УДК 004.042
DOI 10.21661/r-554668

О.И. Филимонов

МЕТОДЫ И ИНСТРУМЕНТЫ АНАЛИЗА ТРАФИКА НА ВЕБ-СЕРВЕРЕ

Аннотация: *в статье дан обзор методов и инструментов анализа вэб-трафика, формирующий представление о различных подходах аналитиков к изучению информации, предоставляемой трафиком. Эта информация позволяет выявить не только характеристики трафика (такие как его объем и интенсивность, сроки предоставления услуги и др.), но и широкий спектр потребительских и даже психологических свойств пользователей сети (частота посещения сайта, интересы, коммуникабельность, платежеспособность и пр.). Такая информация важна для маркетологов, менеджеров и оценщиков бизнеса. Она необходима для улучшения структуры сайта, разработки стратегии развития виртуального бизнеса и повышения его эффективности. Исследование проводится с использованием сравнительного анализа, результаты которого позволяют вэб-аналитику в каждом конкретном случае, зная преимущества и недостатки каждого метода и его инструментов, использовать именно тот метод (или совокупность методов), который даст наилучший результат для изучения особенностей функционирования различных видов виртуального бизнеса, в частности, виртуальных бирж. Проблема использования инструментов анализа веб-трафика становится более сложной, когда объем веб-трафика огромен, что делает задачу его анализа весьма важной и актуальной.*

Ключевые слова*: вэб-сервер, интернет-трафик, журнал трафика, методы и инструменты анализа трафика.*

Введение

Существуют различные модели и инструменты для обнаружения и анализа полезных знаний из доступных данных трафика. Они предлагают аналитикам большие возможности для составления отчетов о результатах анализа. Этот анализ используется в первую очередь для получения общего представления о том,

что произошло на веб-сайте. Веб-мастерам и системным администраторам часто нужна информация, обычно для целей управления веб-сайтом: сколько трафика они получили, на сколько запросов не удалось ответить, и какие ошибки были при этом сгенерированы.

На эту тему проводился ряд исследований [10, с. 171–193; 18, с. 900–907; 22], в результате которых изучены некоторые из таких инструментов. Каждый инструмент предлагал ту или иную функцию, которая реализуется им лучше остальных. Большинство доступных в настоящее время инструментов анализа трафика веб-сервера предоставляют и статистическую информацию.

Далее предлагается сравнительное исследование различных методов анализа и инструментов измерения трафика на веб-сервере (подробный анализ сущности понятия «трафик», его особенностей, а также вопрос формирования классификаций видов трафика проводится в статье автора [1]).

А. Анализ файла журнала веб-сервера HTTP. Технология анализа журналов веб-сервера HTTP – самая старая технология, разработанная в 1995 году [4]. Дадим ее описание.

При посещении веб-сайта пользователь подключается к веб-серверу, который обслуживает файлы, запрашиваемые пользователем. Веб-сервер непрерывно записывает свой HTTP-трафик (все HTTP-запросы и ответы) в журнал, который создает текстовый файл (одна запись для каждого запроса файла) для записи конкретного действия, причем каждый текстовый файл состоит из файла веб-журнала для этого веб-сервера. Информация в файле журнала обычно зависит от того, какой веб-сервер их генерирует. Как правило, разные веб-серверы поддерживают разные форматы журналов.

В журналах веб-сервера хранятся записи о *потоках кликов*, которые могут быть полезны для целей анализа веб-трафика [17, с. 17–21]. Это файлы в виде простого текста, которые содержат информацию об *имени пользователя, его IP-адресе, отметке времени, запросе доступа, количестве переданных байтов, URL-адресе,* на который ссылается пользователь, а также о *кодах ошибок* (если они есть) и т. д.

Данные веб-сервера – это фактически *журналы пользователей,* которые позволяют аналитику отслеживать и анализировать поведение пользователей, посещающих веб-сайт.

Файлы веб-журнала сервера предоставляют ценную информацию об использовании веб-сайта. Основываясь на этих данных, можно выявить [4]:

– количество сделанных запросов («обращений»);

– сколько всего файлов успешно обслужено;

– количество запросов по типу файлов (например, просмотров HTML-страниц);

– обслуживаемые IP-адреса и количество сделанных запросов;

– количество запросов по кодам состояния HTTP (успешные, неудачные, перенаправленные, информационные);

– ссылки на страницы данного сайта;

– браузеры и версии, отправившие запросы.

Из файлов веб-журнала можно получить и некоторые данные расширенного анализа [4]. Этот анализ может ответить на следующие вопросы:

1) кто посетил данный сайт? данные сеанса помогут определить, возвращается ли уникальный посетитель на данный сайт или нет;

– какие группы пользователей получают доступ к веб-сайту?

– какие ресурсы просматриваются чаще всего?

2) по какому пути (ссылкам) посетители переходят на сайт? зная каждую страницу, которую просматривал посетитель, и порядок просмотра страниц, можно определить тенденции перемещения посетителей по страницам, и какой элемент html (ссылка, значок и изображение) посетитель нажимал на каждой странице, чтобы перейти на следующую страницу;

3) сколько времени посетитель провел на каждой странице? по продолжительности пребывания большинства посетителей на странице можно определить, интересна она или нет;

4) где посетитель покинул сайт? последняя страница указывает на место завершения посещения. это может быть страница, которая не понравилась посетителю, – поэтому он захотел покинуть сайт;

5) какова была степень успешности взаимодействия пользователей с сайтом? чтобы выяснить это, можно посмотреть, сколько покупок было совершено и сколько загрузок было завершено.

Эта информация о трафике собирается и отправляется сборщику трафика, который затем пересылает ее анализатору. Лог-анализатор веб-сервера создает *отчеты со сводной статистикой о веб-сайте*. Он использует информацию из файлов журналов сервера, тем самым помогает обрабатывать большой объем сведений, касающихся трафика на веб-серверах.

Традиционно существует *четыре типа журналов сервера*:

– журнал передачи,

– журнал агента,

– журнал ошибок и

– журнал реферера (реферер – это строка заголовка HTTP-запроса, в которой содержится информация о странице, с которой клиент перешел на данный сайт. Анализируя, откуда идет трафик, владелец сайта получает представление о том, откуда на его сайт перешло больше пользователей).

Первые два считаются стандартными, а два других журнала используются для анализа не всегда [14].

Каждая запись журнала регистрирует переход от одной страницы к другой, сохраняя IP-адрес пользователя и всю связанную с этим информацию [5].

Если *журнал* используется правильно, это может быть очень полезно для изучения процесса превращения *посетителей* веб-сайтов в *клиентов* виртуального бизнеса. Он помогает аналитику определить *схему навигации пользователя,* то есть. какие страницы часто посещает пользователь, какие ошибки возникают у пользователя и т. д.

Процесс *анализа производительности* сервера начинается со сбора файлов журналов, охватывающих определенный период времени, поскольку чтобы

понять тенденции трафика, необходимо изучить журналы за определенный период. *Инструменты* веб-анализатора журналов являются частью программного обеспечения веб-аналитики. Они воспринимают файл журнала доступа к веб-сайту в качестве входных данных, анализируют его и генерируют отчеты в качестве результата анализа. Эти инструменты предоставляют аналитику всевозможную информацию, начиная от количества посещений сайта и заканчивая количеством посетителей, заходящих на сайт, и браузерами, которые они используют, а также продолжительностью их пребывания на сайте и др. Действительно, файлы журнала могут быть очень большими, и в них подробно указывается, какие файлы были запрошены из сети. Каждый раз, когда страница, изображение, фильм или любой другой файл загружается с веб-сервера пользователя, дата/время и IP-адрес отправителя запроса регистрируются в файле журнала веб-сервера.

Примечание: речь идет о таких инструментах как [11, с. 29–33; 21]:

1) Google Analytics – это бесплатная утилита, предоставляемая Google, которая помогает отслеживать уникальных посетителей. Она также помогает определить, какие предлагаются лучшие маркетинговые пакеты. Для использования этого инструмента установка не требуется, требуется только учетная запись Google. Средство создания отчетов по электронной почте доступно в Googleanalytics;

2) AWStats: доступно бесплатно. Этот инструмент работает как сценарий CGI или из командной строки. Он отображает всевозможную информацию, содержащуюся в журнале;

3) WebLogExpert – это еще один инструмент для анализа журналов доступа к сети, который обеспечивает их тщательный анализ. Он предоставляет аналитику конкретную и точную информацию о пользовательской статистике, но требуется создание профиля;

4) Аналог – это простой в использовании и устанавливаемый свободно доступный инструмент анализа журналов. Он очень быстрый, хорошо масштабируемый, работает с любой операционной системой и прост в установке.

Следует отметить, что первоначальная цель файлов журнала – производить *статистику производительности* на уровне сайта. Сначала они использовались IT-администраторами как способ обеспечить адекватную полосу пропускания и емкость сервера на веб-сайтах своих организаций. Однако в последние годы компании электронной коммерции значительно развили этот подход, собирая файлы журналов для получения подробной информации о профилях посетителей и их покупательной деятельности и узнавая об удобстве использования своих веб-сайтов (то есть о том, насколько успешно они достигают своих целей, и насколько посетители удовлетворены) [22].

Преимущества анализа файлов веб-журнала. Веб-сервер естественным образом фиксирует фактические данные об использовании сайта в своей рабочей среде. Файл веб-журнала отражает действия большинства посетителей данного сайта в течение потенциально длительного периода времени. Поэтому для инженеров веб-дизайна чрезвычайно важно проводить оценку удобства использования сайта.

Проблемы с анализом файла веб-журнала. Инструменты анализа веб-сайта обычно импортируют данные файла журнала во встроенную базу данных, которая, в свою очередь, преобразует данные в удобочитаемые графические отчеты для различных целей. При этом файлы журналов веб-сервера содержат много информации [20]. Так, файл журнала размером в один мегабайт (1 МБ) обычно содержит около 4000–5000 запросов страниц. Тем самым, процесс передачи и обработки информации сложен и требует много времени, поскольку некоторые данные журнала являются сокращенными и непонятными, а значит, в этом случае они не могут предоставить точную и эффективную информацию об удобстве использования данного сайта.

Отмечают еще две проблемы [21], которые затрудняют анализ файла журнала как индикатора использования. Так:

– иногда отслеживаемых данных в файле журнала *недостаточно*, что может быть вызвано отсутствием определенных типов данных об использовании сайта, которые должны были бы регистрироваться.

– в файл журнала могут попасть *посторонние данные*.

Эти проблемы приводят к недостаточной, необоснованной и вводящей в заблуждение информации, что не позволяет отвечать на большинство вопросов, сформулированных выше, необходимых для расширенного анализа.

Отметим также ряд технических *недостатков* применения метода анализа файла веб-журнала, которые состоят в следующем:

– файл журнала не может отслеживать ID (идентификатор) посетителя, даже если на сайте есть страница входа, что затрудняет идентификацию уникального посетителя несмотря на то, что данные сеанса и IP-адреса доступны. Это происходит потому, что многие из посетителей используют динамические IP-адреса, предлагаемые их поставщиками интернет-услуг;

– в случае, когда владелец сайта или поставщик интернет-услуг имеет центральный кэш для веб-страниц или когда посетитель использует кэш браузера (кэш – это блок и способ хранение данных на сервере для их повторного использования. При заходе пользователя на сайт, кэширование собирает все данные веб-страницы, превращает их в файл HTML и открывает их в браузере. В следующий раз, когда пользователь откроет этот же сайт, кэш загрузит сохранённую копию. Благодаря этому, сервер работает быстрее и не перегружается, а пользователь получает доступ к содержимому сайта без задержки). (Страницы, полученные посетителем путем нажатия кнопки «Назад» во время последующих посещений в том же сеансе), – цепочка этих кэшей не может быть зарегистрирована (или записана) в файл журнала. Эта проблема приводит к неправильному анализу пути. Кроме того, кэширование страницы может дать неверную информацию о том, куда ушел посетитель. Например, если последняя страница была из кэша, файл журнала не может ее отразить;

– веб-сервер ничего не записывает в файл журнала, когда посетитель перешел на страницу, набрав URL-адрес в поле адреса, используя закладку или перейдя по ссылке электронной почты [9]. Эта проблема приводит к неверной информации в реферерах.

– в файл журнала записывается только время, когда передача данных была инициирована, а не время, когда передача данных завершилась. Следовательно, время, проведенное на странице этим посетителем, можно приблизительно оценить только на основе сравнения временных меток текущего запроса и следующего запроса [7; 15].

Процесс анализа также усложняется из-за таких характеристик файлов веб-журнала, как большой размер и разные форматы для разных серверов. Тем не менее, *анализ веб-журналов – это наиболее часто и широко используемый подход для получения информации о посетителях в Интернете* на основе данных из журналов веб-сервера.

Б. Конфигурация CiscoNetflow – это протокол мониторинга трафика, разработанный компанией Cisco для захвата и анализа трафика, проходящего через устройства Cisco. Этот инструмент мониторинга трафика используется также для измерения скорости сетевого трафика и может быть полезен в целях безопасности *Netflow*. При этом протокол работает не только с устройствами Cisco, но и с сетевыми устройствами других производителей.

Netflow – один из важнейших компонентов программного обеспечения, адаптируемых каждой компанией *для мониторинга своего сетевого трафика*. Это программное обеспечение проделывает важную работу по выявлению и устранению *DDoS-атак* (DDoS-атака отправляет на атакуемый веб-ресурс (сайты) большое количество запросов с целью превысить способность сайта обрабатывать их все, и тем самым вызвать отказ в его обслуживании. Чаще всего, это мера экономического давления, недобросовестная конкуренция, шантаж, так как сайт, на который осуществлялась атака, не сможет работать или будет работать крайне медленно) (показывая различия между нормальным сценарием и сценарием атаки).

Netflow регистрирует подробную информацию о пакетах, проходящих через маршрутизатор, в файле журнала, к которому можно получить доступ из удаленной системы для анализа. Он отслеживает трафик в соответствии с различными метриками, такими как *IP-адрес источника, адрес назначения, тип службы,*

порт источника, порт назначения и *т. д.* Эта информация о трафике собирается и отправляется сборщику, который затем пересылает ее анализатору.

Отметим, что существует также аналогичное программное обеспечение – *Internet Protocol FlowInformation Export (IPFIX),* созданное группой IETF для той же цели мониторинга сети [13].

В. Гибридный нейро-нечеткий подход эффективен для интеллектуального анализа и прогнозирования веб-трафика. Это параллельная нейро-нечеткая модель *для обнаружения и анализа полезных знаний* из доступных данных веб-журналов.

Гибридная структура сочетает в себе *самоорганизующуюся карту (SOM)* и *систему нечеткого вывода (FIS),* работающую в параллельной среде.

Примечание: FIS – это популярная вычислительная среда, основанная на концепциях теории нечетких множеств, нечетких правилах «если-то» и нечетких рассуждениях.

Базовая структура FIS состоит из трех концептуальных компонентов:

1) база правил, которая содержит набор нечетких правил;

2) база данных, которая определяет функции принадлежности, используемые в нечетком правиле, и

3) механизм рассуждений, который выполняет процедуру вывода разумного результата или заключения на основе указанных правил и заданных фактов.

Информация о трафике собирается и отправляется сборщику трафика, который затем пересылает ее анализатору. В этом анализе *Netflow* также является одной из программ, собирающих информацию о трафике.

Существует и еще несколько подходов (и соответственно, инструментов) к проведению отслеживания поведения посетителей на веб-сайтах, применяемых в рамках анализа веб-трафика. Информация, полученная в результате такого отслеживания и анализа трафика, может помочь онлайн-продавцам ориентироваться на определенную аудиторию с помощью индивидуальных продуктов и услуг, а также расширять ее.

Некоторые из таких подходов отслеживания и анализа имеют сравнительно меньше ограничений и *больше преимуществ*, а их *недостатки* в основном связаны с (1) некоторыми ограничениями в отношении типов данных, которые они могут отслеживать, и (2) характеристиками среды пользователей. Такие ситуации мотивируют необходимость разработки иных инструментов, которые могут отслеживать столько же данных, но с меньшими ограничениями и большими преимуществами, чем рассмотренные выше методы и инструменты.

К этим подходам относятся: *улучшенное однопиксельное изображение, отслеживание JavaScript и прокси-сервер HTTP* (протокол передачи гипертекста), которые часто работают вместе для отслеживания действий пользователя. В дополнение к базовому анализу также используют *расширенный анализ*, включающий построение дерева *анализа путей* и *кластеризацию пользователей.*

Г. Однопиксельный анализ собирает данные с каждой страницы, которую просматривает посетитель. Одиночный пиксель известен под разными именами, такими как *пошаговые GIF-файлы, невидимые GIF-файлы, маяковые GIF-файлы и веб-ошибки*.

Однопиксельный метод вставляет тег IMG (*тег IMG* используется при написании сайта, а именно, предназначен для отображения на веб-странице изображений в графическом формате GIF, JPEG или PNG. Тег содержит адрес файла картинки и текст, который замещает картинку, если она не может загрузиться) на веб-страницу. Этот тег включает в себя некоторые коды HTML и JavaScript, которые интерпретируются браузером пользователя, когда пользователь получает такую страницу с сервера сайта и, таким образом, захватывает соответствующие данные с этой страницы.

Однопиксельная технология обычно используется с файлами cookie и JavaScript и основана на механизме загрузки изображений. Когда страница загружается, браузер не только создает соединение с веб-сервером, на котором находится страница, но также создает соединение с веб-сервером (на котором находится изображение) для запроса изображения.

Преимуществом однопиксельной техники является то, что файлы веб-журналов фиксируют активность пользователя по «обращениям». Это означает, что если на странице есть три изображения, то в файле журнала будет создано не менее четырех отдельных записей для запроса этой страницы. Эти четыре записи представляют собой результаты запроса для этой страницы и по одной записи для каждого изображения.

В отличие от файлов веб-журнала, однопиксельный метод собирает информацию по просмотрам каждой из страниц и делает по одной записи для каждой страницы. Подробная информация собирается, когда страница полностью загружена. Одним из очевидных *преимуществ* однопиксельной техники является то, что информацию можно легко собрать и поместить в базу данных, в то время как для анализа веб-журналов обычно требуются данные с нескольких веб-серверов, которые могут быть расположены в разных местах. Это означает, например, что если страница имеет пять тегов IMG, и эти пять изображений расположены на пяти серверах, то запрос на эту страницу приведет к записи в каждом файле журналов этих пяти серверов. Чтобы проанализировать данные, придется собрать эти записи с этих пяти серверов, тогда как при однопиксельном анализе создается одна запись, потому что она связана с одним изображением. После загрузки изображения запускается программа отслеживания, которая записывает все необходимые данные в виде одной записи.

При работе с кодом JavaScript на *стороне клиента*, который может захватывать больше данных на стороне клиента, однопиксельный метод позволяет захватывать больше данных, чем инструменты, работающие *на стороне сервера* (такие как файлы веб-журналов и мониторы пакетов), и даже те, которые не могут быть получены этими серверными инструментами.

Недостатки однопиксельного метода таковы:

1) этот подход, основанный на JavaScript, может не захватить необходимые данные, когда браузер посетителя отключает JavaScript, или когда браузер не поддерживает JavaScript [3; 8; 16];

2) как и при прослушивании пакетов, однопиксельный захват данных осуществляется в реальном времени, и если что-то повлияет на загрузку изображения, данные будут потеряны;

3) при использования этой технологии, поскольку она является скрытой, в отличие от файлов cookie (файл cookie – это небольшой фрагмент текста, передаваемый в браузер с сайта, который посещает пользователь. С его помощью сайт запоминает информацию об этом посещении, что удобно для пользователя. Кроме входов в аккаунты, файлы cookie умеют запоминать: предпочтения пользователей, а также язык, валюту, размер шрифта; товары, которые они просматривали или добавили в корзину; текст, который вводили на сайте раньше; IP-адрес и местоположение пользователя; дату и время посещения сайта; версию ОС и браузера; клики и переходы.), веб-браузеры не предлагают настройки предпочтений, чтобы предупреждать пользователей о том, что страницы содержат веб-ошибки [12]. Некоторые потребители могут быть недовольны осознанием того, что за ними наблюдают и отслеживают их действия в сети [2; 19].

Д. Модель отслеживания и анализа трафика на языке программирования Java. В ней используется строка *пути* для представления *полного пути* в сеансе пользователя. Для эффективного хранения всех путей и визуального отображения данных разработана древовидная структура путей и алгоритм построения такого *дерева путей*. Результаты ее использования показывают, что эта модель может предоставить много полезной информации о навигации пользователей и использовании веб-сайта. Кроме того, она может быть использована для разработки систем веб-рекомендаций и персонализации.

В этой модели применяется формальная техника для руководства императивами обеспечения качества при моделировании подхода к отслеживанию и анализу, основанная на использовании дискретной математической теории (теоретико-множественных представлений для описания структурных компонентов, а также логики предикатов – для описания требований).

Е. Мониторы пакетов. Мониторы пакетов собирают данные трафика непосредственно из пакетов TCP/IP, отправляемых на данный веб-сервер и уходящих

с него. В них используется технология прослушивания пакетов, которая схожа с прослушиванием телефонных разговоров [3]. Когда веб-браузер подключается к веб-серверу, он взаимодействует с сервером, отправляя запросы и получая ответы. Все запросы и ответы передаются в виде данных, которые фактически разделяются на пакеты TCP/IP по протоколу HTTP вместе с другими протоколами Интернет-приложений. Эти пакеты можно «прослушивать», когда они перемещаются по сети, для сбора данных о трафике веб-сайта.

Существует два типа мониторов пакетов, различаемых по месту сбора данных о трафике. Это *монитор сервера* и *монитор сети*.

Монитор сервера запускается как подключаемый модуль к веб-серверу. Он получает информацию о каждом событии, происходящем на отслеживаемом им веб-сервере, через интерфейс прикладного программирования. При этом то, какие события и данные видны серверным мониторам, зависит от веб-сервера [6]. Обычно монитор сервера может получить идентификатор посетителя, страницы реферера и некоторую другую информацию о событиях на сервере.

Сетевой монитор – это *сниффер пакетов*, который может захватывать все пакеты данных, проходящие через сеть (сниффер – это «анализатор трафика», то есть программа или устройство для перехвата и анализа сетевого трафика. Часто используется для анализа сетевого трафика в целях обнаружения и устранения отклонений и обеспечения бесперебойной работы. Однако сниффер может быть использован с недобрым умыслом. Снифферы анализируют все, что через них проходит, включая незашифрованные пароли и учетные данные, поэтому хакеры, имеющие доступ к снифферу, могут завладеть личной информацией пользователей. Сетевые мониторы используются чаще, чем серверные, для отслеживания трафика. Они могут видеть почти все, что находится между посетителем и данным веб-сервером. Отслеживаемая *информация включает запросы, ответы сервера, файлы cookie и переданные файлы HTML*. Кроме того, сетевой монитор может отслеживать события остановки, генерируемые браузером, что позволяет владельцу сайта знать те страницы, создание и отображение которых занимает слишком много времени в браузере клиента. Он также может записывать время

ответа веб-сервера на каждый запрос. Некоторые сетевые мониторы могут захватывать «данные формы» HTML, передаваемые методом POST, когда посетитель нажимает кнопку отправки [6]. *Сетевой монитор* может быть установлен на каждом веб-сервере, хотя такой конфигурацией сложно управлять, если веб-серверы находятся в разных географических точках.

В целом, *преимущества анализа пакетов над файлами веб-журналов* таковы:

1) данные можно собирать и анализировать *в режиме реального времени*;

2) можно отслеживать почти всю информацию, находящуюся в файле Web-журнала, а также информацию на сетевом уровне, такую как события «остановки», которых нет в файлах журнала;

3) *анализ пакетов* может быть совместим практически с любым настраиваемым веб-сервером, поскольку он не зависит от формата файла журнала и базовой операционной системы (ОС);

4) организации с распределенными веб-серверами могут легко (автоматически) собирать информацию о трафике в централизованном хранилище данных для последующего их анализа.

Недостатками использования мониторов пакетов являются:

1) мониторинг пакетов сервера может вызвать неожиданные проблемы на самом сервере (например, может привести к выключению сервера при сбое монитора пакетов), поскольку монитор действует как подключаемый модуль;

2) типы данных, которые может отслеживать серверный монитор пакетов, зависят от веб-сервера;

3) хотя сетевой монитор может отслеживать больше данных, чем файлы веб-журнала, он потребляет много процессорного времени и, следовательно, вызывает большую нагрузку на веб-серверы;

4) метод анализа пакетов захватывает данные в реальном времени, и данные не регистрируются сразу, поэтому, если что-то пойдет не так с анализатором пакетов, данные будут потеряны.

Другие подходы. Помимо упомянутых выше методов отслеживания и сбора данных о посетителях в Интернете, также используются некоторые другие подходы. Опишем кратко некоторые из них.

1. Использование HTML-форм – это самый прямой метод сбора данных от посетителей Интернета. Он позволяет сайту собирать информацию о своих посетителях и о том, что они хотят. После того, как посетитель заполняет форму и отправляет ее, информация, содержащаяся в форме, попадает в программу на *стороне сервера*, которая анализирует и сохраняет данные в базе данных для последующего их анализа.

Основное *преимущество* этого подхода состоит в том, что веб-сайт может собирать интересные для анализа данные, благодаря использованию заранее разработанных форм, и отслеживаемые данные легко сохранить и проанализировать позже. Однако этот подход *малопривлекателен*, поскольку большинство пользователей Интернета не желают тратить несколько минут на заполнение формы, если она не является обязательной.

2. Скрытые поля HTML фиксируют данные сеанса посетителя о нем. При отправке формы указанное имя и значение включаются в данные GET или POST. Скрытые поля HTML просты в использовании, но они работают только в том случае, если каждая страница создается в динамическом режиме.

3. Файлы HTTP куки.(cookie) – содержат небольшие фрагменты текстовой информации, которую веб-сервер отправляет веб-браузеру при первом подключении браузера к веб-сайту. При последующих посещениях веб-браузер отправляет тот же идентификатор обратно на веб-сервер, сообщая веб-сайту, что вернулся тот же конкретный пользователь. Разработчики веб-сайтов могут легко идентифицировать отдельных посетителей с помощью файлов cookie, что позволяет лучше понять, как используется сайт. Например, посетителям большинства сайтов заказов не нужно повторно вводить уже сохраненную информацию о сеансе или некоторые личные данные каждый раз при последующих посещениях. Этот метод широко используется для сбора данных от посетителей. Основным *недостатком* этой техники является тот факт, что «браузеры обычно принимают

всего 300 файлов cookie» [12]. Следовательно, файлы cookie не могут использоваться для хранения большого количества информации о каждом посетителе.

4. Перезапись URL-адресов – это еще один метод управления сеансом, при котором браузер добавляет дополнительные данные, идентифицирующие сеанс, в конец каждого URL-адреса, а веб-сервер связывает идентификатор сеанса с данными, которые он хранит о сеансе. Этот подход работает даже тогда, когда браузер не поддерживает файлы cookie или когда посетитель отключает файлы cookie. *Недостатком* этого метода является то, что он добавляет утомительные задачи обработки веб-серверу. Кроме того, если посетитель покидает сеанс и возвращается по закладке или ссылке, информация о сеансе может быть потеряна.

Сводные *результаты исследования* таковы:

Рассмотрены модели и инструменты отслеживания и анализа трафика для эффективного управления транзакциями электронной торговли. Этот анализ предназначен для того, чтобы помочь организациям, занимающимся электронной коммерцией, больше узнать о своих пользователях, чтобы иметь возможность разработать *эффективные маркетинговые стратегии*. В частности, можно сделать *следующие выводы*:

1. Комбинация трех подходов к отслеживанию – улучшенного однопиксельного подхода, подхода JavaScript и прокси-сервера HTTP – всегда позволяет отслеживать пользователя.

2. *Улучшенный однопиксельный метод* имеет больше явных преимуществ по сравнению с обычным однопиксельным методом, поскольку не требует поддержки JavaScript для работы, тогда как обычный метод требует поддержки JavaScript в браузере пользователя.

3. Две основные особенности делают *метод JavaScript* более эффективным и отличающимся от других, поскольку:

– он может отслеживать большинство событий, вызываемых манипуляциями клавиатуры и мыши онлайн-пользователей. При этом аналитик может сам настроить, какие именно события о пользователе он хочет отслеживать;

– метод способен отслеживать ввод любой формы на странице.

4. *Улучшенные однопиксельные подходы и JavaScript* предназначены для работы как со статическими, так и с динамическими веб-страницами.

5. Как в *улучшенном однопиксельном подходе, так и в подходе JavaScript* использование технологии сервлетов Java (сервлет – это механизм, который умеет получать запросы от клиента и возвращать ему ответы. Сервлет взаимодействует с клиентами посредством принципа запрос-ответ. По сути, это механизм, обслуживающий сайты, его пишут разработчики, потому термин узкоспециализированный.) позволяет этим подходам быть более совместимыми со многими платформами, которые используются в большинстве методов отслеживания. Кроме того, технология управления сеансами сервлетов Java позволяет легко управлять сеансами пользователей и приводит к сравнительно простому процессу последующего анализа отслеживаемых данных.

6. В дополнение к *базовому анализу* на основе отслеживаемых данных в настоящее время используется *расширенный анализ*, такой как *модель дерева путей* и *кластеризация* пользователей.

Изложенный материал обзора полезен для дизайнеров и разработчиков маркетинговых *приложений электронной коммерции*, поскольку в нем освещаются некоторые ключевые фундаментальные и/или прагматические проблемы, возникающие при разработке приложений для отслеживания трафика *виртуальной торговли*. Изучение операционной эффективности и результативности работы сети, безусловно, важно, потому что позволяет понимать, применять и продвигать известные принципы, процессы и практики для преодоления выявляемых проблем отслеживания вэб-трафика, которые в конечном итоге влияют на общее качество функционирования виртуального бизнеса и, следовательно, на уровень удовлетворенности клиентов веб-сайтов.

Список литературы

1) Филимонов О.И. Интернет-трафик: понятие, свойства и типология / О.И. Филимонов, Т.Г. Касьяненко // Вопросы инновационной экономики. – 2021. – Том 11. – №3. – doi: 10.18334/vinec.11.3.113230

2) Ackerman S.M., Cranor L.F., Reagle J. Beyond concern: understanding net users' attitudes about online privacy // AT&T Labs-Research Technical Report TR 99.4.3, AT&T Labs-Research, 1999.

3) Analyzing web site traffic, White Paper, Sane Solutions, LLC, North Kingstown, Rhodes Island, USA, 2002.

4) Assessing web site usability from server log files. White Paper, Tec-Ed, Inc., Ann Arbor, MI, USA, 1999.

5) Aye T. Web Log Cleaning for Mining of Web Usage Patterns. IEEE, 2011.

6) Building Confidence Electronic Commerce and Development, in United Nations Conference on Trade and Development (UNCTAD), UNCTAD/SDTE/MISC.11, UNCTAD (Geneva, 2000). Pp. 18.

7) Davison B.D. Web traffic logs: an imperfect resource for evaluation // Proceedings of Ninth Annual Conference of the Internet Society, INET (San Jose, CA, 1999).

8) Driving business decisions in web time, White Paper, Accrue Software, Inc., Fremont, CA, USA, 2000.

9) Drott M.C. Using web server logs to improve site design // Association for Computing Machinery (ACM), in Proceedings on the 16th Annual International Conference on Computer Documentation (Quebec City, Canada). 1998. Pp. 43–50.

10) Ehikioya S.A., Lu S. A Traffic Tracking Analysis Model for the Effective Management of E-commerce Transactions // International Journal of Networked and Distributed Computing. 2020. Vol. 8. Pp. 171–193.

11) Neha Goel, C.K. Jha. Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool // International Journal of Computer Applications. 2013. №2. Pp. 29–33.

12) Hall M. Core Servlets and JavaServer Pages // Sun Microsystems Press, Menlo Park, CA, USA. 2000. 181 p.

13) Introduction to Cisco IOS Netflow – A Technical Overview [Electronic resource]. 2012. URL: https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html (date of treatment: 01.08.21).

14) Joshila Grace L.K., Maheswari V., Nagamalai D. Analysis of Web Logs and Web User In Web Mining // International Journal of Network Security & Its Applications (IJNSA). 2011. Vol. 3. №1.

15) Krishnamurthy B., Rexford J. Software issues in characterizing web server logs // World Wide Web Consortium Workshop on Web Characterization, Cambridge, MA, Cambridge, 1998.

16) Malacinski A., Dominick S., Hartrick T. Measuring web traffic, part 1 and part 2, DeveloperWorks, IBM Corporation, Armonk, NY, USA, 2001.

17) Navin K. Tyagi, Solanki A.K., Wadhwa M. Analysis of Server Log by Web Usage Mining for Website Improvement // International Journal of Computer Science Issues. 2010. Vol. 7. №8. Pp. 17–21.

18) Pande P.V., Tarbani N.M., Ingalkar P.V. A Study of Web Traffic Analysis // International Journal of Computer Science and Mobile Computing. 2014. Vol.3. Pp. 900–907.

19) Privacy notice research, final results, Harris Interactive, Inc., Privacy Leadership Initiative (PLI), 2001. Study No. 15338.Seymour J. Mining for gold in your web traffic logs, Sharper Edge International Pty Ltd., Beecroft NSW, Australia, 2014.

20) Seymour J. Mining for gold in your web traffic logs, Sharper Edge International Pty Ltd., Beecroft NSW, Australia, 2014.

21) Sylvanus A. Ehikioya, Shenghong Lu // A Traffic Tracking Analysis Model for the Effective Management of E-commerce Transactions. 2020.

22) Wilson T. Web traffic analysis turns management data to business data // TechWeb Network, San Francisco, CA, USA, 1999.

Филимонов Олег Игоревич –соискатель ФГБОУ ВО «Санкт-Петербургский государственный экономический университет»; финансовый директор ИП Adlook.me, Россия, Санкт-Петербург.
