DOI: 10.15514/ISPRAS-2023-35(5)-2

Исследование возможности идентификации веб-сайтов, посещаемых пользователем, на основе HTTP/2 трафика

[1,2,3,4] *А.И. Гетьман, ORCID: 0000-0002-6562-9008 <ever@ispras.ru>*
[1,2] *И.А.Степанов,* ORCID: 0009-0003-1964-5001 *<stepanov.ia@phystech.edu>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25.*
[2] *Московский физико-технический институт, 141700, Россия, Московская область, г. Долгопрудный, Институтский пер., 9.*
[3] *Национальный исследовательский университет «Высшая школа экономики», 101978, Россия, г. Москва, ул. Мясницкая, д. 20.*
[4] *Московский государственный университет имени М.В. Ломоносова 119991, Россия, г. Москва, Ленинские горы, д. 1.*

Аннотация. Конфиденциальность является важным свойством безопасности при обмене данными по сети. Для её реализации используется семейство протоколов SSL/TLS, которые, однако, в полной мере не скрывают ни посещаемого сайта, ни действий пользователя. Помимо конфиденциальности приватность также играет значимую роль для пользователей сети. Для обеспечения дополнительной приватности были реализованы некоторые программные решения, такие как Tor и I2P. В качестве меры приватности соответствующих решений может использоваться их устойчивость к специализированному классу атак. Одной из атак является Website Fingerprinting, позволяющая по трафику, отправляемому и получаемому известным пользователем, определять, какие именно сайты он посещал. Website Fingerprinting — это задача классификации, где объектом является посещение пользователем веб-сайта, а классом сам веб-сайт. В данной статье исследуется атака Website Fingerprinting для HTTP/2 трафика. В работе присутствует описание и вычисление популярных признаков, используемых при классификации трафика, и оценивается их применимость к задаче Website Fingerprinting. Для реализации атаки Website Fingerprinting строится несколько классификаторов, среди которых выбирается алгоритм, дающий лучший результат на собранном наборе данных. Точность лучшего классификатора составляет 97.8% в определённых допущениях. Кроме того, в работе присутствует оценка и анализ некоторых ограничений реального мира, влияющих на точность классификации.

Ключевые слова: Website Fingerprinting; HTTP/2; машинное обучение.

Для цитирования: Гетьман А.И., Степанов И.А. Исследование возможности идентификации веб-сайтов, посещаемых пользователем, на основе HTTP/2 трафика. Труды ИСП РАН, том 35, вып. 5, 2023 г., стр. 23–36. DOI: 10.15514/ISPRAS–2023–35(5)–2.

Investigation of the Possibility of Identifying Websites Visited by the User Based on Http/2 Traffic

[1,2,3,4] *A.I. Getman*, ORCID: 0000-0002-6562-9008 <*ever@ispras.ru*>
[1,2] *I.A.Stepanov*, ORCID: 0009-0003-1964-5001 <*stepanov.ia@phystech.edu*>

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences,*
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.
[2] *Moscow Institute of Physics and Technology (National Research University)*
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia.
[3] *National Research University «Higher School of Economics»*
20, Myasnitskaya ulitsa, Moscow 101978, Russia.
[4] *Lomonosov Moscow State University*
1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. Confidentiality is an important security feature when exchanging data over a network. To implement it, a family of SSL/TLS protocols is used, which, however, do not fully hide either the visited site or the user's actions. In addition to privacy, privacy also plays a significant role for network users. To provide additional privacy, some software solutions have been implemented, such as Tor and I2P. As a measure of the privacy of the relevant solutions, their resistance to a specialized class of attacks can be used. One of the attacks is Website Fingerprinting, which allows the traffic sent and received by a known user to determine which sites he visited. Website Fingerprinting is a classification task, where the object is the user's visit to the website, and the class is the website itself. This article examines the Website Fingerprinting attack for HTTP/2 traffic. The paper contains a description and calculation of popular features used in traffic classification, and assesses their applicability to the Website Fingerprinting task. To implement the Website Fingerprinting attack, several classifiers are built, among which an algorithm is selected that gives the best result on the collected dataset. The accuracy of the best classifier is 97.8% under certain assumptions. In addition, there is an assessment and analysis of some real-world constraints affecting the accuracy of classification.

Keywords: Website Fingerprinting; HTTP/2; Machine learning.

For citation: Getman A.I., Stepanov I.A. Investigation of the possibility of identifying websites visited by the user based on HTTP/2 traffic. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 5, 2023. pp. 23-36 (in Russian). DOI: 10.15514/ISPRAS-2023-35(5)-2.

1. Введение

С появлением Всемирной паутины и протокола HTTP, работающего поверх TCP/IP, встал вопрос о конфиденциальности и целостности данных TCP протокола. В 1995 году появился протокол SSL 2.0, который пытался решить эту проблему. Чуть позже в 1999 появился TLS, который и по сей день используется для обеспечения защищённой передачи между узлами во Всемирной паутине.

Однако семейство протоколов SSL/TLS не может в полной мере обеспечить конфиденциальность, так как по косвенным признакам можно примерно восстановить отправляемые запросы без доступа к их содержимому. Помимо конфиденциальности приватность также играет значимую роль для пользователей сети. Для обеспечения ещё большей приватности был создан математический аппарат mixed-сетей и луковой маршрутизации, реализованный в некоторых программных решениях. Для оценки того, насколько хорошо достигается приватность mixed-сетями и луковой маршрутизацией был разработан ряд атак, одной из которых является Website Fingerprinting. Теоретически Website Fingerprinting позволяет по перехваченному трафику, который пользователь отправляет или получает, определить, какие веб-сайты посещал пользователь. С точки зрения машинного обучения Website Fingerprinting – это задача классификации, где объектом является посещение пользователем веб-сайта, а классом сам веб-сайт.

Атакующий, отследив веб-сайты, посещаемые пользователем, может узнать некоторую конфиденциальную информацию о нём: уровень материального благополучия, состояния здоровья и так далее. С другой стороны, высокий уровень конфиденциальности может создавать проблемы провайдерам и системам безопасности, затрудняя защиту от DDoS-атак, фильтрацию небезопасных ресурсов, организацию безопасного интернета и родительского контроля. Таким образом, задача исследования уровня приватности, предоставляемого современными протоколами, и возможности идентификации посещаемых ресурсов без анализа содержимого передаваемых пакетов, является актуальной на данный момент.

Первые работы, посвящённые задачи Website Fingerprinting, использовали для классификации лишь размеры веб-объектов HTML-файла. В случае HTTP/1.1 трафика у атакующего есть больше информации, так как HTTP-запросы обрабатываются последовательно, в отличие от HTTP/2 трафика. Поэтому, в случае HTTP/1.1 трафика помимо размеров и временных меток пакетов, могут быть получены размеры веб-объектов, принадлежащих конкретной веб-странице. Однако в последнее время большинство веб-сайтов используют HTTP/2 вместо HTTP/1.1. Идентификация страниц HTTP/2 трафика является более сложной задачей, так как в этом случае происходит передача нескольких асинхронных HTTP-запросов по одному TCP-соединению. Поэтому в этом случае выделить размеры объектов уже не так просто.

Кроме того, не стоит забывать о различных ограничениях реального мира, которые могут уменьшать точность работы классификатора. В некоторых работах было показано, что время между обучением и использованием классификатора, версия браузера, наличие нескольких вкладок, открытый мир, наличие кэша могут негативно влиять на способность предсказания классификатором веб-сайтов, которые посещал пользователь.

Стоит отметить, что доля работ, посвященных атаке вида Website Fingerprinting для HTTP/2 трафика, невелика. Возникает вопрос о том, могут ли признаки и алгоритмы, используемые при решении данной задачи для других протоколов (HTTP/1.1) и других технологий (VPN, Tor), быть эффективны для HTTP/2 трафика.

Данная работа посвящена исследованию данной проблемы: эффективности атаки вида Website Fingerprinting к современному веб-трафику на основе протокола HTTP/2. Основными результатами работы являются:

- изучение эффективности различных признаков HTTP/2 трафика для атаки Website Fingerprinting
- сбор набора данных для построения классификации веб-сайтов
- построение различных классификаторов и выбор среди них оптимального
- оценка влияния различных ограничений реального мира на точность работы оптимального классификатора

В разделе 2 представлены и описаны различные термины и понятия, касающиеся задачи Website Fingerprinting. Затем в разделе 3 представлен обзор существующих решений. В разделе 4 содержится описание построения различных классификаторов и приведены результаты работы наиболее эффективных из них. Оценки ограничений реального мира описаны в разделе 5. Наконец, в разделе 6 подводятся итоги и обозначаются дальнейшие направления развития данной работы.

2. Описание задачи

В этом разделе представлены термины и понятия, использующиеся в задаче Website Fingerprinting.

2.1 Модель атаки

На рис. 1 представлена модель атаки Website Fingerprinting. Атакующий может только записывать сетевые пакеты, но при этом не может их задерживать, изменять и расшифровывать. Используя собранную информацию, атакующий может попытаться классифицировать веб-сайты, которые посещал пользователь.

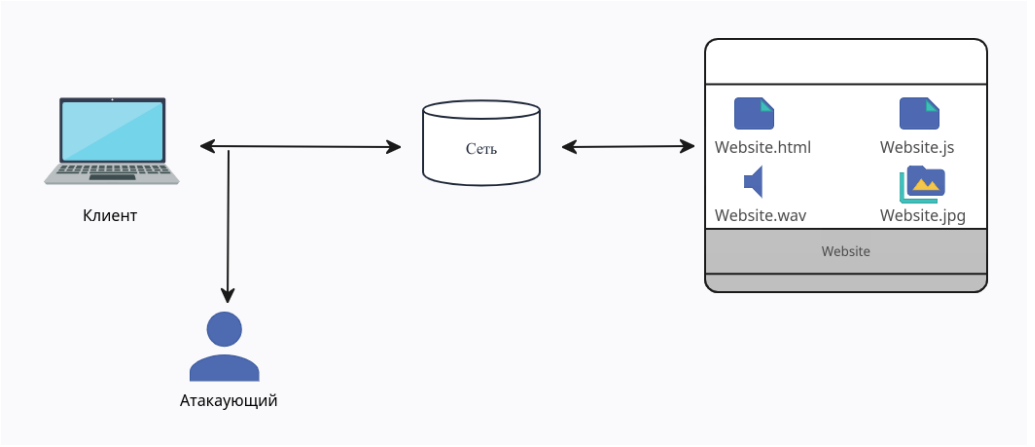

Рис. 1. Модель атаки.
Fig. 1. Attack model.

В данной работе используется термин веб-взаимодействие. Под этим термином мы понимаем набор пакетов, соответствующий посещению пользователем конкретного веб-сайта.

2.2 Метрики

В задаче Website Fingerprinting используемыми метриками являются: accuracy, recall, precision. Стоит понимать, что числовые значения данных метрик для атак Website Fingerprinting сильно зависят от набора данных, на котором обучалась и тестировалась модель машинного обучения. Поэтому, сравнивать эффективность атак Website Fingerprinting, построенных и протестированных на различных наборах данных, между собой, основываясь только на числовых значениях данных метрик, не стоит.

2.3 Сценарий открытого и закрытого мира

В сценарии закрытого мира присутствуют N классов. При обучении у атакующего есть экземпляры каждого из N классов. При тестировании классификатора предполагается, что встречаются веб-сайты только из известного набора (набора из N классов). Наиболее популярной и логичной в данном сценарии является метрика accuracy, так как чаще всего классы сбалансированы. Теперь рассмотрим сценарий открытого мира. Сначала дадим два определения:

Отслеживаемый набор – набор экземпляров веб-сайтов, о наличие которых знает атакующий при обучении и тестировании. Кроме того, при обучении атакующий знает класс, к которому принадлежит каждый экземпляр отслеживаемого набора.

Не отслеживаемый набор – набор экземпляров веб-сайтов, о наличии которых не знает атакующий при тестировании. Кроме того, при обучении атакующий не знает класс (веб-сайт), к которому принадлежит каждый экземпляр не отслеживаемого набора.

Главное отличие сценария закрытого мира от сценария открытого мира заключается в наличии не отслеживаемого набора во втором случае. При этом, классификация в данном сцена-

рии бывает двух видов. Первый: классификация трафика на отслеживаемый и не отслеживаемый набор, второй: классификация идентичная классификации в сценарии закрытого мира, но не на N классов, а на N + 1 класс, где вводится дополнительный класс, соответствующий не отслеживаемому набору.

2.4 Классификация веб-сайтов или веб-страниц

В задаче Website Fingerprinting существует некоторая разница между веб-сайтом и веб-страницей. Под веб-сайтами подразумеваются индексные или фоновые страницы веб-сайтов и классификатор настроен именно на их классификацию. Классификация же веб-страниц более сложный процесс, так как в данном случае у классификатора есть меньше информации для определения конкретной веб-страницы. Кроме того, всё усложняется при классификации веб-страниц одного и того же веб-сайта. В данном случае посещение различных веб-страниц будет иметь схожее "поведение" с точки зрения задачи Website Fingerprinting.

2.5 Отличие HTTP/1.1 от HTTP/2

Поведение протоколов HTTP/2 и HTTP/1.1 отличается друг от друга, что может приводить к тому, что методы для классификации веб-сайтов HTTP/1.1 трафика могут быть неэффективны для классификации HTTP/2 трафика. Теперь рассмотрим подробнее эти отличия.

В HTTP/2 присутствует мультиплексирование в отличие от HTTP/1.1. Мультиплексирование означает передачу нескольких асинхронных HTTP-запросов по одному TCP-соединению, в отличие от HTTP/1.1, где HTTP-запросы обрабатываются последовательно. Это свойство является принципиальным с точки зрения задачи Website Fingerprinting. В HTTP/1.1 каждый объект страницы, такой как файлы HTML, CSS, JS, JPEG и т.д., загружается последовательно в рамках отдельного TCP-соединения. В HTTP/2 же объекты загружаются параллельно, что затрудняет их выделение. Поэтому алгоритмы, основанные на коэффициенте Жаккара и размерах объектов HTML-файла, могут быть не эффективны в случае HTTP/2 трафика.

Server Push. В HTTP/2 присутствует технология Server Push, которая позволяет отправить объект клиенту, не дожидаясь запроса от него. Однако на практике эта технология используется нечасто.

Сжатие заголовков. При передаче данных часть передаваемой информации, а именно заголовки, повторяется. Для решения данной проблемы в HTTP/2 вводят сжатие заголовков с помощью алгоритма Хаффмана.

2.6 Признаки

В задаче Website Fingerprinting наиболее популярными признаками являются: направления пакетов, временные интервалы, размеры пакетов, накопленная длина пакетов. В данной работе подробно исследуется, как те или иные признаки эффективны для классификации. Представим наше веб-взаимодействие массивом пакетов F:

$$F = (p_1, \dots p_n)$$

Направления пакетов. Пусть веб-взаимодействие представлено массивом пакетов F, где $p_i = 1$ представляет пакет нисходящей линии связи (от сервера), а $p_i = -1$ представляет пакет восходящей линии связи (к серверу). Тогда массив F будет массивом признаков направления пакетов.

Размеры пакетов. Пусть веб-взаимодействие представлено массивом пакетов F, где p_i – это размер пакета с учётом направления. Тогда массив F будет массивом признаков размеров пакетов.

Временные интервалы. Пусть веб-взаимодействие представлено массивом пакетов F. Пусть t_i - время отправки пакета с номером i. Введём временной интервал между двумя пакетами:

$$d_i = t_{i+1} - t_i$$

Тогда массивом признаков временных интервалов будет следующий массив:

$$D = (d_1, \dots d_{n-1})$$

Однако признаки, связанные с временными характеристиками, могут быть сильно связаны с некоторыми свойствами сети: скорость интернета, браузер и так далее, что может сильно мешать при классификации веб-сайтов.

Сумма байт до первого встречного пакета. Пусть веб-взаимодействие представлено массивом пакетов F, где p_i – это размер пакета с учётом направления. Введём a_i – сумма подряд идущих элементов одного знака массива F до первого элемента с другим знаком. Тогда массивом данных признаков будет следующий массив:

$$A(F) = (a_1, \dots a_k)$$

Накопленная длина пакетов. Пусть веб-взаимодействие представлено массивом пакетов F, где $p_i > 0$, если пакет идёт от сервера и $p_i = 0$, если к серверу. Пусть A:

$$A(F) = (a_1, \dots a_n) : a_1 = p_1, a_i = p_i + p_{i-1}$$

Тогда массив A будет накопленной суммой пакетов.

2.7 Коэффициент Жаккара.

Пусть есть два множества A и B. Коэффициентом Жаккара для множеств A и B называется отношение:

$$J = \frac{c}{a+b-c}, \text{ где}$$

- c – количество элементов, общих для множества A и B
- a и b – количество элементов множества A и B соответственно

Данный коэффициент для алгоритма классификации был популярен в ранних работах, посвящённых Website Fingerprinting. Однако, как отмечалось ранее, с появлением новых протоколов «выделение размеров объектов» HTML страницы веб-сайта стало более проблематично, что заметно уменьшило популярность коэффициента Жаккара.

3. Обзор литературы

В этом разделе рассмотрены наиболее популярные работы и методы, посвящённые задаче Website Fingerprinting. В конце раздела представлена сравнительная таблица различных работ по данной теме.

3.1 Раннее развитие WF

Первые Website Fingerprinting атаки пытались определять URL-адреса, который пользователь посещает через зашифрованные SSL-соединения. Так в 1998 [1] была предложена атака, которая на основе объёма передаваемых данных (число байт) определяла веб-сайт, так как в случае HTTP/1.0 размер HTML-файла веб-сайта был главным признаком для классификации данного веб-сайта.

Так как в случае HTTP/1.1 трафика файлы передаются по TCP-соединению последовательно, существует возможность определить размеры отдельных объектов веб-сайта. Используя это в [2], авторы предложили классифицировать веб-сайты не по общему числу переданных байт, а по числу и размерам отдельных объектов веб-сайта. Используя эти же признаки и коэффициент Жаккара, авторы в [3] показали, что Website Fingerprinting атака может достигать приемлемой точности для значительного числа веб-сайтов в определённых допущениях.

Позже появились работы для классификации не только SSL трафика, но и VPN трафика. Так в [4] Liberatore и др. классифицируют веб-сайты VPN трафика, используя только размеры пакетов, интервалы между пакетами и коэффициент Жаккара.

3.2 WF на классических алгоритмах ML

С развитием ML алгоритмов постепенно стали появляться статьи, посвящённые Website Fingerprinting атакам на основе алгоритмов машинного обучения. Так в [5] D.Herrmann и др., используя в качестве признаков частоты распределений размеров ip-пакетов, а в качестве ML алгоритма алгоритм Наивного байеса, достигают точности в 97% для 775 различных сайтов. Помимо SSL и VPN трафика, начиная с определённого момента, стали появляться статьи, посвящённые Tor трафику. К примеру, в [6] авторы, используя направления, размеры и временные интервалы трафика, а в качестве алгоритма SVM, достигли точности 55% для закрытого мира. Кроме того, в данной работе проведено исследование сценария открытого мира.

В последствии появились работы [7], посвящённые не только построению атак, но и изучению влияния различных факторов, таких как сценарий открытого или закрытого мира, время с момента сбора трафика, наличие нескольких вкладок у пользователя, на точность атаки. В [8] авторы предложили алгоритм, основанный на расстоянии Махаланобиса и классифицирующий трафик с двумя вкладками. Точность классификации первой страницы составила 75,9%, точность классификации второй - 40,5%. При этом, интервал задержки между двумя страницами составил 2 секунды.

В 2016 A. Panchenko и др. [9], используя в качестве классификатора SVM, исследовали задачу Website Fingerprinting для открытого и закрытого мира и показали, что с увеличением числа экземпляров открытого мира точность классификатора значительно падает. При этом, в случае закрытого мира точность составила более 90% процентов. Алгоритм классификации с данным набором признаков, основанных на числе и размере ячеек в сети Tor, получил в литературе название CUMUL.

Одними из первых кто показал, что в качестве признаков для классификации могут выступать лишь направления пакетов были Avdoshin S. и др. [10]. В своей работе авторы использовали алгоритм на основе SVM, а исследуемая выборка состояла из 7 веб-сайтов. При этом, точность составляла чуть больше 70%.

С другой стороны, существуют работы, посвящённые анализу трафика с использованием информации только о временных метках. К примеру, в [11], используя в качестве алгоритма метод ближайших соседей, авторы классифицируют трафик без знания о начале и конце исследуемого веб-взаимодействия.

3.3 Эпоха глубокого обучения

С развитием глубокого обучения стало появляться большое число статей, посвящённых классификации веб-сайтов с помощью нейронных сетей. В [12] авторы, классифицируют веб-сайты, используя в качестве алгоритма сверточную нейронную сеть. Работа характерна детальным подбором оптимальных гиперпараметров сверточной нейронной сети, большим числом экземпляров набора данных, исследованием как открытого, так и закрытого мира. Стоит отметить, что в качестве признаков алгоритма выступали лишь направления пакетов (ячеек Tor) трафика. Rimmer и др. [13] в качестве алгоритма классификации использовали не только сверточную нейронную сеть, но и автоэнкодер и сеть долгой краткосрочной памяти.

Полученные результаты в различных работах говорят о том, что для классификации веб-сайтов, содержащих быстро меняющийся контент, наиболее подходят признаки, основанные лишь на размерах, направлениях и временных интервалах пакетов, в то время как признаки, созданные вручную (средний размер пакета, максимальный размер пакета и др.), могут быть неэффективны.

Работы, посвящённые атаке Website Fingerprinting для HTTP/2 трафика, стали появляться лишь в последнее время. Так в [14] авторы используют в качестве признаков накопленную длину пакетов, а в качестве алгоритма метод ближайших соседей.

В 2020 году появилась работа [15] классификации веб-сайтов на основе анализа HTTP/2 трафика. При этом, алгоритм основан на уже забытом коэффициенте Жаккара, который достаточно долго не использовался для оценки близости. Хотя точность классификации не высока относительно других алгоритмов 62,21%, однако это можно объяснить огромным числом классов 55 212. Главным достоинством работы является классификация именно HTTP/2 трафика, так как во многих других работах он почти не встречается.

В табл. 1 приведено сравнение по признакам и алгоритмам различных работ, которые упоминались выше.

Число работ, в которых встречается классификации веб-сайтов HTTP/2 трафика невелико, а количество тех, где классифицируется только HTTP/2 трафик (без HTTP/1.1) ещё меньше. Однако в последнее время доля HTTP/2 трафика заметно выросла. В данной работе исследуется атака Website Fingerprinting имено для HTTP/2 трафика.

Табл. 1. Сравнительная таблица работ
Table 1. Comparative table of works

№	Год	Трафик	Признаки	Алгоритм
1	1998	HTTP/1.0	размер html-файла	-
2	2002	HTTP/1.1	размер объектов веб-страницы	коэф. Жаккара
3	2002	HTTP/1.1	размер объектов веб-страницы	коэф. Жаккара
4	2006	VPN	размеры пакетов и временные интервалы	коэф. Жаккара
5	2009	HTTP(s)/1.1	размеры пакетов	наив. Байес
6	2011	Tor	размеры пакетов	SVM
7	2014	Tor	размеры пакетов	SVM решающее дерево наив. Байес
8	2015	HTTP(s)/1.1	размеры пакетов и временные интервалы	SVM
9	2016	HTTP(s)/1.1	размеры пакетов	SVM
10	2016	Tor	направления пакетов	SVM
11	2016	HTTP(s)/1.1 Tor	временные интервалы	другой алгоритм
12	2017	Tor	направления пакетов	сверточная нейронная сеть автоэнкодер LSTM
13	2018	Tor	направления пакетов	сверточная нейронная сеть
14	2019	HTTP(s)/2	накопленная сумма пакетов	K-NN
15	2020	HTTP(s)/2	размеры пакетов	коэф. Жаккара

4. Построение классификатора

В этом разделе будет описан процесс сбора данных и построения оптимального классификатора, то есть выбор признаков, алгоритма и гиперпараметров алгоритма, дающих лучший результат на собранном наборе данных.

Отметим, что в данной работе решается задача закрытого мира для классификации веб-сайтов HTTP/2 трафика. Кроме того, предполагается, что клиент посещает в выбранный момент времени лишь одну веб-страницу, и что нам известны время начала и конца конкретного веб-взаимодействия. Также предполагается, что клиент использует ту же версию браузера, что и атакующий. Эти допущения встречаются во многих работах и заметно упрощают задачу, но

при этом делают её менее реалистичной. В разделе 5 рассмотрены эти допущения и то, как они влияют на точность классификации.

4.1 Сбор набора данных

Для снятия трафика и сбора данных использовался Wireshark для снятия трасс и Selenium для автоматизации процесса. Сбор набора данных осуществлялся следующим образом:

Шаг 1. Открытие браузера.
Шаг 2. Открытие сайта из списка сайтов автоматическим способом.
Шаг 3. Ожидание загрузки веб-сайта.
Шаг 4. Закрытие веб-сайта, сохранение времени начала и конца сессии.
Шаг 5. Повторение 2-4 шага N раз.
Шаг 6. Фильтрация трафика, удаление пакетов, не относящихся к исследуемым веб-сайтам.
Шаг 7. Сохранение отфильтрованных pcap-файлов.
Шаг 8. Разбиение pcap-файлов на веб-взаимодействия, соответствующие посещению пользователем (клиентом) индексной страницы веб-сайта, с помощью информации о начале и конце веб-взаимодействия.

В конечном итоге было собрано 9405 веб-взаимодействий для 15 популярных веб-сайтов. При этом число экземпляров веб-взаимодействий для каждого веб-сайта одинаково. Таким образом, можно говорить о сбалансированности нашей выборки относительно классов.

4.2 Вычисление признаков

Всего было исследовано 5 различных признаков для обучения: направления пакетов, размеры пакетов, временные интервалы, сумма байт до первого встречного пакета, накопленная длина пакетов. На рис. 2 представлена схема вычисления признаков.

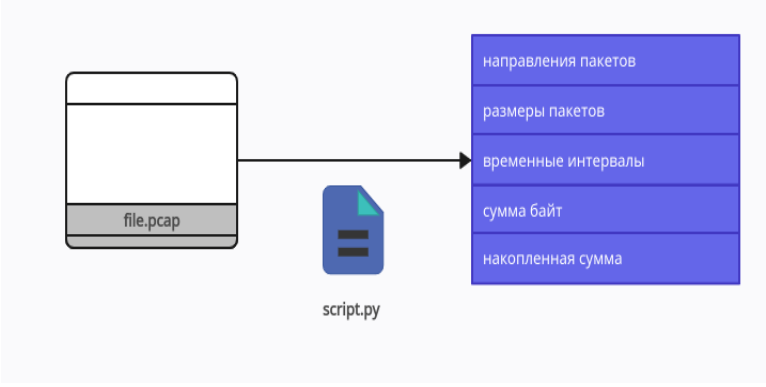

Рис. 2. Вычисление признаков.
Fig. 2. Calculating features.

4.3 Построение оптимального классификатора

Исследовались следующие алгоритмы:

- Решающее дерево
- Случайный лес
- Бустинг
- Коэффициент Жаккара

4.3.1 Решающее дерево

В качестве решающего дерева использовалось решающее дерево из scikit-learn. Точность рассчитывалась с помощью кросс-валидации с разбиением на 10 фолдов. При этом, исследовались следующие гиперпараметры алгоритма: глубина решающего дерева (depth).

На рис. 3 представлена точность работы алгоритма от глубины дерева для различных признаков.

Результаты показывают, что лучший результат при использовании решающего дерева даёт признак накопленная сумма.

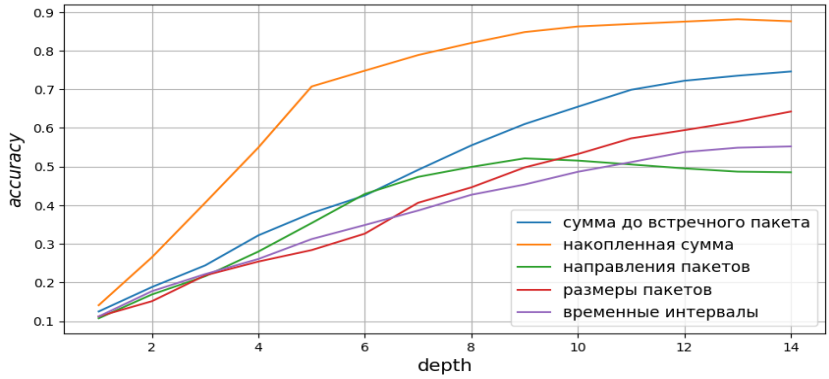

Рис. 3. Точность решающего дерева.
Fig. 3. The accuracy of the decision tree.

4.3.2 Случайный лес

В качестве случайного леса использовался случайный лес из scikit-learn. Точность рассчитывалась с помощью кросс-валидации с разбиением на 10 фолдов. При этом, исследовались следующие гиперпараметры алгоритма: число деревьев (n_estimators). На рис. 4 представлена точность работы алгоритма от числа деревьев для различных признаков.

График показывает, что лучший результат при использовании случайного леса даёт признак накопленная сумма.

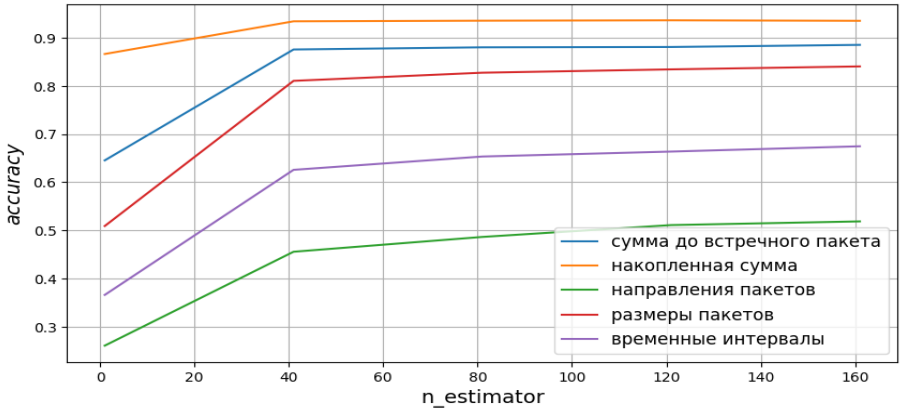

Рис. 4. Точность случайного леса.
Fig. 4. Random Forest Accuracy.

4.3.3 Бустинг

В качестве бустинга использовался CatBoostClassifier из catboost. Точность рассчитывалась с помощью кросс-валидации с разбиением на 10 фолдов. При этом, исследовались следующие

гиперпараметры алгоритма: число деревьев (iterations), глубина дерева (depth). На рис. 5 представлена точность работы алгоритма от числа деревьев для различных признаков.

График показывает, что лучший результат при использовании бустинга даёт признак накопленная сумма и при достижении 200 деревьев точность увеличивается незначительно. На рис. 6 представлена точность работы алгоритма от глубины деревьев для различных признаков.

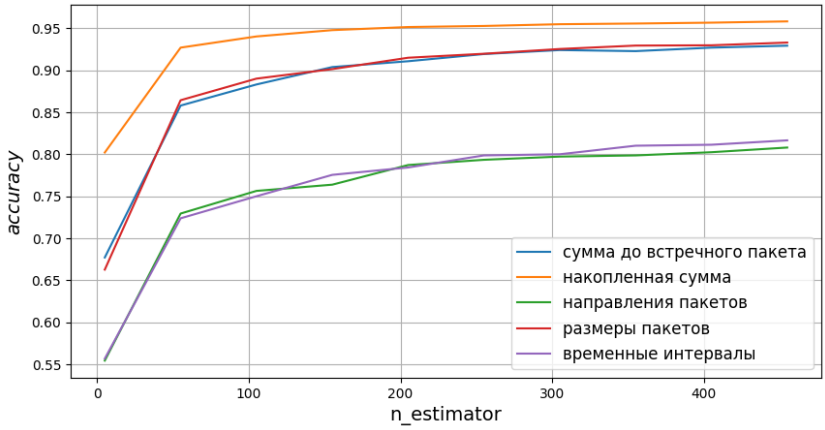

Рис. 5. Точность бустинга (число деревьев).
Fig. 5. Boost Accuracy.

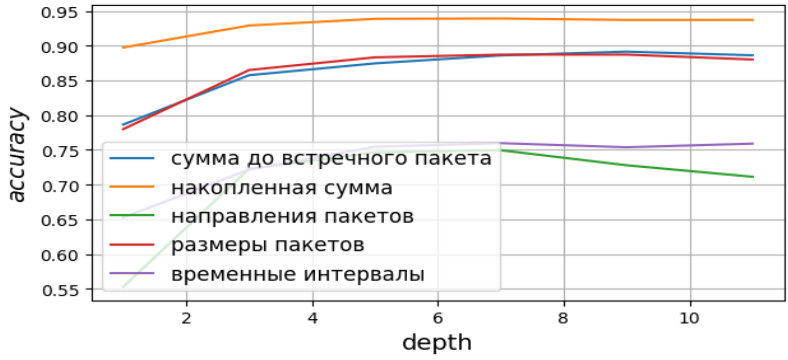

Рис. 6. Точность бустинга (глубина деревьев).
Fig. 6. Boost Accuracy.

4.3.4 Коэффициент Жаккара

Алгоритм, основанный на коэффициенте Жаккара для классификации веб-сайтов, выглядит следующим образом:

Шаг 1. Пусть дано n известных веб-сайтов (1 ... n) для каждого из которых найдены размеры его объектов.

Шаг 2. Пусть дан неизвестный веб-сайт - x, который нужно соотнести с одним из известных, и даны размеры его объектов.

Шаг 3. Рассчитываем коэффициент Жаккара для следующих пар *(1;x) ... (n;x).*

Шаг 4. Максимальное значение коэффициента Жаккара будет указывать на нужный веб-сайт.

Однако стоит понимать, что точность может сильно зависеть от того, какие веб-сайты будут установлены в качестве известных. Поэтому нужно попробовать в качестве известных все комбинации веб-сайтов и выбрать те, что дают лучший результат.

4.4 Итоги

Результаты показывают, что накопленная сумма пакетов даёт лучший результат для всех алгоритмов. В табл. 2 представлены результаты работ различных алгоритмов на собранном наборе данных. Таким образом, лучший результат на собранном наборе данных составил 97.8%.

Табл. 2. Результаты работы алгоритмов
Table 2. Results of the algorithms

№	Алгоритм	Оптимальные гиперпараметры	Лучший признак	Точность
1	Решающее дерево	max_depth = 11	Накопленная сумма	0.891
2	Случайный лес	n_estimators = 40 max_depth = 11	Накопленная сумма	0.925
3	CatBoostClassifier	iterations = 400 depth = 5 learning_rate = 0.2	Накопленная сумма	0.978
4	Коэффициент Жаккара	-	Накопленная сумма	0.907

5. Оценка влияния ограничений реального мира на точность классификации

В данном разделе будут рассмотрены различные ограничения реального мира и оценено их влияние на точность классификации.

5.1 Браузер

Оценим, как версия браузера пользователя влияет на точность классификации. Для этого обучим классификатор CatBoostClassifier на трафике, собранном с помощью браузера1. Протестируем же классификатор на трафике, собранном с помощью браузера2, который принципиально отличается от браузера1. Таким образом моделируется ситуация, когда браузеры атакующего и клиента не совпадают. В табл. 3 представлена точность работы алгоритма, обученном на наборе данных браузера1. Результаты показывают, что алгоритм имеет сильную зависимость от браузера клиента.

Табл. 3. Влияние браузера на точность классификации
Table 3. Browser influence on classification accuracy

	Тестовый набор данных	Точность
1	Браузер 1	0.975
2	Браузер 2	0.340

5.2 Классификация нескольких вкладок

Как уже было сказано, в данной работе предполагалось, что в определённый момент времени клиент имеет только одну открытую вкладку. Однако это предположение заметно отклоняет задачу от реальной. Смоделируем ситуацию, когда у клиента открыто несколько вкладок. Для этого соберём наборы данных, в которых у пользователя фоном открыты одна, две и три вкладки соответственно. Затем протестируем данные наборы данных на классификаторе, построенном в разделе 4. Результаты представлены в табл. 4.

Результаты показывают, что метод имеет сильную зависимость от числа вкладок, открытых у пользователя.

Табл. 4. Влияние нескольких вкладок на точность классификации
Table 4. The effect of multiple tabs on classification accuracy

	Тестовый набор данных (число вкладок)	Точность
1	Одна вкладка	0.975
2	Одна вкладка и одна фоновая	0.480
3	Одна вкладка и две фоновые	0.330
4	Одна вкладка и три фоновые	0.275

6. Заключение

В данной работе представлено исследование классификации веб-сайтов HTTP/2 трафика, посещаемых пользователем. В предположениях, сформулированных выше, было построено несколько классификаторов на популярных алгоритмах машинного обучения. Полученные результаты говорят о теоретической возможности классифицировать веб-сайты, посещаемые пользователем.

Однако стоит понимать, что в условиях реального мира данная классификация может иметь очень невысокую предсказательную способность, что было показано в разделе 5.

В будущей работе планируется значительно увеличить набор данных для классификации, а также более детально исследовать влияние ограничений реального мира на точность работы классификации.

Кроме того, планируется исследование возможности классификации веб-сайтов без знания о начале и конце веб-взаимодействия, что как было сказано ранее, является сильным допущением.

Список литературы / References

[1]. Mistry S. Traffic Analysis of SSL-Encrypted Web Browsing //http://bmrc. berkeley. edu/people/shailen/Classes/SecurityFall98/paper. ps. – 1998.
[2]. Hintz A. Fingerprinting websites using traffic analysis //International workshop on privacy enhancing technologies. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. – С. 171-178.
[3]. Sun, Q., Simon, D. R., Wang, Y. M., Russell, W., Padmanabhan, V. N., & Qiu, L. (2002, May). Statistical identification of encrypted web browsing traffic. In Proceedings 2002 IEEE Symposium on Security and Privacy (pp. 19-30). IEEE.
[4]. Liberatore, M., & Levine, B. N. (2006, October). Inferring the source of encrypted HTTP connections. In Proceedings of the 13th ACM conference on Computer and communications security (pp. 255-263).
[5]. Herrmann, D., Wendolsky, R., & Federrath, H. (2009, November). Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. In Proceedings of the 2009 ACM workshop on Cloud computing security (pp. 31-42).
[6]. Panchenko, A., Niessen, L., Zinnen, A., & Engel, T. (2011, October). Website fingerprinting in onion routing based anonymization networks. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society (pp. 103-114).
[7]. Juarez, M., Afroz, S., Acar, G., Diaz, C., & Greenstadt, R. (2014, November). A critical evaluation of website fingerprinting attacks. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (pp. 263-274).
[8]. Gu, X., Yang, M., & Luo, J. (2015, May). A novel website fingerprinting attack against multi-tab browsing behavior. In 2015 IEEE 19th international conference on computer supported cooperative work in design (CSCWD) (pp. 234-239). IEEE.
[9]. Panchenko, A., Lanze, F., Pennekamp, J., Engel, T., Zinnen, A., Henze, M., & Wehrle, K. (2016, February). Website Fingerprinting at Internet Scale. In NDSS.
[10]. Avdoshin, S. M., & Lazarenko, A. V. (2016). Deep web users deanonimization system. Труды Института системного программирования РАН, 28(3), 21-34.
[11]. Feghhi, S., & Leith, D. J. (2016). A web traffic analysis attack using only timing information. IEEE Transactions on Information Forensics and Security, 11(8), 1747-1759.
[12]. Sirinam, P., Imani, M., Juarez, M., & Wright, M. (2018, October). Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 1928-1943).

[13]. Rimmer, V., Preuveneers, D., Juarez, M., Van Goethem, T., & Joosen, W. (2017). Automated website fingerprinting through deep learning. arXiv preprint arXiv:1708.06376.

[14]. Shen, M., Liu, Y., Chen, S., Zhu, L., & Zhang, Y. (2019, May). Webpage fingerprinting using only packet length information. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

[15]. Ghiëtte, V., & Doerr, C. (2020, June). Scaling website fingerprinting. In 2020 IFIP Networking Conference (Networking) (pp. 199-207). IEEE.

Информация об авторах / Information about authors

Александр Игоревич ГЕТЬМАН – кандидат физико-математических наук, старший научный сотрудник ИСП РАН, доцент ВШЭ. Сфера научных интересов: анализ бинарного кода, восстановление форматов данных, анализ и классификация сетевого трафика.

Aleksandr Igorevich GETMAN – PhD in physical and mathematical sciences, senior researcher at ISP RAS, associate professor at HSE. Research interests: binary code analysis, data format recovery, network traffic analysis and classification.

Иван Александрович СТЕПАНОВ – студент МФТИ. Сфера научных интересов: анализ сетевого трафика, машинное обучение.

Ivan Alexandrovich STEPANOV is a student at MIPT. Research interests: network traffic analysis, machine learning.