НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2023 Том 23 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2023 Vol. 23 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2023-23-3-500-505
УДК 004.056

Метод выявления групп атакующих на основании анализа полезной нагрузки сетевого трафика по протоколу HTTP

Артем Валерьевич Павлов[1]✉, Наталия Викторовна Волошина[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] artempavlov1@gmail.com✉, https://orcid.org/0000-0001-8567-5469
[2] nataliv@yandex.ru, https://orcid.org/0000-0001-9435-9580

Аннотация

Введение. Атаки на веб-приложения являются частым направлением атаки на информационные ресурсы злоумышленниками различного уровня подготовки. Подобные атаки возможно исследовать с помощью анализа HTTP-запросов, произведенных атакующими. В работе исследована возможность выявления групп атакующих по данным событий систем обнаружения вторжений. Выявление групп атакующих позволяет улучшить работу аналитиков безопасности, расследующих и реагирующих на инциденты, снизить влияние событийной усталости при анализе событий безопасности, а также поможет выявить шаблоны и ресурсы атак злоумышленников, что повысит качество защиты системы в целом. **Метод.** Выявление групп атакующих в рамках предложенного метода выполнено на основании последовательности этапов. Проведено разбиение запросов на токены по регулярному выражению, основанному на особенностях протокола HTTP и атак, которые часто встречаются и выявляются системами обнаружения вторжений. Выполнено взвешивание токенов алгоритмом TF-IDF (Term Frequency-Inverse Document Frequency), что позволяет выделить редкие токены. На следующем этапе произведено отделение основного ядра запросов, не содержащих в себе редкие слова, совпадение по которым позволяет говорить о связанности событий. Таким образом происходит отделение использования публичных, доступных открыто от разработанных или модифицированных атакующими, инструментов атаки. Для определения расстояния применено манхэттенское расстояние. На последнем этапе проведена кластеризация методом DBSCAN (Density-based Spatial Clustering of Applications with Noise). **Основные результаты.** Показано, что данные полезной нагрузки HTTP-запросов могут использоваться для выявления групп атакующих. Предложен эффективный метод токенизации, взвешивания и кластеризации рассматриваемых данных и использование метода DBSCAN для кластеризации в рамках метода. Проведена оценка метрик: однородности, полноты и V-меры кластеризации, получаемых различными методами на наборе данных CPTC-2018. Выявлено, что предлагаемый метод позволяет получить кластеризацию событий, обладающую высокой однородностью и достаточной полнотой. Представлено комбинирование кластеризации с кластерами, образованными другими методами, с большой однородностью кластеризации для получения высокого показателя полноты и V-меры при сохранении большой однородности. **Обсуждение.** Предложенный метод может найти применение в работе аналитиков безопасности в SOC (Security Operations Center), CERT (Computer Emergency Response Team) и CSIRT (Cybersecurity Incident Response Team) как при противодействии вторжениям, так и в сборе данных о техниках и тактиках атакующих, включая атаки уровня APT (Advanced Persistent Threat). Метод позволит выявлять шаблоны следов инструментов, используемых атакующими, что даст возможность проводить атрибуцию атак.

Ключевые слова

группы атакующих, сложные атаки, обнаружение вторжений, корреляция событий безопасности, кибербезопасность

Ссылка для цитирования: Павлов А.В., Волошина Н.В. Метод выявления групп атакующих на основании анализа полезной нагрузки сетевого трафика по протоколу HTTP // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 3. С. 500–505. doi: 10.17586/2226-1494-2023-23-3-500-505

Attacker group detection method based on HTTP payload analysis

Artem V. Pavlov[1]✉, Natalia V. Voloshina[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] artempavlov1@gmail.com✉, https://orcid.org/0000-0001-8567-5469

[2] nataliv@yandex.ru, https://orcid.org/0000-0001-9435-9580

Abstract

Attacks on web applications are a frequent vector of attack on information resources by attackers of various skill levels. Such attacks can be investigated through analysis of HTTP requests made by the attackers. The possibility of identifying groups of attackers based on the analysis of the payload of HTTP requests marked by IDS as attack events has been studied. The identification of groups of attackers improves the work of security analysts investigating and responding to incidents, reduces the impact of alert fatigue in the analysis of security events, and also helps in identifying attack patterns and resources of intruders. Identification of groups of attackers within the framework of the proposed method is performed based on the sequence of stages. At the first stage, requests are split into tokens by a regular expression based on the features of the HTTP protocol and attacks that are often encountered and detected by intrusion detection systems. Then the tokens are weighted using the TF-IDF method, which allows to further give a greater contribution when comparing requests to the coincidence of rare words. At the next stage the main core of requests is separated based on their distance from the origin. Thus, requests not containing rare words, the coincidence of which allows us to talk about the connectedness of events, are separated. Manhattan distance is used to determine the distance. Finally, clustering is carried out using the DBSCAN method. It is shown that HTTP request payload data can be used to identify groups of attackers. An efficient method of tokenization, weighting and clustering of the considered data is proposed. The use of the DBSCAN method for clustering within the framework of the method is proposed. The homogeneity, completeness and V-measure of clustering obtained by various methods on the CPTC-2018 dataset were evaluated. The proposed method allows obtaining a clustering of events with high homogeneity and sufficient completeness. It is proposed to combine the resulting clustering with clusters obtained by other methods with high clustering homogeneity to obtain a high completeness metric and V-measure while maintaining high homogeneity. The proposed method can be used in the work of security analysts in SOC, CERT and CSIRT, both in defending against intrusions including APT and in collecting data on attackers' techniques and tactics. The method makes it possible to identify patterns of traces of tools used by attackers, which allows attribution of attacks.

Keywords

attacker groups, complex attacks, intrusion detection, alert correlation

For citation: Pavlov A.V., Voloshina N.V. Attacker group detection method based on HTTP payload analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 500–505 (in Russian). doi: 10.17586/2226-1494-2023-23-3-500-505

Введение

Анализ HTTP-запросов играет важную роль в обнаружении вторжений. Направления атак, которые можно в них обнаружить, включает SQL-инъекции, SSRF, XXE, инъекции команд и др. Многие из атак могут привести к удаленному выполнению кода на сервере, где размещено приложение, превращая его в плацдарм для дальнейшего продвижения злоумышленников. Согласно отчету Ростелеком-Солар[1], в 2021 году атаки на веб-приложения использовались для взлома инфраструктуры в 20 % атак группировок среднего уровня и в 50 % — высокого уровня.

Данные HTTP-запросов можно использовать не только для обнаружения вторжений, но и для последующего их анализа, в частности выделения групп атакующих. Понятие группы атакующих несколько шире сложных атак или APT (Advanced Persistent Threat), поскольку включает различные вредоносные кампании, проводимые одной группой. Таким образом, понятие группы атакующих охватывает атаки, проводимые одним злоумышленником, на одну цель (сложная атака), схожие цели (вредоносная кампания) и различные цели.

Выделение групп атакующих может улучшить работу SOC (Security Operations Center), CERT (Computer Emergency Response Team) и CSIRT (Cybersecurity Incident Response Team) благодаря преобразованию одномерного массива событий в двумерный массив действий групп. Это позволяет снизить влияние явления усталости от сигналов тревоги (alert fatigue) [1] при работе аналитиков. Выделение групп дает возможность точнее оценить уровень угрозы и глубину компрометации системы для принятия соответствующих мер реагирования. Также выделение групп и шаблонов их действий помогает при проведении форензиологических исследований, атрибуции атак, превентивном выявлении ресурсов, используемых атакующими.

Современные методы выделения групп атакующих включают подходы, основанные на атрибутах событий или на шагах, предпринимаемых атакующими при продвижении внутри системы [2]. Первая категория включает методы, основанные на свойствах событий, временных метках, статистических взаимосвязях и фильтрации, вторая — основанные на модели предпосылок-последствий и сценариях атак. Некоторые методы используют комбинированный подход с применением методов обеих категорий. Выделяемое сходство может быть основано на использовании схожего инструментария, совпадении целей и статуса атак, а также на основании обмена информацией внутри груп-

[1] [Электронный ресурс]. Режим доступа: https://rt-solar.ru/upload/iblock/01e/Otchet-ob-instrumentarii-professionalnykh-gruppirovok_2021-god.pdf (дата обращения: 31.03.2023).

пы атакующих. Детальный обзор методов корреляции событий безопасности при решении схожих задачах представлен в работах [3–5].

Среди методов работы с данными в кибербезопасности и обнаружении вторжений отметим взвешивание токенов с использованием алгоритма TF-IDF (Term Frequency-Inverse Document Frequency). Например, в работе [6] использована комбинация алгоритмов BPE (Byte Pair Encoding) и TF-IDF для обнаружения вторжений по данным HTTP-запросов. На наборе данных CSIC-2010[1] точность их метода с классификатором Isolation Forest достигла 89,48 %, а F-мера — 90,17 %. В [7] предложена система BotTokenizer, содержащая алгоритм TF-IDF. На основании URL (Uniform Resource Locator) запросов [7] с высокой точностью были определены в трафике известные семейства ботнетов. Среди других работ в области кибербезопасности и обнаружения вторжений с использованием TF-IDF можно выделить [8], в которой рассмотрена комбинация алгоритмов TF-IDF и SVM (Support Vector Machine) для обнаружения вторжений. Отметим, что проведенный анализ научных работ по теме показал, что подход TF-IDF не был изучен для решения задачи выделения групп атакующих или схожих задач.

В результате выполненного исследования предложен метод разбиения HTTP-запросов на токены и их взвешивания на основе алгоритма TF-IDF с последующей кластеризацией для выявления групп атакующих. Проведено исследование эффективности метода с применением различных методов кластеризации на наборе данных соревнований по тестированию на проникновение CPTC-2018. Представленный подход показал высокий уровень однородности получаемой кластеризации и достаточную полноту. Полученные результаты показали практическую значимость подхода и зарекомендовали его для использования в комплексных методах по выявлению групп атакующих.

Метод

Атакующие, использующие один и тот же инструмент, часто оставляют схожие следы в системе. Поскольку HTTP относится к текстовым протоколам, данные запросов, помеченных как попытка вторжения, могут использоваться для определения сходства инструментария, применяемого в рамках исследуемых атак. Также многие типы атак на веб-приложения требуют нескольких запросов для достижения цели атаки, что позволяет выявить атаку группы [9].

Известно, что многие группы атакующих используют инструменты, находящиеся в открытом доступе. Таким образом, следы данных инструментов будут схожи, однако это сходство не будет говорить о том, что атакующие представляли одну группу. Для решения задачи разделения случаев использования одного и того же непубличного и открытого инструментов подходит алгоритм TF-IDF [10]. TF-IDF работает по следующему принципу: после разбиения текста на токены, для каждого токена (слова) вычислим отдельно показатели TF и IDF:

$$TF(t, d) = \frac{n_t}{\sum_{i=1}^{k} n_i}, \quad (1)$$

где n_t — количество слов t в коллекции; k — общее количество уникальных слов; $\sum_{i=1}^{k} n_i$ — общее количество слов в документе; d — исследуемый документ.

$$IDF(t, D) = \ln\frac{n_c}{\sum_{j=1}^{m} n_j}, \quad (2)$$

где n_c — количество коллекций; $\sum_{j=1}^{m} n_j$ — количество коллекций, содержащих требуемое слово t; D — массив коллекций.

Рассчитаем общий показатель TF-IDF на основании выражений (1) и (2):

$$TF - IDF(t, d, D) = TF(t, d)IDF(t, D).$$

Таким образом, часто встречающиеся следы открытых инструментов будут иметь меньший вес при оценке сходства запросов, поскольку подобные артефакты будут встречаться чаще. На основании этого возможно реализовать метод, выполняющий выявление групп атакующих по содержимому запросов, помеченных системами обнаружения вторжений, как события атаки.

Предлагаемый метод разделим на три этапа.

Подготовка данных и токенизация. Для применения метода используем поле с нагрузкой (payload) HTTP-запроса событий безопасности. Приведем данные к текстовому виду: выполним, если необходимо, раскодирование из Base64, а также операцию URL-декодирования. Далее проведем токенизацию данных запросов по следующему регулярному выражению:

[^\s\/'\.\+&\?=<>\[\]:\(\)\-”’,]+.

Данный набор символов позволяет разделить слова внутри самого запроса с использованием синтаксиса HTTP значимые составляющие внутри нагрузки таких видов атак, как SQL-инъекция или эксплуатация XSS.

После токенизации оценим слова методом TF-IDF и отбросим слова, частота в документах которых больше порогового значения. В настоящей работе выбрано значение пороговой частоты 0,7. Произведенный отбор не оказал значительного влияния на результаты последующих этапов, но позволил существенно сократить время, требуемое для вычислений. Оптимальное значение порога выберем при реализации метода в рамках конкретного окружения для сокращения времени выявления.

Отбор запросов. Для оценки расстояния между запросами применим манхэттенское расстояние. Данный выбор связан с тем, что матрица TF-IDF является разреженной и имеет большую размерность с точки зрений количества свойств. Это приводит к неэффективности использования евклидова расстояния из-за «проклятия» размерности [11].

[1] [Электронный ресурс]. Режим доступа: https://www.tic.itefi.csic.es/dataset/ (дата обращения: 20.05.2023).

Отметим, что применение методов кластеризации к получившемуся набору даст в результате низкие показатели однородности и полноты кластеризации. Это связано с тем, что многие запросы не содержат в себе редкие слова. Таким образом, существует группа запросов, близких к началу координат по совокупности весов входящих в них слов, и близость этих запросов в геометрическом смысле не отражает соответствие одной и той же группе атакующих, а говорит лишь об отсутствии в запросе редких, аномальных слов. Более того, применение кластеризации ко всем запросам может быть сильно ограничено с точки зрения выбора методов кластеризации, поскольку матрица расстояний, используемая во многих методах, является ресурсоемкой, и размер потребляемых ресурсов растет квадратично от количества элементов.

Для отбора запросов, существенно отстоящих от центра координат, выполним оценку плотности распределения суммы расстояний слов в запросах от начала координат, определим границу, на которой происходит существенное падение плотности, и выберем запросы, находящиеся справа от этой границы, для которых верно выражение

$$dist_{Manh}(x_i, 0) > \tau,$$

где $dist_{Manh}$ — манхэттенское расстояние; τ — значение координаты границы; x_i — оцениваемый элемент выборки.

Кластеризация. К отобранным данным используем различные методы кластеризации. Для определения расстояния между объектами применим манхэттенское расстояние.

Экспериментальное исследование показателей оценки метода

Для подтверждения работоспособности метода выполним эксперимент с применением различных методов кластеризации данных.

Выберем набор данных CPTC-2018 [12]. Набор был собран в рамках соревнований по тестированию на проникновение, когда перед различными командами была поставлена задача получить максимальные привилегии в сети, моделирующей реальную компанию. Такой подход имеет ряд преимуществ по сравнению с другими. Так, при сборе данных из реального трафика пришлось бы вести очень длительное наблюдение, убирать конфиденциальные данные из трафика и проводить долгую и сложную разметку данных. При искусственном моделировании атак их разнообразие было бы существенно ограничено. Более того, в современных наборах данных, построенных на этом принципе, практически отсутствуют сложные атаки. В данном случае, благодаря природе сбора данных, не требуется длительное наблюдение и удаление конфиденциальных данных. Большая инфраструктура и различия в подходах команд-участниц позволили получить различные атаки, реализованные разными инструментами. Использование копий инфраструктуры для команд дало возможность провести точную разметку данных относительно принадлежности к командам, представляющим различные группы атакующих.

В наборе представлены данные о срабатывании системы обнаружения вторжений, статистические данные по трафику, метрики, собранные с конечных устройств и другие данные. Согласие на публикацию своих атак дали 7 команд, участвовавших в соревнованиях. Для проверки предлагаемого в данной работе метода использованы данные системы обнаружения вторжений Suricata[1]. В наборе представлено более 330 тыс. срабатываний, относящихся к 225 видам и 14 категориям атак.

При обработке набора данных дополнительно произведем следующие действия: данные команд объединим в единый набор с добавлением метки команды; удалим следующие события: с IP-адресом назначения 169.254.169.254, который используется для управления инфраструктурой организаторов; из категории «Not Suspicious Traffic»; не относящиеся к протоколу HTTP.

В результате обработки осталось 89 тыс. событий безопасности. Переведем нагрузку из кодировки Base64 в текст, затем применим URL-декодирование. Для последующего исследования возьмем колонки с метками команд и HTTP-нагрузку.

После токенизации и применения TF-IDF осуществим оценку плотности распределения расстояний запросов от начала координат. Для дальнейшей работы отберем запросы вида:

$$dist_{Manh}(x_i, 0) > 5{,}3.$$

Полученный график плотности показан на рисунке.

Сдвиг границы вправо приводит к уменьшению числа рассматриваемых запросов, тогда как сдвиг влево ведет к уменьшению показателей оценки, получаемой в итоге кластеризации. В выборку со значением границы 5,3 попало 17 тыс. запросов.

К полученным данным применим различные методы кластеризации. Для оценки эффективности методов используем метрики: однородность, полнота и V-мера [13].

Определим однородность (h) (гомогенность), которая определяет, насколько элементы в кластерах схожи друг с другом:

$$h = 1 - \frac{H(C|K)}{H(C)},$$

где C и K – эталонная и полученная кластеризации; $H(C)$ — энтропия; $H(C|K) = -\sum_{c,k} \frac{n_{ck}}{N} \log\left(\frac{n_{ck}}{n_k}\right)$; $H(C) = -\sum_{c} \frac{m_c}{N} \log\left(\frac{m_c}{N}\right)$, $H(C|K)$ — условная энтропия, N — общее количество объектов в выборке; n_k — количество объектов в кластере k; m_c — количество объектов с меткой c; n_{ck} — количество объектов с меткой c в кластере k.

[1] [Электронный ресурс]. Режим доступа: https://suricata.io/ (дата обращения: 31.03.2023).

Таблица. Оценка различных методов кластеризации
Table. Clustering methods evaluation

Метод	Гиперпараметры	Метрики, %		
		Однородность	Полнота	V-мера
Спектральная кластеризация [14]	$N = 8$	0,7	**0,45**	**0,55**
DBSCAN (Density-based Spatial Clustering of Applications with Noise) [15]	*eps*=1, *min_samples*=1	**0,98**	0,04	0,08

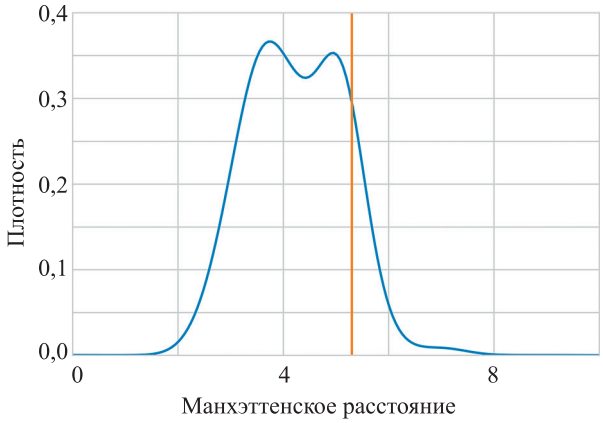

Рисунок. График плотности распределения манхэттенского расстояния запросов от начала координат с указанием границы отсечки
Figure. Density plot of the Manhattan distance of requests from the origin indicating the cutoff boundary

Таким образом, значение однородности показывает степень изменения энтропии за счет известной принадлежности объектов к выделенным алгоритмом кластерам. Лучший результат отражает значение $h = 1$, тогда каждый кластер содержит только элементы одного класса.

Рассчитаем полноту по следующей формуле:

$$c = 1 - \frac{H(C|K)}{H(K)}.$$

Полнота отражает значение доли событий с определенный меткой, которая попадает в один кластер. Лучший результат получим в том случае, если все значения с одной меткой находятся в одном кластере.

Поскольку однородность и полнота, как и точность и полнота при классификации, могут принимать значение 1 в том случае, если число кластеров равно числу событий (тогда максимальна однородность) или если число кластеров равно 1 (тогда максимальна полнота), для оценки сбалансированности этих параметров используем гармоническую V-меру, которая имеет вид

$$V = 2\frac{h \times c}{h + c}.$$

В таблице приведены лучшие методы с точки зрения однородности получаемой кластеризации и V-меры. Выбор гиперпараметров осуществим для исследуемого набора данных. При указанных в таблице значениях предложенный в данной работе метод показал свою эффективность. Задачу оптимизации значений гиперпараметров решим отдельно при решении практических задач.

Несмотря на то, что отсутствуют общепринятые границы показателей, при которых оцениваемые методы эффективны, исходя из особенностей решаемой задачи можно сделать предположение о том, что наибольшую практическую применимость будут иметь подходы, обладающие высокой однородностью. Это связано с тем, что наличие в полученных кластерах большого количества неверно отмеченных событий приведет к неверной оценке возможностей и глубины проникновения атакующих и, соответственно, недостаточности или избыточности предпринимаемых мер. Более того, результаты кластеризации множества методов с высокой однородностью можно объединить в комплексном методе без существенной потери однородности. Таким образом, кластеризация при помощи DBSCAN — наиболее подходящий метод кластеризации при анализе данных полезной нагрузки HTTP-запросов предложенным методом среди всех рассмотренных.

Заключение

Предложен метод выявления групп атакующих на основании токенизации данных нагрузки HTTP и применения алгоритма TF-IDF. Показано, что высокая эффективность работы метода достигается при использовании DBSCAN в качестве метода кластеризации. Улучшить результат кластеризации возможно, объединив полученные кластеры с кластерами других методов с высоким показателем однородности.

Предложенный метод может быть применен при выделении групп атакующих при анализе других текстовых протоколов, таких как FTP или LDAP. Отметим, что значение регулярного выражения, применяемого для токенизации, требует уточнения с учетом особенностей конкретного рассматриваемого протокола. Количество и разнообразие атак, проводимых через подобные протоколы, существенно меньше, чем для HTTP. Метод также применим для анализа HTTPS-трафика, при условии настройки передачи трафика в IDS через обратный прокси или файрвол.

В рамках дальнейших исследований возможна разработка комплексного метода выявления групп атакующих, включающего предложенный, а также проверка предложенного метода на данных реальных атак.

Литература

1. Hassan W., Guo S., Li D., Chen Z., Jee K., Li Z., Bates A. NoDoze: Combatting threat alert fatigue with automated provenance triage // Proc. of the 2019 Network and Distributed System Security Symposium. 2019. https://doi.org/10.14722/ndss.2019.23349
2. Pavlov A., Voloshina N. Analysis of IDS alert correlation techniques for attacker group recognition in distributed systems // Lecture Notes in Computer Science. 2020. V. 12525. P. 32–42. https://doi.org/10.1007/978-3-030-65726-0_4
3. Kotenko I., Gaifulina D., Zelichenok I. Systematic literature review of security event correlation methods // IEEE Access. 2022. V. 10. P. 43387–43420. https://doi.org/10.1109/access.2022.3168976
4. Mirheidari S.A., Arshad S., Jalili R. Alert correlation algorithms: A survey and taxonomy // Lecture Notes in Computer Science. 2013. V. 8300. P. 183–197. https://doi.org/10.1007/978-3-319-03584-0_14
5. Navarro J., Deruyver A., Parrend P. A systematic survey on multi-step attack detection // Computers & Security. 2018. V. 76. P. 214–249. https://doi.org/10.1016/j.cose.2018.03.001
6. Zhan J., Liao X., Bao Y., Gan L., Tan Z., Zhang M., He R., Lu J. An effective feature representation of web log data by leveraging byte pair encoding and TF-IDF // Proc. of the ACM Turing Celebration Conference — China (ACM TURC ‘19). 2019. P. 62. https://doi.org/10.1145/3321408.3321568
7. Qi B., Shi Z., Wang Y., Wang J., Wang Q., Jiang J. BotTokenizer: Exploring network tokens of HTTP-based botnet using malicious network traces // Lecture Notes in Computer Science. 2018. V. 10726. P. 383–403. https://doi.org/10.1007/978-3-319-75160-3_23
8. Chen R.-C., Chen S.-P. Intrusion detection using a hybrid support vector machine based on entropy and TF-IDF // International Journal of Innovative Computing, Information & Control (IJICIC). 2008. V. 4. N 2. P. 413–424.
9. Павлов А.В. Анализ сетевого взаимодействия современных эксплойтов // Информационные технологии. 2022. Т. 28. № 2. С. 75–80. https://doi.org/10.17587/it.28.75-80
10. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. V. 24. N 5. P. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
11. Aggarwal C., Hinneburg A., Keim D. On the surprising behavior of distance metrics in high dimensional space // Lecture Notes in Computer Science. 2001. V. 1973. P. 420–434. https://doi.org/10.1007/3-540-44503-x_27
12. Muniah N., Pelletier J., Su S.-H., Yang S.J., Meneely A. A cybersecurity dataset derived from the national collegiate penetration testing competition // Proc. of the HICSS Symposium on Cybersecurity Big Data Analytics. 2019.
13. Rosenberg A., Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure // Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007. P. 410–420.
14. Shi J., Malik J. Normalized cuts and image segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. V. 22. N 8. P. 888–905. https://doi.org/10.1109/34.868688
15. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD). 1996. P. 226–231.

Авторы

Павлов Артем Валерьевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0001-8567-5469, artempavlov1@gmail.com

Волошина Наталия Викторовна — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 55511854200, https://orcid.org/0000-0001-9435-9580, nataliv@yandex.ru

References

1. Hassan W., Guo S., Li D., Chen Z., Jee K., Li Z., Bates A. NoDoze: Combatting threat alert fatigue with automated provenance triage. *Proc. of the 2019 Network and Distributed System Security Symposium*, 2019. https://doi.org/10.14722/ndss.2019.23349
2. Pavlov A., Voloshina N. Analysis of IDS alert correlation techniques for attacker group recognition in distributed systems. *Lecture Notes in Computer Science*, 2020, vol. 12525, pp. 32–42. https://doi.org/10.1007/978-3-030-65726-0_4
3. Kotenko I., Gaifulina D., Zelichenok I. Systematic literature review of security event correlation methods. *IEEE Access*, 2022, vol. 10, pp. 43387–43420. https://doi.org/10.1109/access.2022.3168976
4. Mirheidari S.A., Arshad S., Jalili R. Alert correlation algorithms: A survey and taxonomy. *Lecture Notes in Computer Science*, 2013, vol. 8300, pp. 183–197. https://doi.org/10.1007/978-3-319-03584-0_14
5. Navarro J., Deruyver A., Parrend P. A systematic survey on multi-step attack detection. *Computers & Security*, 2018, vol. 76, pp. 214–249. https://doi.org/10.1016/j.cose.2018.03.001
6. Zhan J., Liao X., Bao Y., Gan L., Tan Z., Zhang M., He R., Lu J. An effective feature representation of web log data by leveraging byte pair encoding and TF-IDF. *Proc. of the ACM Turing Celebration Conference — China (ACM TURC ‘19)*, 2019, pp. 62. https://doi.org/10.1145/3321408.3321568
7. Qi B., Shi Z., Wang Y., Wang J., Wang Q., Jiang J. BotTokenizer: Exploring network tokens of HTTP-based botnet using malicious network traces. *Lecture Notes in Computer Science*, 2018, vol. 10726, pp. 383–403. https://doi.org/10.1007/978-3-319-75160-3_23
8. Chen R.-C., Chen S.-P. Intrusion detection using a hybrid support vector machine based on entropy and TF-IDF. *International Journal of Innovative Computing, Information & Control (IJICIC)*, 2008, vol. 4, no. 2, pp. 413–424.
9. Pavlov A.V. Analysis of network interaction of modern exploits. *Information Technologies*, 2022, vol. 28, no. 2, pp. 75–80. (in Russian). https://doi.org/10.17587/it.28.75-80
10. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 1988, vol. 24, no. 5, pp. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
11. Aggarwal C., Hinneburg A., Keim D. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 2001, vol. 1973, pp. 420–434. https://doi.org/10.1007/3-540-44503-x_27
12. Muniah N., Pelletier J., Su S.-H., Yang S.J., Meneely A. A cybersecurity dataset derived from the national collegiate penetration testing competition. *Proc. of the HICSS Symposium on Cybersecurity Big Data Analytics*, 2019.
13. Rosenberg A., Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 410–420.
14. Shi J., Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, vol. 22, no. 8, pp. 888–905. https://doi.org/10.1109/34.868688
15. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

Authors

Artem V. Pavlov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0001-8567-5469, artempavlov1@gmail.com

Natalia V. Voloshina — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 55511854200, https://orcid.org/0000-0001-9435-9580, nataliv@yandex.ru

Статья поступила в редакцию 27.02.2023
Одобрена после рецензирования 31.03.2023
Принята к печати 25.05.2023

Received 27.02.2023
Approved after reviewing 31.03.2023
Accepted 25.05.2023