УДК 004.8

Юсифов Ф.Ф.
Институт Информационных Технологий НАНА, Баку, Азербайджан
y_ferhad@yahoo.com, yusfar@rambler.ru

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ INTERNET ИСПОЛЬЗОВАНИЕМ ЛОГ-ФАЙЛОВ

В данной статье осуществлен краткий обзор Web mining технологий, применяемых для сбора информации и извлечения знаний об использовании Internet. Рассмотрены наиболее эффективные принципы интерпретирования данных лог-файлов web-серверов.

***Ключевые слова**: Web mining, лог-файл, web-сервер, web-трафик, идентификация.*

Введение

Извлечение знаний можно определить, как нахождение и анализ полезной информации. Естественная комбинация двух областей – интеллектуальный анализ данных с WWW, называемых Web mining, стал центром проводимых в последние годы научно-исследовательских работ.

Термин Web mining использовался в основном в трех направлениях [1–3]. Первое, которое называется Web content mining, описывает процесс обнаружения информации или ресурса из миллионов источников WWW. Второе, называемое Web structure mining, – это процесс извлечения структурной информации из Web. Третий, называемый Web usage mining, является процессом анализа журналов доступа Web или другой пользовательской информации и доступа на одной или нескольких окрестных сетях. Web mining может быть широко определен как нахождение и анализ полезной информации из WWW. Это широкое определение, с одной стороны, описывает автоматический поиск, поиск информации и средств, доступных от миллионов сайтов и взаимосвязанных баз данных, с другой стороны, обнаружение и анализ пользовательских образов доступа от одних или нескольких серверов Web.

Рост объема доступных через Internet данных, хранимых в слабо структурированном виде, способствовал появлению автоматических программных средств поиска информации и получения данных об использовании определенных ресурсов [4, 5].

Лог-файл – это файл, содержащий системную информацию о работе сервера и информацию о действиях пользователей: дату и время визита пользователя, IP-адрес компьютера пользователя, наименование браузера пользователя, URL запрошенной пользователем страницы, реферер пользователя. Лог-файлы используются для анализа и оценки сайтов и их посетителей [4–6].

Использование методов интеллектуального анализа данных положило начало созданию серверных и клиентских интеллектуальных систем, которые могут эффективно добыть знания из WWW. С этой целью возник целый ряд интеллектуальных систем, основная задача которых состоит в эффективном извлечении знаний из Internet [7].

Основные принципы

Процесс автоматического изучения характеристик доступа пользователей к серверам, в основном, может включать изучение наиболее популярного

нахождения ассоциативных правил, путей посещения, кластеризацию, классификацию и т.д. Для решения этих задач можно использовать накопленные Internet журналы регистрации (лог-файлы) web-сервера. Организации собирают огромные объемы информации, автоматически создаваемые серверами и оседающие в журналах. Источниками информации являются также ссылочные журналы, в которых содержится информация для каждой страницы, на которую есть ссылка, журналы браузеров и регистрационные или анкетные данные пользователей, собранные CGI-сценариями [4, 5, 8].

С самого начала существования WWW разработчики web-серверов придерживаются традиционного формата представления данных о регистрируемом сервером трафике. Для регистрации используются четыре основных файла: access log (журнал регистрации доступа), error log (журнал регистрации ошибок), referrer log (журнал ссылок) и agent log (журнал агентов) [4]. Комбинации этих журналов могут варьироваться, но именно они являются единственным источником информации о трафике. Наиболее важным является журнал регистрации доступа, так как он содержит все HTTP-запросы для каждого обращения к web-странице, графическому изображению, CGI-программе, аудиоклипу или другому объекту независимо от того, каким было это обращение – удачным или нет. Однако HTTP-запрос и обращение к web-странице – это не одно и то же. Можно сказать, что даже сама, непосредственная web-страница может содержать с полудюжины графических изображений. При просмотре такой страницы посетителем web-сервер обслуживает не только HTTP-запрос самой страницы, но также HTTP-запросов изображений. Таким образом, при обращении к странице, содержащей одно или несколько графических изображений, в журнале доступа появится такое же количество регистрационных записей [4–6].

Журнал регистрации ошибок – второй, не менее важный файл регистрации. Хотя с точки зрения статистики не в такой степени, как предыдущий, но, тем не менее, он имеет особое значение для администраторов и web-мастеров. В этот журнал заносятся записи в том случае, если web-сервер регистрирует ошибку или аварийное состояние, так как большинство ошибок, в частности таких, как обращение к несуществующей странице, регистрируется в журнале доступа. Однако некоторые события записываются только в журнал ошибок. Например, если в ходе пересылки страницы читатель откажется от ее просмотра, в журнале ошибок появится запись send lost connection, а в журнале доступа для большинства серверов такая запись не будет зарегистрирована. В журналах ссылок и агентов регистрируется дополнительная информация о посетителях: указатель URL-страницы, из которой посетитель «попал» на данный узел, а также тип его web-браузера. При желании привлечь внимание к новому web-узлу такая информация поможет вам получить представление о том, как ваши будущие посетители узнают о его существовании [4, 5].

Основные потребители систем категории Web usage mining – организации, для которых главными задачами являются персонификация наполнения страниц и оптимизация сайта, также подобные системы представляют интерес для провайдеров Internet и сетевых администраторов. Основными областями применения в этом случае являются оптимизация работы Сети, минимизация трафика и оптимизация предоставляемых услуг [4, 8].

Мониторинг Сети оставляет открытой дверь для потока коммуникации данных, позволяя лучше понять происходящие процессы [9]. Большинство

традиционных систем мониторинга web предоставляют возможность фильтрации и получения статистической информации о пользователях. Подобный инструментарий помогает определять количество обращений к разным файлам и серверам, адреса отдельных пользователей, при этом такие системы рассчитаны на малый или ограниченный поток данных и редко предоставляют возможность анализа связи между обращениями к файлам и логикой их расположения.

Сбор и предоставление информации

Обычно сбор информации на уровне сервера представляет собой отбор информации непосредственно из журналов web-сервера. Этот способ используется наиболее часто, поскольку без излишних накладных расходов можно получить достаточно полную картину работы пользователей с сервером. Кроме того, это один из немногих методов, для которого уже существуют заранее накопленные данные. Действительно, все или почти все серверы автоматически ведут журнализацию, при этом журналы, как правило, хранятся годами. Рассмотрим детально, какую именно информацию предоставляет журнал сервера, отвечающий требованиям стандарта Common Log Format (CLF) [6, 10]. Используя этот формат, для каждого HTTP-запроса в журнале заполняется несколько информационных полей, а именно: имя пользователя (user name), обратившегося с запросом с удаленного компьютера; аутентификационное имя, под которым посетитель получил доступ к защищенной информации web-узла; имя хоста (или IP-адрес), с которого был сделан HTTP-запрос; дата и время; текст HTTP-запроса, направленного с удаленного компьютера; статус-код, показывающий, успешным ли был запрос; количество байтов данных, переданных в результате выполнения запроса на удаленный компьютер [4, 10, 11]. Обычно первые два информационных поля заполняются редко, так как большинство операционных систем не сообщает пользовательских имен, а для большинства Web не требуется аутентификация. Информация остальных пяти полей используется для аналитической обработки и построения на ее основе всевозможных графиков и диаграмм. Одна часть полученной информации может быть использована самостоятельно, например, количество просмотров определенного файла. Другая же можно использовать в совокупности, в частности для определения времени, которое посетитель затрачивает на чтение определенной страницы. Оно равно разности временных отметок двух следующих друг за другом обращений к web-страницам.

Комбинированный или расширенный формат журнала доступа является разновидностью стандартного Common Log Format [5, 6, 10]. При комбинированном формате к концу записи в журнале доступа добавляется информация из журналов ссылок и агентов. Это иногда гораздо удобнее, чем вести данные журналы отдельно, поскольку при отдельном ведении каждого из них зачастую невозможно восстановить соответствие той или иной записи определенным запросам из журнала доступа. Однако проблема заключается в возможности последующей обработки этой информации, так как некоторые пакеты для анализа web-трафика не могут работать с комбинированным форматом данных.

Можно сказать, что большинство современных web-серверов (в том числе Apache или IIS) предоставляют администратору возможность выбирать, какие поля должны включаться в журнал, а какие – нет. Самые распространенные из дополнительных полей, которые при добавлении к Common Log Format образуют

так называемое Combined Log Format, таковы: обратившееся приложение; URL документа, с которого осуществлено обращение; значения cookies [4, 12] .

Заманчиво собирать информацию о посещениях на уровне клиента. Один из способов – использование Java-программ, подгружаемых через страницы интересующего нас сервера; однако функциональность подобных приложений ограничена, кроме того, сам пользователь с помощью настроек своего браузера способен исключить или ограничить возможность подобного сбора информации. Вторым способом могло бы стать внесение изменений в программы для просмотра Сети. Но следует понимать, что вносить в журнал придется все сразу, поскольку если в будущем понадобится собирать данные по какому-либо новому параметру, то внести соответствующие изменения в браузеры всех клиентов будет почти невозможно. Также при таком подходе возникают две неразрешимые проблемы: во-первых, как правило, никто не хочет, чтобы его шаги протоколировались, а затем собранные данные куда-либо отсылались, во-вторых, мало кто станет обновлять свое программное обеспечение из-за нужд третьей стороны, осуществляющей сбор данных. Таким образом, сбор информации на стороне клиента более любых других методов затрагивает проблему сохранения неприкосновенности личной жизни, и пока такие методики мало применимы.

Созданием специализированного программного обеспечения для прокси-серверов можно добиться некоторых преимуществ по сравнению со сбором на стороне сервера или клиента. Решается проблема со снижением быстродействия сервера; кроме того, достаточно просто можно осуществить подключение нового сайта к сбору статистики или обновление системы для взаимодействия с новыми версиями [5].

Как альтернативу сбору информации на стороне сервера или шлюза можно рассмотреть сбор данных на узлах сети. Во-первых, не всегда возможен доступ к журналам сервера, во-вторых, не всегда данные, собираемые на сервере, релевантны к решаемой задаче [4, 6, 12]. Кроме того, добавление на сервер каких-либо программных средств сбора интересующей информации может быть невозможно или может просто замедлить сервер, что крайне нежелательно. Выходом может служить размещение датчиков в узлах сети на подходе к серверу. В таком случае сервер разгружается от излишнего программного обеспечения.

Можно заметить, что независимо от места сбора информации в полном потоке данных содержатся пароли или другие данные. Даже IP-адреса источника или получателя в некоторых случаях могут быть сочтены частной информацией, особенно с учетом того, что по адресу можно определить компьютер, с которого была произведена операция. Можно исключать из обработки подобные данные, но это приводит к потерям ценной информации. Разработчики должны выбирать подходящие варианты для эффективного анализа.

Подготовка данных

Одним из важнейших этапов для эффективного анализа лог-файлов является подготовка данных. На этом этапе могут выполняться некоторые простые интеграционные задачи, например, совмещение нескольких журналов и отсев ненужных для решаемой задачи данных [3, 10, 13]. Найденные ассоциации полезны только в том случае, если данные в журнале показывают точную картину доступа пользователей к сайту, иногда удаление записей о файлах с «неважными» суффиксами (jpg, gif, map и др.) может существенно очистить записи. Во многих

случаях требуется также очистить записи от неудачных запросов (например, оставить только строки, в которых ответ сервера имеет код 200) [5, 8]. Но гораздо более сложная проблема состоит в определении обращений, которые не заносятся в журнал, механизмы локального кэша или прокси-сервера искажают картину перемещений пользователей в Internet. Сейчас для преодоления этой проблемы используются различные методы. На данный момент методы борьбы с этой проблемой используют топологию сайта и ссылочные журналы вкупе с временной информацией для обнаружения пропущенных ссылок. Более-менее точную картину перемещений пользователя можно составить, только если пропуски страниц были единичными (например, если клиент использовал в своем браузере переход по журналу назад), в таком случае можно дополнять путь пользователя; эта проблема получила название «заполнение пути» (path completion) [6, 11, 13].

После того как данные очищены, возникает задача разбиения журнала на различные сеансы различных пользователей. Для того чтобы однозначно различать обращения различных пользователей из рассмотренных выше полей журнала, можно использовать IP-адрес, агент пользователя и адрес пользователя, вызвавшего документ.

Рассмотрим четыре основных типа спорных ситуаций для идентификации различных пользователей [5, 6, 10, 11]:

- один IP-адрес/много пользователей. Очень распространенная ситуация возникает при использовании провайдером прокси-сервера, кроме того, когда любому пользователю при установке связи с провайдером выделяется случайный адрес (очень характерно при связи по телефонной линии), два разных пользователя могут получить одинаковый адрес;
- много IP-адресов/один пользователь. Также весьма распространенный случай, возникает при динамическом выделении адресов провайдером. В этих ситуациях IP-адреса пользователей изменяются в каждом соединении;
- много IP-адресов/один сеанс. В некоторых случаях (широко известный пример – AOL) новый адрес выделяется пользователю при каждом новом обращении к странице. Для двух этих ситуаций можно выделять разных пользователей, основываясь на типе браузера, и отслеживать путь пользователя за один сеанс, находя для каждого документ, вызвавший его, и таким образом выделять отдельные сеансы от входа на сайт до страницы, с которой не было перехода внутри сайта;
- один пользователь использует различные браузеры. В таком случае, если IP-адрес не дает достоверных данных, можно воспользоваться только двумя методами, описанными ниже, при этом надо учесть, что файлы cookie будут далеко не всегда корректно работать.

В любом из упомянутых случаев, если для идентификации не хватает данных журнала, можно использовать файлы cookies и уникальную регистрацию пользователей. У каждого из этих методов есть недостатки: пользователь может удалить файлы, находящиеся на его компьютере, а обязательная регистрация, помимо очевидных недостатков, не обязательно получает точные данные [8]. Другая важная задача – идентификация сеанса (session identification) доступа. Перед тем как будет выполнен какой-либо анализ использования, необходимо разделить данные на логические части, представляющие разные сеансы или транзакции. Сеанс пользователя – весь набор использованных страничных ссылок, сделанных им за одно посещение сайта. Проблема определения сеансов сходна с

определением отдельных пользователей. Самым популярным методом решения этой проблемы является выделение сеансов использования по временному принципу, когда два последовательных обращения с одного адреса считаются принадлежащими одному сеансу, если перерыв между этими обращениями не превысил заданный порог. Вторым широко используемым способом является поддержание per session cookies (на стороне пользователя хранятся данные только от первого визита на страницу до выключения браузера; анализ этих данных позволяет отличить одно посещение пользователя от другого).

Транзакции отличаются от пользовательского сеанса тем, что в них могут входить от одной до всех страниц маршрута пользователя за один или несколько сеансов, в зависимости от заданного условия. Основная задача разбиения работы пользователя на транзакции состоит в выделении групп семантически близких обращений одного пользователя, поэтому для разбиения могут использоваться и операция разбиения, и операции слияния. Таким образом, транзакция может быть меньше или больше, чем один сеанс [5].

Методы анализа и программное обеспечение

Как правило, по данным журнала сервера подсчитываются самая популярная страница, наибольшее количество посетителей за день, неделю, месяц, также могут быть выделены страницы, обращения к которым вызвали наибольшее количество ошибок. Можно применять статистический анализ к уже очищенному и разбитому на транзакции журналу. В этом случае функциональная полезность резко возрастает, появляется возможность подсчитывать статистические данные для продолжительности пребывания на различных страницах или длины транзакции. Конечно, сбор статистики не дает требуемой для методов извлечения знаний глубины, но любая система принятия решений предоставляет пользователю такого рода информацию как потенциально интересную и полезную. В качестве удобного интерфейса анализа получаемых данных часто используются методы data mining.

После идентификации отдельной транзакции аналитик может применить один из методов извлечения знаний из схемы доступа: анализ пути, нахождение ассоциативных правил и последовательностей образцов, кластеризация или классификация [3, 10, 11].

Для анализа пути используются различные виды графов, так как граф представляет некоторое отношение, определенное на странице (или другом объекте) [5]. Самым распространенным способом является построение графа, соответствующее физической структуре сервера, при этом страницы являются узлами, а ссылки между ними – направленными ветвями. Также могут быть использованы графы, основанные на типах страниц, когда ребра представляют совпадения между страницами, или ребрам соответствует количество пользователей, переходящих с одной страницы на другую. Большая часть исследований, посвященных нахождению частых путей или последовательностей ссылок, проведена для графов, отражающих физическую структуру. С помощью этой методики можно определять наиболее посещаемые пути в Сети.

В связи с тем, что подобные базы транзакций содержат множество информации, обычно технологии поиска ориентируются только на записи, доступ к которым осуществлялся не менее определенного числа раз. Обнаружение этих правил для организаций, занятых в электронной коммерции, может помочь в

разработке эффективного маркетинга. Также эта информация помогает при улучшении организации сетевого пространства.

Нахождение последовательностей образцов – обнаружение связи между различными операциями, происходящими на протяжении одного временного интервала. В серверных журналах транзакций каждое посещение клиента записывается с некоторым интервалом времени [10, 11, 13].

Другой важный тип связанности данных, также обнаруживаемый при помощи этой методики, – сходные временные последовательности. Например, нам может быть интересно найти общие характеристики у клиентов, обращавшихся к одному файлу в определенный период времени, или временной интервал, на протяжении которого интересующий нас файл чаще всего используется.

Обнаружение классификационных правил позволяет создать описание записей, принадлежащих к определенной группе в связи с общностью атрибутов. Это описание затем используется для классификации вновь добавляемых записей. При изучении использования Сети можно разрабатывать описание для клиентов, обращавшихся к определенным файлам, используя имеющуюся для этих клиентов демографическую информацию или схемы доступа.

Кластерный анализ позволяет сгруппировать клиентов или данные, которые имеют сходные характеристики. Кластеризация информации о клиенте с данными в журналах может позволить разработать и осуществить ряд маркетинговых стратегий.

Основной областью применения для кластер-анализа в Web usage mining является персонификация наполнения страниц. Пользователь распределяется в одну из категорий, после чего соответствующим образом изменяется выводимая для данного пользователя информация. Еще одной традиционной для кластеризации областью применения является поддержка принятия решений.

Подготовка и анализ данных для анализа лог-файлов осуществляется в такой последовательности:

1. Проверить время и дату заинтересованных записей.
2. Выполнить трассировку для определения IP-адреса.
3. Проверить маршрут и имя запрашиваемого файла.
4. Изучить коды сообщений.
5. Определить, каким браузером пользуется посетитель.
6. Проверить ссылки.

Основная задача программных средств анализа web-трафика – извлечение полезной информации из регистрационных журналов сервера, и самое удивительное, они с ней успешно справляются. Программное обеспечение для анализа web-трафика существует почти столько же, сколько и сама Всемирная паутина. Но лишь недавно разработчики коммерческих программных продуктов заметили эту нишу рынка и быстро заполнили ее. Некоторые разработчики концентрируют свои силы на отдельных программных продуктах, другие же тратят все свое время и деньги на попытки изменить направление анализа трафика в целом [8, 14].

Такие компании, как, например, Software Inc. и Interse Corp., занимаются созданием программных продуктов для анализа лог-файлов. Их продукты (WebTrends и Market Focus соответственно) обрабатывают журнал доступа в формате Common Log Format и, насколько он позволяет, генерируют на языке HTML детализированный отчет. Самым лучшим считается пакет WebTrends,

главным образом из-за его цветных графиков и диаграмм, а также формата отчетов, которые очень удобны для публикации статистики непосредственно на web-узле [4, 8, 14]. С помощью пакета WebTrends генерируются пять видов отчетов: статистика подключений, статистика оплаты, краткая статистика, техническая статистика и полный отчет – комбинация всех перечисленных отчетов. С помощью пакета Market Focus создаются 14 стандартных отчетов, включая отчет по запросам, географическому местоположению, тенденциям в изменении запросов и частоты обращений, дневной и почасовой пропускной способности, а также полный отчет, включающий все вышеперечисленное.

Для повышения эффективности работы в реальном времени создание новых программных средств с применением Web mining способно принести максимум пользы. Создание таких интеллектуальных анализаторов может помочь эффективно анализировать лог-файлы.

Заключение

Методы извлечения знаний об использовании Internet, на данный момент, становятся все более популярными. Хорошим показателем может служить возросшее число научно-исследовательских работ в этой области. Можно сказать, что на данный момент хорошо работающих систем, позволяющих проводить точный анализ Web, практически нет, а существующие системы мало эффективны. При этом, в связи с резким ростом числа пользователей Web, потребность рынка в подобных информационных системах крайне велика. С этой целью, для решения важных проблем в этой области, применение методов интеллектуального анализа данных может помочь решить ряд таких задач, как идентификация пользователей, идентификация сеанса доступа, сохранение конфиденциальности, идентификация транзакции и др.

Литература

1. Cooley R., Mobasher B., and Srivastava J. Web mining: Information and pattern discovery on the World Wide Web. Proc. 9th IEEE Int. Conf. Tools with Artificial Intelligence, pp. 558–567, Nov., 1997.
2. Kosla R. and Blockeel H. Web mining research a survey, SIG KDD Explorations, vol. 2, pp. 1–15, July, 2000.
3. Srivastava J., Cooley R., Deshpande M., and Tan P., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 1(2):12–23, 2000.
4. Стаут Р. Анализ трафика Web-узла. http://www.ccc.ru/magazine/depot/97_03/read.html?0803.htm, 1997.
5. Щербина А. Основы извлечения знаний из Internet. http://www.osp.ru/os/2003/04/049.htm, 2003.
6. Iváncsy R., Vajk I. Different Aspects of Web Log Mining. 6th International Symposium of Hungarian Researchers on Computational Intelligence. Budapest, November, 2005.
7. Wanga X., Abrahamb A., Smitha K. A. Intelligent web traffic mining and analysis. Journal of Network and Computer Applications, vol. 28, pp. 147–165, 2004.

8. Рабин Д. Изучайте журналы посещений. Сети и системы связи №1 (121), http://ccc.ru/magazine/depot/05_01/read.html?0201.htm, 2005.
9. Уилсон Э. Мониторинг и анализ сетей. Методы выявления неисправностей. Изд. «Лори», 2002.
10. Baglioni M., Ferrara U., Romei1 A., Ruggieri S., and Turini F. Preprocessing and Mining Web Log Data for Web Personalization. www.di.unipi.it/~ruggieri/Papers/aiia2003.pdf , 2003.
11. Bartolini G. Web usage mining and discovery of association rules from HTTP servers logs. www.prato.linux.it/~gbartolini/en/view-a/2/pdf/wum.pdf, 2001.
12. Воннакот Л., Тэлли Б., Ванг Ю. Web-серверные платформы, http://www.osp.ru/nets/1997/09/108.htm, 1997.
13. Iváncsy R., Vajk I. Frequent Pattern Mining in Web Log Data. www.bmf.hu/journal/Ivancsy_Vajk_5.pdf, 2006.
14. Марков Р. WrSpy – считаем и контролируем трафик почтовых и прокси-серверов. «Системный администратор», Август, 2005.

UOT 004.8
Yusifov F.F.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
y_ferhad@yahoo.com, yusfar@rambler.ru
Log-fayllardan istifadə etməklə internetdən biliklərin aşkarlanması
Məqalədə İnternetdən biliklərin toplanması və aşkarlanması məqsədilə Web mining texnologiyasından istifadə olunması məsələsi tədqiqi olunur. Web-serverdə toplanan log-fayl verilənlərinin daha effetktli emal üsulları araşdırılır.
***Açar sözlər**: Web mining, loq fayl, Web-server, Web-trafik, identifikasiya.*

Yusifov F.F.
Institute of Information Technology ANAS, Baku, Azerbaijan
y_ferhad@yahoo.com, yusfar@rambler.ru
Knowledge discovering from Internet with use log files
In given paper is carried out the brief review the web mining technologies applied to collection of the information and discovering knowledge of use Internet. It is considered the most effective principles of interpretation data of log files of web servers.
***Key words:** Web mining, log file, Web-server, Web-traffic, identification.*