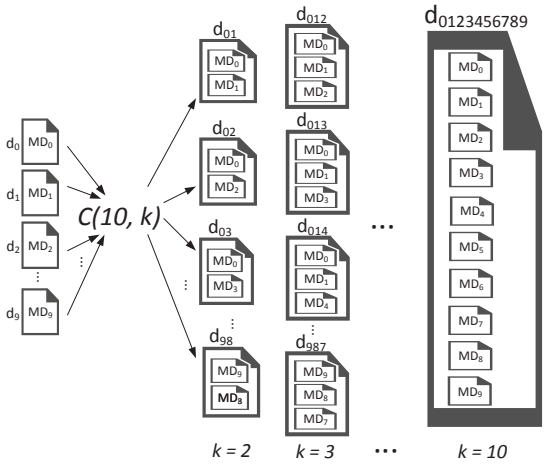

Figure 1. Evaluation dataset generation.

The 10 Master Documents, $MD_0, MD_1, MD_2, ..., MD_9$, are then opportunely combined in order to obtain further text documents to be included into the evaluation dataset. The goal is to have a clear overlap between each pair of documents. To achieve this we generate new documents concatenating a variable number of different MDs. The higher the number of MDs mutually present in two documents is, the larger is the overlapping portion of two documents, and hence the more similar are the two documents.

As shown in the diagram in Figure 1, we create one document for each *Combination* $\mathcal{C}(n, k)$, where n is the number of MDs, and $k \in \{1, ..., 10\}$. For example, the document d_{01} is created by concatenating the Master Documents MD_0 and MD_1, i.e., $d_{01} = \{MD_0, MD_1\}$; similarly, d_{012} is created by concatenating MD_0, MD_1, and MD_2; i.e., $d_{012} = \{MD_0, MD_1, MD_2\}$. The evaluation dataset $\mathcal{D}$ is therefore composed by 1023 different text documents whose contents are partially overlapping[7].

4.2. Evaluation Metric Definition

A correlation method is considered accurate if it derives the most similar documents, and it rates their similarity with an appropriately high score. The more similar two documents are, the higher the *linking score* should be. Ideally the linking score of two completely different documents is 0, whereas the score for two identical documents is 1. In order to have a *ground truth* and evaluate the results obtained adopting the different linking methods, we calculated the ideal linking score between each pair of documents in the evaluation dataset, using the following approach.

Let $\mathcal{Q}_{xy}$ be the set of MDs mutually present in d_x and d_y; i.e., $\mathcal{Q}_{xy} := \{MD_i | MD_i \in d_x \wedge MD_i \in d_y; i = 0, ..., 9\}$. The linking score between two documents $d_x =$

[6]The documents have been selected from the US-CERT security bulletins repository - https://www.us-cert.gov/ncas/bulletins

[7]The 10 documents obtained with $k = 1$ are exactly the 10 MDs, i.e., $d_i = MD_i$, whereas the document generated with $k = 10$ is obtained by appending all the MDs within the same document, $d_{0123456789} = \{MD_0, MD_1, MD_2, MD_3, MD_4, MD_5, MD_6, MD_7, MD_8, MD_9\}$

$\{MD_0, MD_1, ..., MD_{n_x}\}$ and $d_y = \{MD_0, MD_1, ..., MD_{n_y}\}$ is proportional to the number $q_{xy} = |\mathcal{Q}_{xy}|$ of MDs mutually present in the two documents. The expected linking score $ls(d_x, d_y)$ is calculated[8] as:

$$ls(d_x, d_y) = \frac{q_{xy}}{max(|d_x|, |d_y|)} \quad (4)$$

where $|d_x|$ (respectively $|d_y|$) is the amount of MDs comprised in d_x (d_y).

It is important to notice that the length of the MDs does not influence the calculation of the linking scores. We assume, in fact, that the MDs are opportunely selected such that they are entirely dissimilar from one-another, i.e., independently from their length, two different MDs are characterized by two non-overlapping text corpora. Hence, the similarity (and thus the linking score) between two generated documents depends exclusively on the number of MDs mutually present in the two documents. This implies that the linking score between any pair of different MDs is always 0, and that the linking score between two identical documents is always 1.

If we consider, for example, the document d_{012} (obtained by concatenating MD_0, MD_1, MD_2), and the document d_{023} (obtained by the concatenation of MD_0, MD_2, MD_3), they have 2 mutual MDs out of 3. This means that the they have $2/3$ of their MD content in common. Their linking score is therefore $ls(d_{012}, d_{023}) = 0.66$. Similarly, if we consider the documents d_{0123} and d_{012}, their linking score is $ls(d_{0123}, d_{012}) = 0.75$.

With this approach we calculate the linking score for every pair of documents in the evaluation dataset, and we store them in the *Linking Score Matrix*: $L(|\mathcal{D}| \times |\mathcal{D}|)$. Where $l_{ij} = ls(d_i, d_j)$ are the elements of L.

4.3. Accuracy Assessment and Methods Evaluation

To assess the accuracy of a linking method we consider each pair of documents in the evaluation dataset $\mathcal{D}$ and we calculate their similarity. Hence, we obtain a similarity matrix $S(|\mathcal{D}| \times |\mathcal{D}|)$, where $s_{ij} = s(d_i, d_j)$ are the elements of the matrix S, representing the cosine similarity between the documents d_i and d_j, calculated with the linking method under test.

Given that, alike the linking scores in L, the cosine similarities in S are based on the frequency of the features and their occurrence in the dataset documents (i.e., their TFIDF), in order to evaluate how accurate each linking method is, it is appropriate to compare the similarities in S (calculated with the method under test), against the linking scores included in the *ground truth* matrix L.

Thus, for each linking method, we measure the distance between the similarities in S, and the linking scores in L. We define a *tolerance* t, with $0 \leq t \leq 1$, as the maximum allowed difference between a value in S and its respective value in L, to consider the link precise.

[8]Note that if the two analyzed documents have different length (i.e., $|d_x| \neq |d_y|$), we consider the longest document to calculate the overlapping portion.

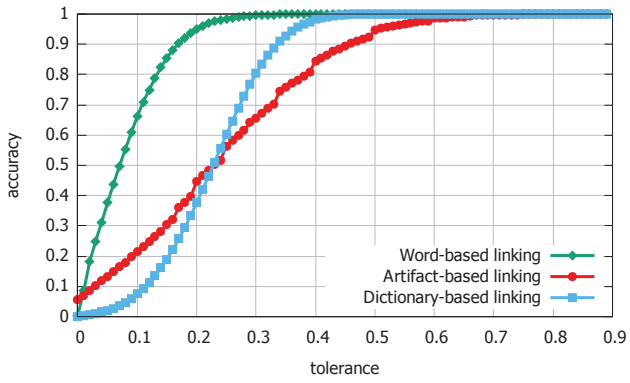

Figure 2. Accuracy of the 3 linking methods for values of tolerance t between 0 and 0.9. The accuracy reaches 90% with t between 0.16 and 0.46, depending on the method.

We then study the accuracy of each linking method, against the ground truth, by observing the portion of elements in matrix S which do not differ more than t from their respective elements in matrix L, i.e. the ratio of elements falling within a given tolerance range:

$$\text{accuracy} = \frac{|\{s_{ij}| \ |s_{ij} - l_{ij}| < t\}|}{|S|} \quad (5)$$

where $i, j = 1, ..., |D|$ and $t = 0, ..., 1$.

Figure 2 displays the accuracy of each linking method for a tolerance varying between 0 and 0.9. Every linking method reaches a high accuracy level with an acceptable tolerance ($0.2 < t < 0.3$). Moreover, when the tolerance is set to 0 the *artifact-based* linking method is the most accurate, while the other two methods have no similarity score matching exactly the ground truth scores. However, when the tolerance increases (from $t = 0.01$ onwards) the *word-based* linking method provides the highest accuracy, and reaches the maximum accuracy with a tolerance $t \geq 0.3$. The *artifact-based* linking method performs better than the *dictionary-based* method for $t < 0.22$, but is less accurate when the tolerance is higher, reaching the maximum accuracy only with $t = 0.7$. The reason for this lies in the fact that the *artifact-* and *dictionary-based* methods leverage a limited set of words and phrases from the ICT security domain, whereas the *word-based* linking relies on the words actually present in the analyzed documents.

In order to further evaluate the quality of the three presented methods we define a *discrimination threshold* t_r, and the binary conditions reported in Table 1, where: *True Positives* (*TP*) are the elements that, both in the similarity matrix S and in the ground truth matrix L, indicate strong similarities between two documents; *True Negatives* (*TN*) are the elements that, both in the similarity matrix S and in the ground truth matrix L, indicate weak similarities between two documents; *False Positives* (*FP*) are the elements that indicate strong similarities in the similarity matrix S, but weak similarities in the ground truth matrix

Table 1. BINARY CONDITIONS DEFINITION, AT A GIVEN DISCRIMINATION THRESHOLD t_r; s_{ij} ARE THE ELEMENTS OF THE SIMILARITY MATRIX S, AND l_{ij} ARE THE ELEMENTS OF THE GROUND TRUTH MATRIX L.

Condition	Definition
TP	$TP = \|\{s_{ij} \| s_{ij} > t_r \wedge l_{ij} > t_r\}\|$
TN	$TN = \|\{s_{ij} \| s_{ij} \leq t_r \wedge l_{ij} \leq t_r\}\|$
FP	$FP = \|\{s_{ij} \| s_{ij} > t_r \wedge l_{ij} \leq t_r\}\|$
FN	$FN = \|\{s_{ij} \| s_{ij} \leq t_r \wedge l_{ij} > t_r\}\|$

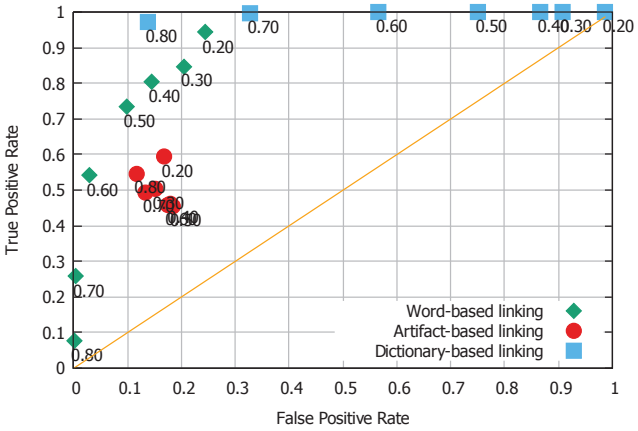

Figure 3. ROC space of the 3 linking methods for values of discrimination threshold between 0.2 and 0.8. The closer to the upper-left corner the points are, the more precise is a linking method in identifying similar documents.

L; *False Negatives* (*FN*) are the elements that indicate weak similarities in the similarity matrix S, but strong similarities in the ground truth matrix L.

We then calculate the *False Positive Rate (FPR)*, the *True Positive Rate (TPR)*, and the *F-measure*[9] and we represent them, for each linking method, for different values of t_r. These statistical metrics indicate how precisely the methods identify highly similar documents, and how precisely they distinguish those different.

Figure 3 shows the *Receiver Operating Characteristics* (*ROC space*) [16] for discrimination threshold values between 0.2 and 0.8. This graph, in combination with the plot reporting the F-measure trends (Figure 4), allow us to identify the optimal discrimination threshold for each method, i.e., the value of t_r providing the highest TPR, the smallest FPR and the highest F-measure.

By looking at Figure 3, one can notice that the *word-based* linking method provides low FPRs even for relatively high TPRs. For example, when the FPR is around 0.2, the TPR is around 0.85; this means that the method is able to correctly identify 85% of the documents that are very similar to a given one, and it erroneously considers similar 20% of the actually *dissimilar* documents. Given that the discrimination threshold to obtain these values is $t_r = 0.3$,

[9]In statistical analysis, FPR $= \frac{\text{FP}}{\text{FP+TN}}$, TPR $= \frac{\text{TP}}{\text{TP+FN}}$, and F-measure $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$, where: Precision $= \frac{\text{TP}}{\text{TP+FP}}$, and Recall $=$ TPR

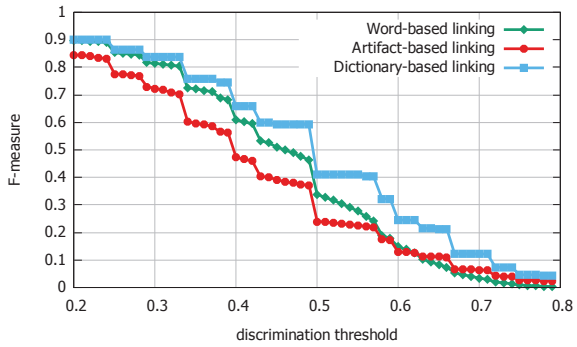

Figure 4. F-measure of the 3 linking methods for values of discrimination threshold between 0.2 and 0.8. The precision of the methods decreases similarly in the 3 methods when the discrimination threshold increases. The highest degradation is observable around $t_r = 0.5$.

we can observe in Figure 4 that the correspondent value of F-measure for the *word-based* method is around 0.81; this proves that this method provides a high precision when distinguishing dissimilar documents.

If we analyze the same statistics for the *artifact-based* method, we recognize a cloud of points in the central-left part of the ROC space (see Figure 3). More precisely, the values of FPR are concentrated between 0.1 and 0.2, while the TPR values lie between 0.45 and 0.6. This implies that the discrimination threshold does not influence much the performance of the method. The highest TPR (0.6) is reached for a FPR of 0.17, and a discrimination threshold of 0.2. This corresponds to an F-measure of 0.85 in Figure 4 and it is the best trade-off for this method.

Observing the ROC space for the *dictionary-based* method one could quickly conclude that this method allows to obtain the highest TPR values, for corresponding relatively low values of FPR (e.g., TPR=0.98 when FPR=0.14 and t_r=0.8), however, when considering the respective F-measure, it is clear that such promising results have a drawback: a very low value of F-measure, that is a low precision in distinguishing dissimilar documents. This is due to the fact that this method adopts binary frequencies, it does not consider the document frequency, and it only includes a relatively small set of features (i.e. 1000 words). Therefore, the documents are represented by smaller feature vectors and are considered similar to one-another more easily than in other methods. On the other hand, it is more difficult for two documents to be considered very different; this implies a low True Negative Rate (TNR) and hence a low F-measure. This issue can be mitigated by adopting a larger and more accurate set of words in the dictionary.

4.4. Performance Assessment

After evaluating the proposed linking methods qualitatively, we measured their performance in terms of speed[10].

[10]The three methods have been implemented within the framework described in Section 5. The evaluation has been executed running the system on a general purpose personal computer with Intel(R) Xeon(R) CPU E5-1620 v2 at 3.70GHz, 8 cores and 16GB of memory, running Ubuntu 12.04.5 LTS operating system.

In particular we observed how the dataset size influences the average time required to calculate all the similarities for a given document. Table 2 collects the results of this evaluation.

Table 2. AVERAGE LINKING TIME (IN SECONDS) WITH DIFFERENT DATASET SIZES

Method	$\|D\| = 100$	$\|D\| = 500$	$\|D\| = 1023$
word-based	0.369	0.920	1.789
artifact-based	0.444	1.377	3.724
dictionary-based	0.067	0.310	0.625

When considering a dataset of 1023 documents, with an average linking time of 0.6 seconds, the *dictionary-based* method is by far the fastest method. In fact, it benefits of smaller feature vectors and lower computational load (it does not calculate the TFIDF). As demonstrated in the previous subsection this implies the drawback of a lower accuracy and poorer precision. However this method results to be the most efficient if high precision is not a requirement, e.g., if all the related documents need to be determined, rather than a short list of most similar ones. Among the other two methods, the *word-based* method performs better because it looks for exactly those words extracted from the dataset, whereas the *artifact-based* method looks for concepts within external artifacts.

Table 2 also outlines that the three methods' performance scale-up when the dataset size increases. The longest linking time measured, with a dataset of 1023 documents, is in fact 3.7 seconds. This is an acceptable speed for non real-time operations such as human-driven incident handling (as described in Section 5).

5. System Implementation and Illustrative Application

The methods presented in Section 3 have been designed with a specific application in mind: to correlate IT-security information and facilitate the establishment of cyber situational awareness in a computer network. In this section we describe how we integrated the proposed approaches within *CAESAIR*, a Collaborative Analysis Engine for Situational Awareness and Incident Response introduced in our previous work [8]. We moreover demonstrate how such a system can be employed in a European network of cooperating CI's security operation centers.

5.1. System Implementation

The previously discussed document correlation methods were implemented as part of the CAESAIR system[11]. As depicted in Figure 5, CAESAIR imports security-data from diverse input sources and in several standard formats (such as STIX, IODEF 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 document's feature vector; it then performs the document linking, examining all the other documents present in the knowledge base. Through an intuitive graphical web client, CAESAIR displays the rated list of the derived most relevant documents, sorted according to their similarity to the selected one. This allows the analysts to faster and more extensively analyze significative security information, to identify meaningful relations between reported incidents, discovered vulnerabilities, targeted systems, and involved actors, allowing to achieve shorter incident response times. In upcoming versions of the software the user will be additionally able to provide feedback on the goodness of the calculated similarities, and therefore influence future correlation results.

It should be noticed that several optimization expedients have been put in place when implementing the linking methods within the CAESAIR system. Although the theory behind all the three approaches implies handling large sparse matrices with feature vectors, in practice these are actually not stored; indeed, in case of *artifact-* and *word-based* linking methods we only save the non-zero elements of every feature vector as separate rows in a database table. To calculate the cosine similarity between two documents d_1 and d_2, we dynamically compose two smaller vectors describing just the features that have a non-null value for at least one of the documents. In case of dictionary-based approach we however store the complete feature vectors, that are considerably shorter due to a smaller size of the dictionary.

5.2. A Pan-European Cyber Incident Analysis Framework

Handling IT incidents occurring in a critical infrastructure is nowadays as essential as complex. This task can be greatly facilitated if CI operators and qualified authorities cooperatively exchange security-relevant information in their possession [3]. In the ECOSSIAN[12] project we propose a pan-European system to collaboratively detect and react to cyber threats targeting CIs. The ECOSSIAN ecosystem foresees a three-tiered architecture of SOCs: at organization level (*O-SOC*), at national level (*N-SOC*) and at European level (*E-SOC*) [11]. Incidents affecting the CIs and detected at organization level are reported by the O-SOCs to their respective N-SOC; the N-SOC performs incident handling, establishes the national cyber situational awareness [5], and derives and distributes tailored mitigation strategies to the involved CIs.

In [9] we presented a possible N-SOC architecture and we discussed the necessary system components, in order to support such operations and provide: efficient threat data acquisition and aggregation, privacy-preserving information sharing, effective cyber threat intelligence analysis and correlation, intuitive visualization and evaluation of the analysis results. As exhaustively demonstrated by the use-case

[11]The logic is entirely implemented in Python; Elasticsearch is used for storing and querying the analyzed text documents and artifacts. The remainder of the internal data is persisted using PostgreSQL.

[12]www.ecossian.eu

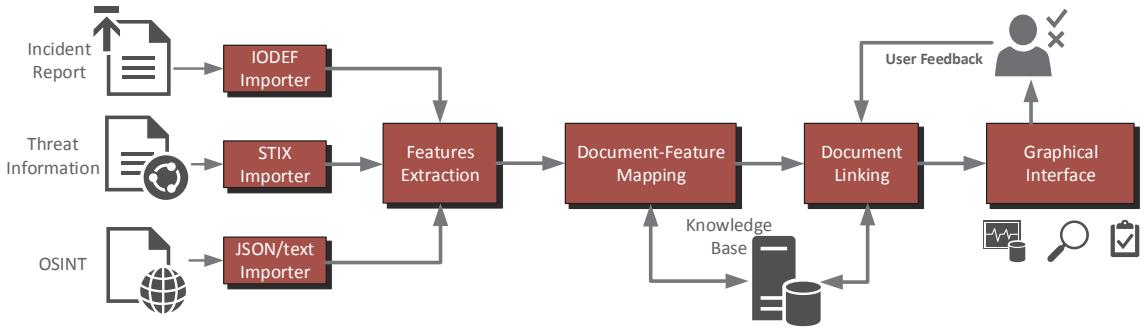

Figure 5. CAESAIR process diagram.

described in [8], CAESAIR plays a fundamental role in this architecture. Thanks to its powerful correlation capability, CAESAIR provides the N-SOC analysts with the necessary support to handle reported incident information; it quickly identifies related threats and possible existing solutions by examining numerous Open Source INTelligence (OSINT) feeds; it allows to establish situational awareness by keeping track of security incidents reported by the CIs deployed on the national territory.

6. Conclusion

Cyber incident handling tasks performed by SOC operators are often supported by automated analysis; this process needs to involve correlation of natural language documents to identify similarity amongst the collected information and transform it into threat intelligence. In this paper we evaluated three previously presented term-document VSM methods for correlating large amounts of IT security information. We observed that higher accuracy requires more resources; on the other hand, the fastest methods resulted to be less accurate. In applications where a quicker correlation is necessary, but high accuracy is not required, the *dictionary-based* performs the best; on the other hand, when high accuracy is desired, but the time constraints are not stringent, the *word-based* approach suits better. Furthermore, we demonstrated the integration of these techniques into an operational analysis engine (CAESAIR), and we discussed the applicability of such an approach into ECOSSIAN, a pan-European incident analysis network for critical infrastructure protection. The CAESAIR system is completely integrated within the ECOSSIAN architecture and, thanks to the methods presented in this paper, it has proven high quality correlation results when applied on diverse large-scale datasets.

Future work include the evaluation of the system with real datasets of IT incidents and threat information, as well as the application of clustering techniques on the correlation results to identify classes of relevant information.

Acknowledgments

This work was partly funded by the European Union FP7 project ECOSSIAN (607577), and by the Austrian FFG research projects synERGY (855457) and CISA (850199).

References

[1] ESET, "BlackEnergy trojan strikes again: Attacks Ukrainian electric power industry," http://www.welivesecurity.com/2016/01/04/blackenergy-trojan-strikes-again-attacks-ukrainian-electric-power-industry/ , 2016.

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

[3] 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.

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

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

[6] 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.

[7] A. D'Amico *et al.*, "Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts." in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 49, 2005, pp. 229–233.

[8] G. Settanni, F. Skopik *et al.*, "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.

[9] G. Settanni, F. Skopik, Y. Shovgenya, R. Fiedler *et al.*, "A collaborative cyber incident management system for European interconnected critical infrastructures," *Journal of Information Security and Applications*.

[10] G. Settanni, Y. Shovgenya, F. Skopik *et al.*, "Correlating cyber incident information to establish situational awareness in critical infrastructures," in *Proceedings of the 14th Annual Conference on Privacy, Security and Trust*, 2016.

[11] 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.

[12] ENISA, "Incident Handling Automation Project," 2015.

[13] P. D. Turney, P. Pantel *et al.*, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010.

[14] 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.

[15] 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.

[16] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.