УДК 004.738
DOI: 10.25729/ESI.2023.32.4.015

Выделение групп используемых Интернет-ресурсов для обнаружения внутренних источников киберугроз

Исаев Сергей Владиславович, Донцов Денис Юрьевич
Институт вычислительного моделирования СО РАН,
Россия, Красноярск, *si@icm.krasn.ru*

Аннотация. Защищенность корпоративной сети является важным аспектом успешного функционирования организации. В данной работе исследуется безопасность внутреннего периметра сети на примере Красноярского научного центра СО РАН. Существуют различные средства для предотвращения киберугроз и анализа посещаемых Интернет-ресурсов, но их быстродействие и возможность применения сильно зависят от объема входных данных. В статье рассматриваются существующие методы определения сетевых угроз с помощью анализа журналов прокси-сервера. Исследовано разделение интернет-пользователей на тематические группы для выявления аномалий. Предложен метод кластеризации Интернет-ресурсов, направленный на снижение объема входных данных путем исключения групп безопасных Интернет-ресурсов или выбора только подозрительных Интернет-ресурсов. Предложенный метод состоит из этапов: предобработка данных, выделение сессий пользователей, анализ данных и интерпретация полученных результатов. Исходными данными являются записи журнала прокси-сервера. На первом этапе из исходных данных выбираются полезные для анализа данные, после чего непрерывный поток данных делится на небольшие порции (сессии) при помощи метода ядерной оценки плотности. На втором этапе выполняется мягкая кластеризация используемых Интернет-ресурсов путем применения метода тематического моделирования. Результатом второго этапа являются неразмеченные группы Интернет-ресурсов. На третьем этапе, с помощью эксперта, происходит интерпретация полученных результатов путем анализа наиболее популярных Интернет-ресурсов в каждой группе. Метод имеет множество настроек на каждом этапе, что позволяет сконфигурировать его под любой формат и специфику входных данных. Область применения метода не ограничена. Он может быть использован как в качестве дополнительного шага предобработки с целью снижения количества входных данных, так и при выявлении аномальных данных.

Ключевые слова: кластерный анализ, тематическое-моделирование, кибербезопасность

Цитирование: Исаев С.В. Выделение групп используемых Интернет-ресурсов для обнаружения внутренних источников киберугроз / С.В. Исаев, Д.Ю. Донцов // Информационные и математические технологии в науке и управлении. – 2023. – № 4(32). – С. 170-181. – DOI:10.25729/ESI.2023.32.4.015.

Введение. Повсеместное внедрение информационно-телекоммуникационных технологий в жизнь людей и переход на цифровую экономику требуют повышенного внимания к проблемам кибербезопасности. Выполнение профессиональных обязанностей в настоящее время в большой мере связано с использованием информационно-телекоммуникационных ресурсов. Современные текстовые редакторы, установленные непосредственно на компьютере, требуют подключения к сети Интернет для загрузки справочных материалов, шаблонов оформления или обновления программных модулей. Даже в быту уже сложно себе представить жизнь без различных электронных устройств, облегчающих процесс получения информации и обеспечивающих коммуникации между людьми. Во многих производственных процессах используется сеть Интернет для обеспечения сервисного обслуживания и удаленного контроля состояния оборудования. Повсеместное внедрение технологии «интернета вещей» повышает риск несанкционированного доступа к цифровым активам за счет недостаточной защиты или некорректных настроек отдельных устройств. В связи с массовостью явления невозможно в каждом случае привлечь квалифицированного эксперта в области кибербезопасности для выявления и предотвращения возможных рисков. Необходимостью становятся средства защиты, ко-

торые не требуют специальных знаний по настройке и способны противодействовать изменяющимся угрозам. Этим критериям соответствуют антивирусы, встроенные средства защиты, сканеры безопасности, системы предотвращения вторжений и прочие средства защиты, которые получают периодические обновления от разработчиков. Они решают множество задач, но, к сожалению, не охватывают весь спектр проблем кибербезопасности.

Можно выделить три основных класса источников киберугроз: связанные с действиями человека, проблемы функционирования техники и связанные с форс-мажорными ситуациями [1]. В большинстве случаев именно человек является причиной возникновения киберугроз [2], поэтому перспективным направлением является создание методов и алгоритмов, помогающих снизить риск вторжений по вине человека. Чтобы блокировать доступ к потенциально опасным Интернет-ресурсам, многие крупные организации используют их фильтрацию [3]. Это позволяет существенно снизить риск киберугроз, но может создать проблемы доступа к нужным ресурсам и не дает 100% защиту. В связи с этим, создание дополнительных средств защиты, учитывающих особенности и сценарии использования Интернет в конкретной организации, является актуальным.

Анализ кибербезопасности с точки зрения внутреннего периметра сети требует получения и сохранения в пригодном для обработки виде запросов пользователей и устройств сети, метаданных о структуре сети и т.д. В результате анализа можно определить аномалии сценариев взаимодействия в сети, идентифицировать и оперативно блокировать угрозы [4-6]. В процессе работы телекоммуникационных программ и протоколов образуется большое число всевозможных журналов функционирования сервисов и использования Интернет-ресурсов. Они располагаются как на устройствах пользователей, так и на телекоммуникационном оборудовании. Суммарный объем этих данных для крупной организации за один день составляет десятки Гигабайт. В большинстве работ используются методы анализа, основанные на кластеризации [7, 8] или машинном обучении [9, 10]. Выделение части данных, полезных для анализа безопасности, позволяет снизить время их обработки, увеличить временной диапазон для анализа, что позволяет повысить оперативность и качество решений по защите корпоративной сети. Уменьшить объем сохраняемых для последующего анализа данных можно за счет группировки посещаемых пользователями ресурсов и определения среди них условно безопасных и потенциально опасных. Кроме того, подобная группировка полезна с точки зрения обнаружения аномальных групп ресурсов, сигнализирующих о потенциальных киберугрозах. Множество имеющихся коммерческих сетевых анализаторов (Kaspersky Anti Targeted Attack, PT Network Attack Discovery, Threat Detection System и др.) нацелены в первую очередь на обнаружение и предотвращение атак из внешнего периметра.

В данной работе исследуется возможность анализа данных запросов пользователей к ресурсам сети Интернет с целью идентификации рисков кибербезопасности и сокращения исходного набора данных. Предложена методика мягкой кластеризации посещаемых пользователями Интернет-ресурсов на основе применения тематического моделирования данных журналов доступа прокси-сервера. Основная цель состоит в исключении из дальнейшего анализа найденных групп с безопасным контентом и выделении аномальных групп, несущих повышенные риски кибербезопасности.

Источники данных. В данном исследовании за основу были взяты данные ежедневных журналов прокси-сервера корпоративной сети Красноярского научного центра СО РАН. Сбором и анализом данных занимается отдел информационно-телекоммуникационных технологий Института вычислительного моделирования СО РАН, который более двух десятков лет обеспечивает доступ пользователей к различным Интернет-сервисам. За это время были

накоплены значительные объемы данных, проводились исследования различных аспектов информационной безопасности корпоративной сети научного центра [11]. На основе анализа данных были выявлены источники киберугроз и предложены эффективные способы их блокирования [12]. В большинстве исследований анализируются лишь внешние угрозы кибербезопасности, источники которых расположены в неконтролируемой зоне сети Интернет, но угрозы безопасности внутри корпоративной сети тоже несут значительные риски и не имеют единой точки защиты. К таким угрозам можно отнести: вирусы, попадающие во внутреннюю сеть на флеш-накопителях и дисках, мобильные компьютеры и электронные устройства, бесконтрольно подключаемые ко внутренней сети. Если на таком устройстве имеется вирус или некорректно настроенная программа, то обнаружить их можно по нетипичным действиям в сети, в том числе попыткам получить доступ в Интернет. Так как рассматриваемые устройства и находящиеся на них программы не используют штатные для сети настройки, то их возможно идентифицировать и блокировать путем анализа сетевых запросов и выявления аномалий.

Одним из множества источников данных об использовании ресурсов сети Интернет работниками организации может служить журнал прокси-сервера, который выступает посредником между веб-браузером и веб-сервером или другим Интернет-сервисом. Журналы прокси-сервера содержат данные о доступе пользователей к сети Интернет. На основе их анализа можно решать задачи оптимизации телекоммуникационной подсистемы, улучшения защиты, выявления нетипичных запросов и аномальной активности устройств.

Рассматриваются данные ежедневных журналов прокси-сервера за месяц, общий объем которых составил более 2 Гигабайт. Суммарное количество сохраненных запросов к Интернет-ресурсам превышает десять миллионов, что выводит используемые для анализа алгоритмы в область Big-data. Журнал представляет собой текстовый файл, в каждой строке которого содержатся следующие атрибуты:

- время запроса по часам прокси-сервера;
- длительность отклика;
- IP-адрес клиента источника запроса;
- статус запроса прокси сервера;
- HTTP-код состояния, отправленный клиенту;
- размер переданных клиенту данных;
- метод совершения запроса;
- унифицированный адрес ресурса;
- идентификатор пользователя;
- тип содержимого запроса по MIME.

На основе группировки строк по идентификатору пользователя и IP-адресу клиента можно для каждого источника запросов вычислить дополнительные метаданные, такие, как: время начала и конца активности, продолжительность активности, средняя частота и размер полученных данных.

Выделение групп пользователей. Целью анализа являлось выделение тематических групп пользователей Интернет-ресурсов. Предпосылкой появления угроз кибербезопасности в данной постановке является изменение тематической группы пользователем или наличие аномально малых групп. Последовательность действий для анализа представлена на схеме (рисунок 1).

Для дальнейшей обработки использовались записи об успешно полученных ресурсах, имеющие код 200, так как в случае недоступного ресурса невозможно судить о его содержании и, в дальнейшем, проверить правильность классификации. Из них были выбраны параметры с

унифицированным адресом ресурса (URL) и идентификатором пользователя (UserName). Были созданы: список пользователей, список ресурсов и таблица, связывающая пользователей, ресурсы и количество посещений. После обработки журналов за семь дней объем данных уменьшился примерно на три порядка с 2 ГБ до 3 МБ. Полученный таким образом набор данных для одного пользователя может быть интерпретирован как точка в многомерном пространстве, где оси являются ресурсами, а количество посещений ресурса – координатой точки на оси. Для не посещённых ресурсов некоторым пользователем (отсутствующие данные) принимается значение координаты равным 0.

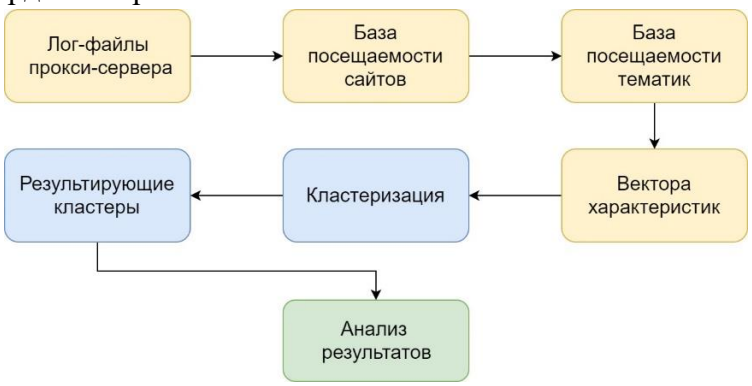

Рис. 1. Последовательность выделения групп пользователей

Ввиду того, что ресурсов, имеющих различный URL, очень много, были просуммированы ресурсы с одинаковым доменным именем, что позволило снизить размерность пространства до 4000. Для решения задачи дальнейшего понижения количества показателей было решено произвести свертку и перейти от доменных имен Интернет-ресурсов к тематикам сайтов. Помимо снижения размерности, это позволяет получить для каждого пользователя его предпочтительные тематики Интернет-ресурсов.

Чтобы решить задачу определения тематики сайтов, был использован сервис Яндекс-Каталог (https://yandex.ru/yaca), позволяющий определить тематическую рубрику для большинства сайтов. К моменту начала исследований это был наиболее полный и удобный для автоматизированного использования сервис. К сожалению, тематики некоторых не популярных или малопосещаемых сайтов не определяются данным сервисом, поэтому их пришлось исключить из дальнейшего анализа, при этом какие-то полезные для решения задачи данные были потеряны, но была достигнута полная автоматизация процесса. В результате, получившийся набор данных имеет порядка 160 атрибутов, что на три десятичных порядка меньше первичного количества различных Интернет-ресурсов. Для такой размерности могут эффективно применяться различные методы кластеризации. Таблица 1 содержит пример фрагмента данных с количеством посещений по тематикам.

Таблица 1. Число посещений по тематикам

Идентификатор пользователя	Тематики из «Яндекс.Каталог»			
	Интернет	Социальные сети	Баннерные сети	Банки
4	3102	58	0	0
7	227	8	73	8
5	142	2	0	0
6	113	14	226	56
2	80	1	125	0
3	78	21	21	15
1	24	7	0	0

Для применения методов кластерного анализа потребовалась нормализация всех характеристик к единому диапазону. После нормализации всех характеристик эти данные можно представить в виде многомерного пространства, в котором каждая координата – это значение показателя. При кластеризации существенным является способ задания расстояния между отдельными кластерами. Были рассмотрены несколько алгоритмов вычисления расстояния, в результате наиболее подходящим был признан метод минимальной дисперсии Уорда. Расстояние вычисляется по формуле:

$$d(U,V) = \sqrt{\frac{|V|+|S|}{Q} * d(V,S) + \frac{|V|+|T|}{Q} * d(V,T) - \frac{|V|}{T} * d(S,T)^2},$$

где кластер U был получен путем объединения кластеров T и S, а $Q = |V| + |T| + |S|$, под операцией $|*|$ понимается нахождение количества элементов в соответствующем кластере.

Для получения групп применялся метод иерархической кластеризации (рисунок 2).

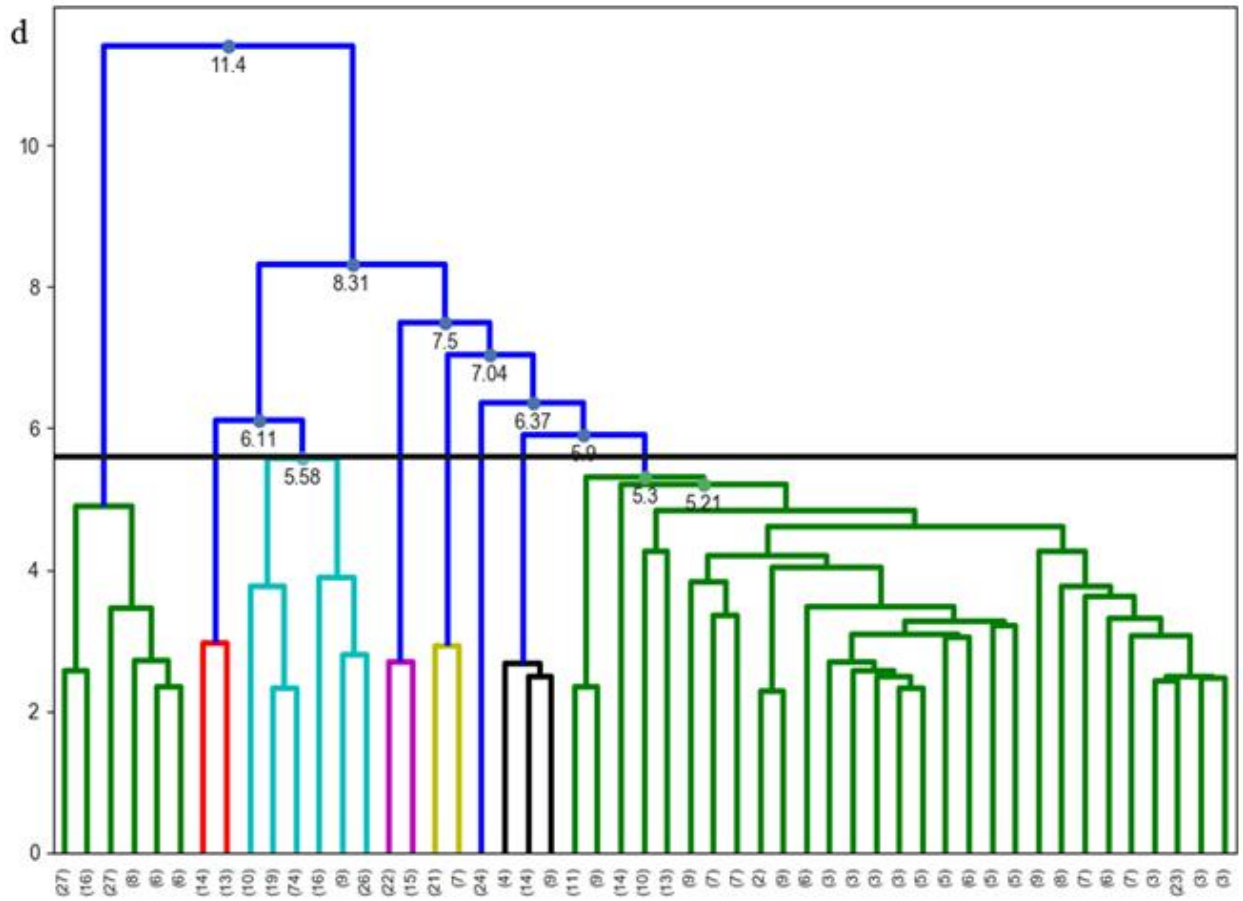

Рис. 2. Дендрограмма иерархической кластеризации источников

Для визуализации полученных групп были использованы методы главных компонент (PCA) и линейного дискриминантного анализа (LDA). Для количества кластеров 3 метод LDA показал более наглядные результаты (рисунок 3), при большем количестве кластеров оба метода не позволяют разграничить кластеры на плоскости.

Для проверки гипотезы о постоянстве предпочтений пользователей произведено сравнение полученных групп за разные промежутки времени. Это необходимо для подтверждения того, что пользователи из одного кластера сохраняют схожие предпочтения за определенный период времени. Смена предпочтений при этом сигнализировала бы о аномалии. В результате анализа групп за несколько интервалов времени был сделан вывод о наличии постоянных

групп, в которые входят более половины пользователей. Таким образом, признаком возможной угрозы при таком способе группировки остается только появление групп из единичных пользователей при небольшой глубине кластеризации. Кроме того, охват посещаемых сайтов сервиса Яндекс-Каталог недостаточен, в связи с чем существенная часть ресурсов выпала из рассмотрения.

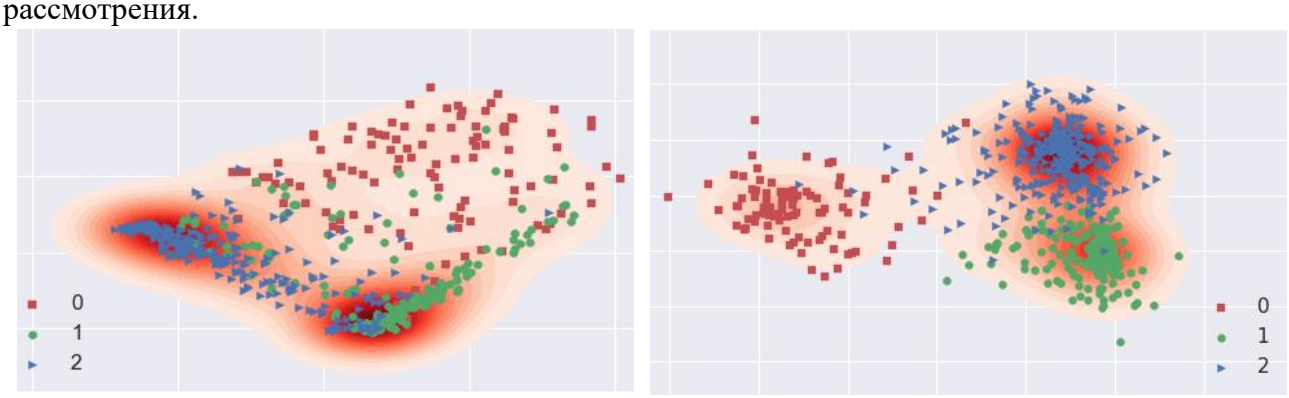

Рис. 3. Результат визуализации 3-х кластеров методами PCA и LDA

Выделение групп Интернет-ресурсов. Для устранения зависимости от внешних классификаторов и включения в рассмотрение всех Интернет-ресурсов был исследован подход по установлению зависимостей между посещаемыми Интернет-ресурсами в пределах некоторых сессий и получению групп Интернет-ресурсов, имеющих зависимости в различных сессиях. Сессией будем называть совокупность последовательно посещенных пользователем Интернет-ресурсов за некоторый период активности. В простейшем случае в качестве сессии можно выбирать одни сутки, но очевидно, что за сутки у пользователя как правило есть несколько сеансов работы с Интернет-ресурсами, и в разные сеансы могут решаться различные задачи. Для улучшения качества кластеризации нужно использовать и прочие способы определения сессий.

Для решения задачи разделения Интернет-ресурсов на группы путем анализа их принадлежности одной сессии будем использовать вероятностное тематическое моделирование [13]. Вероятностные тематические модели осуществляют «мягкую» кластеризацию, позволяя документу или термину относиться сразу к нескольким темам с различными вероятностями. Документом в нашем случае является множество Интернет-ресурсов, посещенных пользователем в течение сессии. Термином является сам Интернет-ресурс, независимо от времени и пользователя, его посетившего. В результате моделирования мы получаем наборы Интернет-ресурсов, которые совместно встречались в разных сессиях. Каждому такому набору можно подобрать условную тематику. Например, если в исходном журнале присутствуют записи:

$$R_{1,1}, R_{2,1}, R_{3,1}, R_{4,2}, R_{5,2}, R_{6,1}, R_{7,1}, R_{8,2}, R_{9,2}, R_{10,1},$$

где первый индекс – номер записи, а второй – номер пользователя, то после разбиения на сессии мы можем получить:

- $S_{1,1} = (R_{1,1}, R_{2,1}, R_{3,1})$ – первая сессия первого пользователя;
- $S_{1,2} = (R_{4,2}, R_{5,2}, R_{8,2}, R_{9,2})$ – первая сессия второго пользователя;
- $S_{2,2} = (R_{6,1}, R_{7,1}, R_{10,1})$ – вторая сессия первого пользователя.

Результатом тематического моделирования является набор неименованных групп (тематик), включающих расположенные в порядке количества встречаемости Интернет-ресурсы: $T_1 = (R_{1,1}, R_{1,2}, R_{1,3}, \ldots)$, $T_2 = (R_{2,1}, R_{2,2}, R_{2,3}, \ldots)$, $T_3 = (R_{3,1}, R_{3,2}, R_{3,3}, \ldots)$, ….

Эксперт может на основании содержания наиболее популярных Интернет-ресурсов группы присвоить ей тематику и определить степень опасности. Безопасные группы мы можем удалить из дальнейшего анализа, а самые опасные исследовать более детально. Задача решалась в несколько этапов.

Этап 1 – Предобработка данных. Современные веб-сайты устроены таким образом, что, открывая один адрес, браузер запрашивает, помимо основной ссылки, еще 10-20 других, в которых могут запрашиваться как необходимые для отображения страницы элементы, так и не имеющие визуального представления счетчики и элементы программного кода. В среднем за рабочий день в журнале прокси-сервера регистрируется около четырех тысяч запросов от каждого пользователя. Целью предобработки является удаление из набора данных, не несущих полезной для дальнейшего анализа информации, за счет чего удается повысить качество результатов и скорость основной фазы анализа [14].

Из имеющихся данных журналов были исключены записи, касающиеся анонимного доступа для разрешенного технического трафика (обновления известных приложений и сервисов) и ресурсы типа css/js/image, которые, как правило, не несут смысловой нагрузки или ее сложно определить автоматически без привлечения эксперта. За счет этого объем данных был снижен примерно в 5 раз. Включение в анализ ресурсов JavaScript требует отдельного рассмотрения и лежит вне пределов данной работы.

Следующим шагом предобработки было выделение из URL частей, содержащих полный домен или IP-адрес использованного ресурса, для суммирования числа посещений ресурсов с одинаковой доменной частью.

Этап 2 – Выделение сессий пользователей. Требуется сгруппировать данные каждого пользователя в сессии с непересекающимися временными интервалами. Имеется несколько возможных алгоритмов получения таких наборов, отличающихся сложностью и достоверностью полученных данных.

1. Наборы с постоянным периодом. Для выделения таких сессий нужно задать постоянный временной интервал и выделить данные пользователя внутри него. При использовании такого подхода в одну сессию могут попадать периоды активности пользователя небольшой длительности. Например, утром пользователь работал с почтовой программой, а вечером того же дня читал новости. При длине периода 24 часа это будет считаться одной сессией, что не согласуется с реальностью.
2. Определение периодов отсутствия активности пользователя и выделений сессий между этими периодами. Такой подход генерирует сессии произвольной длины, в зависимости от определения минимального периода неактивности. Существенным недостатком метода является его неработоспособность при наличии постоянных фоновых процессов, таких, как работа мессенджера, почтового клиента и прочих программ, периодически опрашивающих интернет-ресурс, независимо от активности пользователя.

Для устранения проблем перечисленных подходов предложено использовать метод ядерной оценки плотности (Kernel Density Estimation – KDE), основанный на непараметрической оценке плотности случайной величины [15-17]. Этот метод оценивает плотность распределения одномерного набора данных и обнаруживает точки локального экстремума. Используя, например, точки минимума в качестве границ интервалов активности, мы получаем сессии различной длины, соответствующие реальной активности пользователя. На исследуемых данных средняя длина выделенной сессии составила 4.5 минуты.

Для применения данного метода необходимо задать два параметра – размер ядра (в нашем случае это количество ресурсов, посещенных пользователем за сессию), и ширина полосы (диапазон времени для агрегирования). Изменение этих параметров существенно влияет

на количество выделенных сессий. Их значения подбирались опытным путем, на основе экспертной оценки достоверности размеров получаемых сессий. На рисунке 4 показана гистограмма распределения среднего размера ядра среди всех пользователей. По горизонтальной оси размер ядра, по вертикальной – количество пользователей.

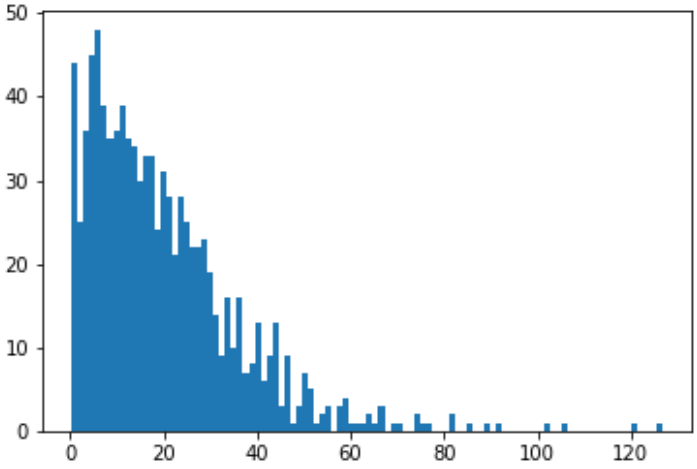

Рис. 4. Гистограмма распределения среднего размера ядра

Этап 3 – Тематическое моделирование. Решив задачу разделения данных журналов на сессии, мы получили все начальные данные для применения тематического моделирования для мягкой кластеризации наших сессий, состоящих из Интернет-ресурсов. В настоящее время известны и используются различные методы тематического моделирования, такие, как: методы сингулярного разложения (SVD), метод моментов, вероятностный латентно-семантический анализ (PLSA) и др. [18-19]. В данной работе использовался наиболее распространённый и хорошо описанный метод латентного размещения Дирихле (LDA) [20-21]. Применение метода LDA для полученных сессий позволяет мягко сгруппировать все Интернет-ресурсы на заданное количество групп, которое должно определяться опытным путем, в зависимости от ожидаемого результата. В таблице 2 приведен результат тематического моделирования ресурсов, посещенных пользователями за один месяц. Ресурсы разделены на 5 групп и расположены в порядке убывания силы принадлежности к группе.

Таблица 2. Результат тематического моделирования для 5 групп

№ группы	Ресурсы в порядке убывания принадлежности группе
1	newslab.ru, 4pda.ru, sfkras.ru, edu.sfu-kras.ru, worldcrisis.ru, libgen.is
2	nowa.cc, ugadalki.ru, scask.ru, forum.rcmir.com, 2baksa.net, autoopt.ru
3	update.eset.com, law-college-sfu.ru, kinoaction.ru, kiwt.ru, dostavka-krasnoyarsk.ru
4	apps.webofknowledge.com, packages.linuxmint.com, http.debian.net, urod.ru, fips.ru, mc.corel.com
5	fitohobby.ru, ib.adnxs.com, allrefs.net, ckp-rf.ru, teammodels.no, profinance.ru

На рисунке 5 представлен результат тематического моделирования с числом групп 30, с проекцией на две главные компоненты. Размер окружности зависит от количества посещений элементов группы.

В результате проведенного тематического моделирования мы выделили потенциально опасные группы с малым количеством посещений и расположенные обособленно от прочих. В большинстве случаев это оказались ресурсы, связанные со специализированными задачами

по загрузке или синхронизации информации через Интернет. Получение информации о таких задачах позволило их упорядочить и регламентировать.

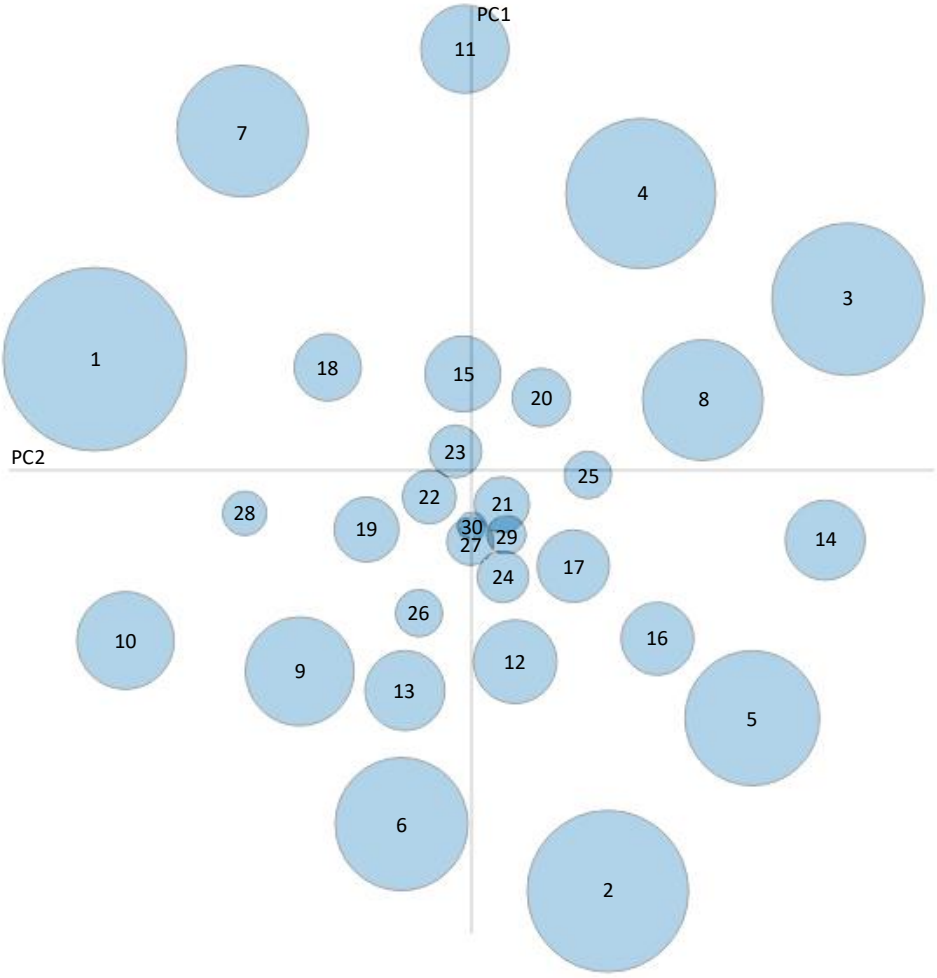

Рис. 5. Визуализация результатов тематического моделирования

Заключение. В работе рассмотрены два способа разбиения Интернет-ресурсов: с помощью внешнего классификатора и методом тематического моделирования. В первом случае в результате кластеризации мы получаем тематические группы пользователей с явно прописанными тематиками. На основании этого выявляются пользователи с аномальным поведением или посещающие ресурсы с потенциально вредоносной тематикой. К сожалению, использование внешнего классификатора накладывает существенные ограничения на возможности автоматического анализа неизвестных классификатору ресурсов, за счет чего возможно пропустить источники рисков. Перспективным представляется использование нескольких сервисов-классификаторов, которые бы покрывали большее множество Интернет-ресурсов, но даже в таком случае новые ресурсы будут в них отсутствовать.

Второй подход, выполняющий разбиение данных на сессии и тематическое моделирование, использует для анализа весь объем данных. Наличие настраиваемых параметров позволяет получать результаты как для организации с малым количеством пользователей и слабым каналом Интернет, так и для высоконагруженной корпоративной сети. Использование подхода возможно как на сетях без существующих аномалий, так и с функционирующими аномальными источниками, так как мы в любом случае получим группы ресурсов, которые должны будем классифицировать автоматически или с привлечением экспертов. Благодаря

мягкому разбиению Интернет ресурсов на тематики можно решать задачи: определение тематик Интернет-ресурсов, выявление сферы интересов пользователей, выделение сайтов с безопасной и несущей риски тематикой, а также определение и блокирование сервисов Интернет-рекламы.

Для улучшения результатов моделирования в дальнейшем планируется использование различного рода метаданных, таких, как длительность сессии, преимущественные виды контента, частота запросов. Описанный подход может быть использован для выявления внешних источников угроз за счет анализа журналов доступа к веб-ресурсам организации. Кластеризация источников с помощью тематического моделирования может помочь в выявлении источников угроз, которые не детектируются обычными системами обнаружения вторжений в связи с малыми временными рамками анализа.

Визуализация полученных групп с проекцией на главные компоненты позволяет получить общее представление о тематиках Интернет-запросов пользователей и, в отдельных случаях, выявить аномалии. Разработанные программные компоненты могут быть использованы при создании инструментальной системы обработки и анализа данных журналов Интернет-сервисов, а предложенные решения по противодействию обнаруженным угрозам встраиваться в функционирующие системы информационной безопасности.

Список источников

1. Mouna J., Latifa B., Latifa B.R., Anis A. Classification of security threats in information systems. Procedia Computer science, 2014, vol. 32, p. 489-496.
2. Дерендяев Д.А. Определение влияния человеческого фактора на основные характеристики угроз безопасности / Д.А. Дерендяев, Ю.А. Гатчин, В.А. Безруков // Кибернетика и программирование, 2019. – № 3. – С. 38-42.
3. Gyorodi R., Cornelia G., Pecherle G., Radu L. Network security using firewalls. Journal of Computer science and control systems, 2008, vol. 1.
4. Kao D.Y., Wang S.J., Huang F. Dataset Analysis of proxy logs detecting to curb propagations in network attacks. Intelligence and security informatics, 2008, pp. 245-250.
5. Marshall B., Chen H. Using importance flooding to identify interesting networks of criminal activity. Lecture notes in computer science, 2006, vol. 3975, pp. 14-25.
6. Mukkamala S., Sung A. Identifying significant features for network forensic analysis using artificial techniques. International journal of Digital evidence, 2003, vol. 1, no 4.
7. Bayraktar C., Karakaya Z., Gökçen H. Real-time anomaly detection system within the scope of smart factories. The journal of Supercomputing, 2023, vol. 79.
8. Wang C., Zhou H., Hao Z., Hu S., Li J., Zhang X., Jiang B., Chen X., Network traffic analysis over clustering-based collective anomaly detection. Computer networks, 2022, vol. 205, pp. 108760.
9. Yang T., Jiang Z., Liu P., Yang Q., Wang W. A traffic anomaly detection approach based on unsupervised learning for industrial cyber-physical system. Knowledge-Based Systems, 2023, vol. 279, pp. 110949.
10. Demertzis K., Tsiknas K., Taketzis D., Skianis C., Iliadis L. Darknet Traffic big-data analysis and network management for real-time automating of the malicious intent detection process by a weight agnostic neural networks framework. Electronics, 2021, vol. 10, pp. 781.
11. Исаев С.В. Кибербезопасность научного учреждения - активы и угрозы / С.В. Исаев // Информатизация и связь, 2015. – №1. – С. 53-57.
12. Исаев С.В. Анализ киберугроз и их источников для корпоративной сети Красноярского научного центра СО РАН / С.В. Исаев / Информационные и математические технологии в науке и управлении, 2016. – № 4-1. – С. 76-85.
13. Blei D.M. Probabilistic topic models. Communications of the ACM, 2012, vol. 55, no. 4, pp. 77-84.
14. Fei, B., Eloff, J., Oliver, M., Venter, H. Analysis of web proxy logs. IFIP international conference on digital forensics, Orlando, 2006, vol. 222, pp. 247-258.
15. Scott D.W. Multivariate density estimation. Theory. Practice and visualization: second edition. New York, 2015.
16. King T.L., Bentley R.J., Thornton L.E. [et al.] Using kernel density estimation to understand the influence of neighbourhood destinations on BMI. BMJ Open, 2016, vol. 6.

17. Kalinic M., Krisp J. Kernel density estimation (KDE) vs. hot-spot analysis - Detecting criminal hot spots in the city of San Francisco. Lund, Sweden, 2018.
18. Воронцов К.В. Обзор вероятностных тематических моделей, 2021.
19. Albalawi R., Yeap T., Benyoucef M. Using topic modeling methods for short-text data: A comparative analysis. Frontiers in artificial intelligence, 2020, vol. 3.
20. Jelodar H., Wang Y., Yuan, Ch., Xia, F. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey, 2017.
21. Tharwat A., Gaber T., Ibrahim A., Hassanien A.E. Linear discriminant analysis: A detailed tutorial. Ai communications, 2017, vol. 30, pp. 169-190.

Исаев Сергей Владиславович. *Кандидат технических наук, доцент, Институт вычислительного моделирования СО РАН, заместитель директора по научной работе, зав. отделом информационно-телекоммуникационных технологий. Область научных интересов: кибербезопасность и защита информации, интернет-технологии, распределённые информационные системы, интеллектуальные системы. AuthorID: 1623, SPIN: 1552-8542, si@icm.krasn.ru, Россия, 660036, Красноярск, Академгородок, дом. 50, стр. 44, ИВМ СО РАН.*

Донцов Денис Юрьевич. *Аспирант, Институт вычислительного моделирования СО РАН. Область научных интересов: машинное обучение, интернет-технологии. AuthorID: 1623, doncov.dy@icm.krasn.ru, Россия, 660036, Красноярск, Академгородок, дом. 50, стр. 44, ИВМ СО РАН.*

UDC 004.738
DOI: 10.25729/ESI.2023.32.4.015

Identification of groups of visited Internet resources for detection of internal cyberthreats source

Sergey V. Isaev, Denis Y. Doncov

Institute of Computational Modeling SB RAS, Russia, Krasnoyarsk, *si@icm.krasn.ru*

Abstract. The protection of the corporate network is an important aspect of the successful functioning of the organization. In this paper, the cybersecurity of the internal network perimeter is studied using the example of the Krasnoyarsk Scientific Center of the Siberian Branch of the Russian Academy of Sciences. There are various tools for preventing cyber threats and analyzing visited Internet resources, but their performance and applicability strongly depend on the amount of input data. The article discusses existing methods for identifying network threats by analyzing proxy server logs. The division of Internet users into thematic groups to detect anomalies is investigated. A method for clustering Internet resources is proposed, aimed at reducing the volume of input data by excluding groups of safe Internet resources or selecting only suspicious Internet resources. The proposed method consists of the following steps: data preprocessing, user session selection, data analysis, and interpretation of the results. The source data is the log entries of the proxy server. At the first stage, useful data for analysis are selected from the initial data, after which the continuous data stream is divided into small portions (sessions) using the kernel density estimation method. At the second stage, soft clustering of the used Internet resources is performed by applying the topic modeling method. The result of the second stage are unallocated groups of Internet resources. At the third stage, with the help of an expert, the results obtained are interpreted by analyzing the most popular Internet resources in each group. The method has many settings at each stage, which allows you to configure it for any format and specifics of the input data. The scope of the method is not limited. It can be used both as an additional preprocessing step to reduce the amount of input data and to detect anomalous data.

Keywords: cluster analysis, topic-modeling, cybersecurity

References

1. Mouna J., Latifa B., Latifa B.R., Anis A. Classification of security threats in information systems. Procedia Computer science, 2014, vol. 32, p. 489-496.
2. Derendyaev D.A., Gatchin Yu.A., Bezrukov V.A. Opredelenie vliyaniya chelovecheskogo faktora na osnovnye kharakteristiki ugroz bezopasnosti [Determining the influence of the human factor on the main characteristics of security threats]. Kibernetika i programmirovaniye [Cybernetics and programming], 2019, no. 3, pp. 38-42.

3. Gyorodi R., Cornelia G., Pecherle G., Radu L. Network security using firewalls. Journal of Computer science and control systems, 2008, vol. 1.
4. Kao D.Y., Wang S.J., Huang F. Dataset Analysis of proxy logs detecting to curb propagations in network attacks. Intelligence and security informatics, 2008, pp. 245-250.
5. Marshall B., Chen H. Using importance flooding to identify interesting networks of criminal activity. Lecture notes in computer science, 2006, vol. 3975, pp. 14-25.
6. Mukkamala S., Sung A. Identifying significant features for network forensic analysis using artificial techniques. International journal of Digital evidence, 2003, vol. 1, no 4.
7. Bayraktar C., Karakaya Z., Gökçen H. Real-time anomaly detection system within the scope of smart factories. The journal of Supercomputing, 2023, vol. 79.
8. Wang C., Zhou H., Hao Z., Hu S., Li J., Zhang X., Jiang B., Chen X., Network traffic analysis over clustering-based collective anomaly detection. Computer networks, 2022, vol. 205, pp. 108760.
9. Yang T., Jiang Z., Liu P., Yang Q., Wang W. A traffic anomaly detection approach based on unsupervised learning for industrial cyber-physical system. Knowledge-Based Systems, 2023, vol. 279, pp. 110949.
10. Demertzis K., Tsiknas K., Taketzis D., Skianis C., Iliadis L. Darknet Traffic big-data analysis and network management for real-time automating of the malicious intent detection process by a weight agnostic neural networks framework. Electronics, 2021, vol. 10, pp. 781.
11. Isaev S.V. Kiberbezopasnost' nauchnogo uchrezhdeniya - aktivy i ugrozy [Cybersecurity of a scientific institution - assets and threats]. Informatizatsiya i svyaz' [Informatization and communication], 2015, vol 1, p. 53-57.
12. Isaev S.V. Analiz kiberugroz i ikh istochnikov dlya korporativnoy seti Krasnoyarskogo nauchnogo tsentra SO RAN [Analysis of cyber threats and their sourceson the corporate network Krasnoyarsk Scientific Center of the SB RAS]. Informatsionnyie i matematicheskie tehnologii v nauke i upravlenii [Information and mathematical technologies in science and management], 2016. № 4-1. p. 76-85. (In Russ.)
13. Blei D.M. Probabilistic topic models. Communications of the ACM, 2012, vol. 55, no. 4, pp. 77-84.
14. Fei, B., Eloff, J., Oliver, M., Venter, H. Analysis of web proxy logs. IFIP international conference on digital forensics, Orlando, 2006, vol. 222, pp. 247-258.
15. Scott D.W. Multivariate density estimation. Theory. Practice and visualization: second edition. New York, 2015.
16. King T.L., Bentley R.J., Thornton L.E. [et al.] Using kernel density estimation to understand the influence of neighbourhood destinations on BMI. BMJ Open, 2016, vol. 6.
17. Kalinic M., Krisp J. Kernel density estimation (KDE) vs. hot-spot analysis - Detecting criminal hot spots in the city of San Francisco. Lund, Sweden, 2018.
18. Vorontsov K. V. Obzor veroyatnostnykh tematicheskikh modelei [Overview of probabilistic thematic models], 2021
19. Albalawi R., Yeap T., Benyoucef M. Using topic modeling methods for short-text data: A comparative analysis. / Frontiers in artificial intelligence, 2020, vol. 3.
20. Jelodar H., Wang Y., Yuan, Ch., Xia, F. Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey, 2017.
21. Tharwat A., Gaber T., Ibrahim A., Hassanien A.E. Linear discriminant analysis: A detailed tutorial. Ai communications, 2017, vol. 30, pp. 169-190.

***Isaev Sergey Vladislavovich**. Cand. Sc., docent, institute of computational modeling SB RAS, deputy director for research, head of the department of information and telecommunication technologies. Research interests: cybersecurity and information protection, Internet technologies, distributed information systems, intelligent systems. AuthorID: 1623, SPIN: 1552-8542, si@icm.krasn.ru, 660036, Russia, Krasnoyarsk, Akademgorodok, bld. 50, p. 44, ICM SB RAS.*

***Dontsov Denis Yurievich**. Postgraduate student, institute of computational modeling SB RAS. Research interests: machine learning, Internet technologies. AuthorID: 1623, doncov.dy@icm.krasn.ru, 660036, Russia, Krasnoyarsk, Akademgorodok, bld. 50, p. 44, ICM SB RAS.*

Статья поступила в редакцию 05.05.2023; одобрена после рецензирования 07.12.2023; принята к публикации 07.12.2023.

The article was submitted 05/05/2023; approved after reviewing 12/07/2023; accepted for publication 12/07/2023.