ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 50 за 2020 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Ревякин Ю.Г.

Возможности web-аналитики для оценки эффективности научных публикаций

Рекомендуемая форма библиографической ссылки: Ревякин Ю.Г. Возможности web-аналитики для оценки эффективности научных публикаций // Препринты ИПМ им. М.В.Келдыша. 2020. № 50. 42 с. http://doi.org/10.20948/prepr-2020-50
URL: http://library.keldysh.ru/preprint.asp?id=2020-50

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

Ю.Г.Ревякин

Возможности web-аналитики для оценки эффективности научных публикаций

Москва — 2020

Ревякин Ю.Г.

Возможности web-аналитики для оценки эффективности научных публикаций

Работа посвящена вопросам использования методов web-аналитики для оценки эффективности размещения научных публикаций в интернет. Дается общее представление о методах сбора и анализа данных, используемых в web-аналитике, представлен сравнительный обзор наиболее популярных программных решений. Обсуждаются вопросы использования методов web-аналитики для сайтов научных организаций. На примере сайта ИПМ РАН рассмотрены практические вопросы построения системы web-аналитики на основе анализа данных log-файла. Обсуждается проблема фильтрации запросов от программ-роботов для корректного определения показателей посещаемости научных интернет-ресурсов.

Ключевые слова: web-аналитика, научная публикация, анализ log-файла, фильтрация запросов, статистика посещений сайта

Yurii Gennadievich Revyakin

The opportunities of web-analysis to estimate the effectiveness of scientific publications

The work is devoted to the use of web-analytics methods for evaluation the effectiveness of scientific publications on the Internet. A general idea of the methods for collecting and analyzing data used in web analytics is given, a comparative overview of the most popular software solutions is presented. The issues of using web analytics methods for websites of scientific organizations are discussed. On the example of the IPM RAS website, practical issues of building a web analytics system based on the analysis of log file data are considered. The problem of filtering requests from robots to correctly determine the attendance metrics of scientific Internet resources is discussed.

Key words: web analytics, scientific publication, filtering requests, log-file analysis, web-site statistics

Оглавление

1. Web-аналитика: история возникновения, предмет изучения и развитие 5
1.1 Возникновение web-аналитики и предмет изучения 5
1.2 Развитие средств web-анализа 5
1.3 Факторы, влияющие на развитие web-аналитики 6
2. Основные понятия web-аналитики 6
2.1 Метрики и параметры 7
2.2 Ключевые показатели эффективности 9
3. Методы получения исходных данных 10
3.1 Анализ данных log-файла 10
3.2 Встраиваемый программный код 12
3.3 Сравнение основных технологий сбора первичных данных: достоинства и недостатки 12
4. Примеры программных систем сбора и анализа web-статистики 13
4.1 Google Analytics 13
4.1.1 Фильтры и сегменты GA 14
4.1.2 Цели и пути 15
4.1.3 Отчеты и средства их настройки 16
4.2 Средства web-аналитики с открытым кодом, основанные на анализе log-файлов 17
5. Web-аналитика для сайта ИПМ им. М.В. Келдыша РАН 20
5.1 Методы web-аналитики для сайтов научных организаций 20
5.2 Сайт ИПМ им. М.В. Келдыша РАН 23
5.3 Выбор программного решения для системы web-статистики 25
5.4 Анализатор log-файлов AWStats – краткая характеристика 26
5.5 Система web-статистики сайта ИПМ РАН 28
5.5.1 Статистика посещаемости за текущие сутки 29
5.5.2 Статистика посещаемости в текущем году 33

5.5.3 Статистика посещаемости электронной библиотеки 33
5.5.4 Архив отчетов посещаемости сайта за прошлые года 33
6. Оценка востребованности публикаций электронной библиотеки 34
6.1 Выбор критерия востребованности публикаций 34
6.2 Задача фильтрации запросов роботов в log-файле web-сервера 34
6.3 Счетчик числа обращений к текстам публикаций – особенности реализации 36
6.3.1 Краткая характеристика программы 36
6.3.2 Общая блок-схема алгоритма 37
6.3.3 Описание работы программы 38
6.3.4 Представление статистики обращений 40
7. Заключение 41
Список литературы 42

1. Web-аналитика: история возникновения, предмет изучения и развитие

1.1 Возникновение web-аналитики и предмет изучения

Как только Всемирная паутина превратилась из исследовательского проекта в глобальную распределенную информационную сеть, сразу же возникла потребность в инструменте, который бы позволял:

- оценить интерес пользователей сети к опубликованным в WWW материалам;
- понять, насколько выбранная форма представления материалов соответствует задачам, поставленным при их публикации.

Ответом на такую потребность стало появление web-аналитики как набора методик для сбора и анализа количественных и качественных данных об использовании интернет-ресурсов.

Согласно ставшему уже общепринятым определению [2, 3], под web-аналитикой понимается процесс сбора, измерения, представления и анализа данных о посещаемости интернет-ресурса для оценки и оптимизации использования ресурса пользователями сети.

1.2 Развитие средств web-анализа

Первые программные средства web-аналитики появились сразу после разработки стандартов World Wide Web (WWW) (1989) и создания первых программ-браузеров для пользователей Всемирной паутины (1990).

Можно выделить следующие основные этапы развития web-аналитики:

1990 – появления WWW и первых браузеров;

1993 – создание первого пакета статистики на основе анализа log-файлов;

1996 – реализация счетчика обращений к странице интернет-ресурса как удаленного сетевого сервиса;

1997 – появление первых систем сбора web-статистики, использующих для сбора исходных данных отслеживающий программный код, включаемый в содержимое страниц интернет-ресурса;

2006 – развитие методов web-аналитики, основанных на анализе характеристик поведения посетителя интернет-ресурса: продолжительность визита, последовательность просмотра страниц, взаимодействие с элементами web-интерфейса;

2012 – появление кросс-платформенных приложений web-аналитики для интегральной оценки активности пользователя на всех устройствах,

применяемых им для доступа в Интернет, – настольном компьютере, планшете или мобильном телефоне.

1.3 Факторы, влияющие на развитие web-аналитики

Вместе со Всемирной паутиной web-аналитика прошла достаточно впечатляющий путь эволюции и развития: от счетчиков посетителей одностраничных сайтов до систем анализа поведения пользователей при посещении страниц интернет-порталов. Общий вектор развития web-аналитики можно определить как переход от количественных оценок к качественным – потребовалось получить ответ не только на вопросы «что» или «сколько», но и «почему».

Для достижения этой цели общие показатели востребованности интернет-ресурса постепенно вытесняются показателями, характеризующими конкретного пользователя. Собираемые данные и генерируемые на их основе отчеты все более ориентируются на оценку интернет-ресурса с точки зрения эффективности решения тех прикладных задач, которые ставились при его создании.

Развитие получает и сравнительная web-аналитика, используемая для сопоставления показателей востребованности интернет-ресурсов, посвященных схожей или родственной тематике.

2. Основные понятия web-аналитики

Основанием для выполнения любого анализа являются исходные данные. Источником данных для web-аналитики является, в первую очередь, информация, содержащаяся в запросах, передаваемых от пользовательского приложения (браузера пользователя) к серверу (web-серверу) при просмотре гипертекстовых документов (html-страниц). Объектом таких запросов (называемых http-запросами) являются сам гипертекстовый документ и его составляющие – мультимедиа-файлы, файлы определения стиля документа и другие. Формат http-запроса определен http-протоколом – основным протоколом прикладного уровня, используемым в WWW для обмена файлами. Именно информация, содержащаяся в заголовке http-запроса, чаще всего используется средствами web-аналитики при анализе посещаемости интернет-ресурса. Кроме того, информация из заголовка http-запроса может быть дополнена:

- информацией уровня сетевого протокола. Примером такой информации является IP-адрес пользователя;

- данными прикладного уровня, включаемыми в http-запрос для обработки внешними web-приложениями. К данным прикладного уровня относятся, в частности, значения параметров, передаваемые при вызове программы-скрипта на стороне web-сервера;
- внешними источниками данных, которые используются для интерпретации информации, содержащейся в http-запросе. Внешние справочники используются, например, для определения по IP-адресу пользователя имени его интернет-провайдера или страны проживания.

2.1 Метрики и параметры

Основным результатом применения средств web-аналитики является построение отчетов, отражающих взаимосвязь двух типов данных: метрик и параметров (dimensions).

Метриками называют количественные показатели использования интернет-ресурса. Метрикой может быть счетчик, среднее значение или отношение двух метрик.

Наиболее часто используются следующие метрики.

Запросы – число http-запросов за определенный период времени.

Просмотры страниц – число запросов от посетителей сайта на передачу гипертекстовых документов за определенный период времени.

Визиты (сессии) – число визитов пользователей интернет-ресурса за определенный период времени. Под визитом понимается период активного обращения одного пользователя к интернет-ресурсу; все запросы к файлам ресурса, поступившие от пользователя в течение этого периода, считаются относящимися к одному визиту пользователя.

Уникальные посетители – число пользователей, определяемых своими уникальными характеристиками и посетивших интернет-ресурс хотя бы один раз за рассматриваемый период времени.

Отметим две основные особенности метрик, используемых в web-аналитике. Во-первых, они ориентированы на анализ активности реальных, «одушевленных» посетителей сайта. Во-вторых, существуют отдельные метрики, значения которых невозможно определить с абсолютной достоверностью. Так, мы никогда не можем точно определить моменты начала и окончания визита пользователя. Причина этого кроется в том, что http-протокол является протоколом передачи данных без установления соединения и сохранения состояния и предполагает, что запросы обрабатываются независимо друг от друга. Другим примером такой неоднозначности вычисления метрик может служить проблема определения уникальных

пользователей, посетивших сайт за некоторый отрезок времени. Для определения уникальных посетителей используется целый ряд признаков: IP-адрес машины пользователя, используемый им браузер, регистрационные данные пользователя. Но ни один из этих признаков не может быть использован со 100% достоверностью для однозначной идентификации пользователя: IP-адрес пользователя чаще всего не является постоянным и меняется от одного визита пользователя к другому; сам пользователь может использовать для доступа к интернет-ресурсу разные программы-браузеры. Для вычисления подобных проблемных метрик не существует единых общепризнанных методик, каждый из разработчиков средств web-аналитики использует собственные алгоритмы их определения, основанные на разных исходных данных и различных начальных допущениях.

Но метрики являются только количественными показателями использования интернет-ресурса и не представляют самостоятельной ценности. Поэтому в web-аналитике они рассматриваются относительно набора атрибутов, характеризующих посетителя сайта и его активность во время посещения web-ресурса. Такие атрибуты получили название параметров (dimensions).

Примерами параметров являются такие характеристики пользователей, как их IP-адрес и географическое положение, используемые ими браузеры и операционные системы, адреса посещенных пользователями страниц, введенные поисковые запросы.

Значения параметров, в отличие от метрик, являются текстовыми величинами.

Для каждого из параметров может быть определена одна или несколько метрик. Так, мы можем для каждого адреса страницы интернет-ресурса вычислить число обращений пользователей к этой странице за определенный отрезок времени.

Типовой отчет web-аналитики является двумерной таблицей, где строкам соответствуют возможные значения параметра, а колонки заполнены величинами метрик, вычисленными для этого значения. При этом, конечно, надо понимать, что не все виды метрик могут быть определены для произвольного параметра.

Большинство существующих средств web-аналитики допускает создание отчетов, включающих несколько параметров одновременно, хотя отчеты с числом параметров больше двух встречаются довольно редко.

Чтобы облегчить анализ представленных в отчетах данных, можно разбить множество вычисленных значений метрик на группы, используя как критерий разбиения значение параметра (параметров).

Несмотря на очевидность идеи разбиения данных на основе значений параметров, интерпретация полученных при этом результатов требует некоторой осторожности. Например, итоговое число «уникальных посетителей» за месяц, как правило, меньше суммы всех «уникальных посетителей» за каждый день месяца. Объяснение этого факта достаточно простое – один и тот же посетитель мог посещать сайт в разные дни месяца. При этом он учитывается как «уникальный посетитель» для каждого дня месяца, когда пользователь обращался к сайту. Аналогичная же статистика за весь месяц рассматривает его как одного «уникального пользователя».

Точно так же сумма новых и повторных посетителей интернет-ресурса за определенный отрезок времени может не совпасть с общим количеством посетителей, зарегистрированных за этот же период. Это неизбежно произойдет, если один из «новых» посетителей обратится повторно к сайту в течение рассматриваемого отрезка времени.

Таким образом, суммирование значений метрик дает правильную оценку только для таких разбиений данных, которые исключают двойной учет событий, рассматриваемых данной метрикой.

2.2 Ключевые показатели эффективности

Стандартные метрики web-аналитики позволяют в первую очередь оценить загруженность web-сервера запросами и объем порождаемого при этом сетевого трафика. Но они не позволяют понять, насколько реализация интернет-ресурса соответствует тем целям, которые ставились при его создании. Нужен набор количественных показателей, дающих понятную и однозначную оценку эффективности использования интернет-ресурса именно с точки зрения решаемых им прикладных задач. Такой набор количественных показателей определяют на основе полученных ранее стандартных метрик и называют **К**лючевыми **П**оказателями **Э**ффективности (KPI – Key Performance Indicators) интернет-ресурса.

В качестве КПЭ можно использовать как сами стандартные метрики, так и более сложные показатели, вычисляемые на их основе. Приведем в качестве примера достаточно часто используемый КПЭ, получивший название «**Показатель отказов**». Он определяется как отношение числа уникальных посетителей, покинувших интернет-ресурс после просмотра одной страницы, к общему числу уникальных посетителей за рассматриваемый период.

Вычисленное таким образом значение показателя позволяет оценить, насколько пользователь, может быть впервые попавший на страницы интернет-ресурса, мотивирован к дальнейшему изучению представленных на ресурсе материалов.

При этом любой КПЭ также является метрикой, но не каждая метрика может и должна рассматриваться в качестве КПЭ.

Выбор набора КПЭ для оценки эффективности работы интернет-ресурса невозможен без четкого понимания задач, которые ставились при его создании. Здесь нет и не может быть единых рецептов и рекомендаций.

3. Методы получения исходных данных

Все программные средства web-аналитики характеризуются в первую очередь используемым методом сбора первичных данных. Основываясь на этом, их можно разделить на несколько видов:

- программные средства, использующие в качестве источника первичных данных файлы журналирования (log-файлы) web-сервера, содержащие информацию о выполнении сервером каждого http-запроса;
- программные средства, использующие для сбора первичных данных специальный отслеживающий программный код, встраиваемый в текст страницы интернет-ресурса. Когда страница после запроса пользователя обрабатывается браузером, программный код выполняется и передает информацию удаленному серверу статистики сведения о выполненном запросе на просмотр страницы;
- программные средства, использующие оба описанных выше способа сбора данных;
- программные средства, основанные на анализе log-файлов прикладного уровня. Такие log-файлы могут вестись на сервере для протоколирования работы web-приложений, реализующих прикладную функциональность сайта. Формат и содержание такого log-файла определяются самим web-приложением; при этом записи log-файла фиксируют события, происходящие на более высоком уровне – уровне взаимодействия пользователей с web-приложением.

3.1 Анализ данных log-файла

Как уже было отмечено выше, любой web-сервер может записывать информацию обо всех выполненных запросах в отдельный log-файл. Формат записи log-файла определяется соответствующей директивой конфигурации web-сервера и, как правило, такая запись содержит достаточно подробную информацию о запросе и его источнике. Приведем в качестве примера

определение так называемого «комбинированного» формата log-файла сервера Apache в том виде, в каком оно задается в конфигурационном файле сервера:

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-
    agent}i\"" combined
```

где символом «%» выделены управляющие элементы строки формата, которые определяют поля каждой записи log-файла:

%h – соответствует IP-адресу или доменному имени узла сети, с которого был отправлен запрос;

%l – удаленное имя пользователя, под которым пользователь зарегистрировался на своем рабочем месте – узле сети;

%u – имя пользователя, определенное после выполнения пользователем аутентификации на web-сервере;

%t – временная отметка выполнения запроса;

%r – строка команды запроса;

%>s – код ответа, возвращенный сервером в ответ на запрос;

%b – размер ответа сервера на запрос в байтах;

%{Referer}i – адрес URL web-страницы – источника запроса (т.е. web-страницы, содержащей гиперссылку, при переходе по которой и был создан текущий запрос);

%{User-agent}i – строка, описывающая удаленное клиентское приложение, использованное для обращения к серверу: имя программы-клиента, идентификацию операционной системы и аппаратной платформы.

Например, обращение к заглавной странице сайта www.keldysh.ru может оставить в log-файле следующую запись:

```
194.226.57.126 - - [11/May/2020:16:48:49 +0300] "GET / HTTP/1.1" 200
    16541 "-" "Mozilla/5.0 (Windows NT 6.1; Win64; x64)
    AppleWebKit/537.36 (KHTML, like Gecko) Chrome/60.0.3112.90
    Safari/537.36",
```

где “-“ отмечены поля записи, не содержащие информации.

Наиболее существенной проблемой при использовании в качестве источника первичных данных web-аналитики записей log-файла является широкое распространение технологии кэширования web-страниц. При кэшировании запрошенная пользователем web-страница интернет-ресурса заменяется ее копией из кэша, при этом сам запрос на страницу до исходного сервера попросту не доходит. И, как неизбежное следствие, никакой информации о таком запросе log-файл web-сервера не содержит. Кэширование web-страниц выполняется как на рабочих местах пользователей используемыми

ими браузерами, так и специальными сетевыми службами (т.н. proxy-серверами). По современным оценкам, примерно в одной трети случаев в ответ на запрос пользователя о передаче web-ресурса выдается его кэшированная копия.

3.2 Встраиваемый программный код

Если исходные статистические данные передаются при обработке html-страницы браузером, то становится совершенно неважным, что пользователь получает в ответ на свой запрос – исходный оригинал страницы или его кэшированную копию.

Используя отслеживающий программный код, можно получить и дополнительную информацию о пользователе, не представленную в http-запросе: характеристики его рабочего места и/или последовательность действий при просмотре web-страницы.

3.3 Сравнение основных технологий сбора первичных данных: достоинства и недостатки

Оба представленных метода сбора первичных данных web-аналитики имеют свои преимущества, но и не лишены определенных недостатков.

Для средств web-аналитики, основанных на анализе log-файлов, можно кратко сформулировать их основные достоинства и недостатки следующим образом:

- нет необходимости вносить изменения в контент интернет-ресурса, вставляя фрагмент программного кода в страницы сайта;
- log-файл содержит информацию обо всех запросах, адресованных web-ресурсу, независимо от их источника. Таким образом, учитывается активность роботов и поисковых машин, которые, как правило, не интерпретируют и не выполняют программный код, включенный в состав html-страницы;
- log-файл содержит информацию обо всех запросах, адресованных интенет-ресурсу, включая запросы документов в формате pdf, doc, xls и мультимедиа-файлов;
- но при этом не учитываются запросы, перехваченные и обработанные службами кэширования интернет-контента.

Хотя в последнее время средства web-аналитики, использующие для сбора первичных данных специальный отслеживающий код, оказались достаточно популярны и распространены, они тоже не лишены известных

недостатков. Перечислим основные характеристики, присущие средствам web-аналитики этого типа:

- не требуют доступа к log-файлу web-сервера;
- не зависят от технологий кэширования, дают более точные и достоверные данные;
- позволяют собирать информацию о событиях, не связанных напрямую с запросом на объект интернет-ресурса, например действия пользователя при просмотре страницы, заполнении формы и т.п.;
- могут быть использованы организациями, не имеющими доступа к серверам своего интернет-ресурса или не располагающими собственными возможностями для установки и сопровождения средств web-аналитики. Модель сбора первичной информации позволяет организовать службу web-аналитики как удаленный сетевой сервис, доступный другим пользователям на платной или бесплатной основе;
- требуют внесения изменений в тексты html-страниц ресурса;
- зависят от настроек браузера пользователя, не работают при установке запрета на выполнения встроенного в html-страницы программного кода;
- не могут быть использованы для анализа активности роботов и поисковых систем;
- требуют дополнительных усилий по настройке сайта для сбора информации об обращениях к объектам, не являющимся html-страницами.

4. Примеры программных систем сбора и анализа web-статистики

4.1 Google Analytics

Основанный на методе включения отслеживающего программного кода в текст html-страниц, сетевой сервис Google Analytics получил столь широкое распространение, что стал на текущий момент почти синонимом термина «web-аналитика».

Сетевой сервис Google Analytics (или сокращенно GA) был создан компанией Google на основе системы web-аналитики Urchin on Demand, приобретенной Google вместе с фирмой-разработчиком Urchin Software. Впервые представленный в 2005 году, сетевой сервис стал одним из первых бесплатных сервисов web-аналитики и с тех пор активно поддерживается и развивается компанией Google.

Безоговорочный успех GA можно объяснить в первую очередь его доступностью. Он доступен с точки зрения приобретения, так как является

бесплатным для всех сайтов с числом просмотров страниц до 10 миллионов в месяц. Он доступен с точки зрения внедрения – для подключения сайта к GA достаточно выполнить несложную процедуру регистрации и затем обычным текстовым редактором включить в состав html-страниц сгенерированный GA фрагмент программного кода. Наконец, не требуется много усилий, чтобы начать использовать GA – сразу после регистрации и редактирования страниц сайта пользователь получает доступ к набору стандартных отчетов, дающих общую оценку востребованности размещенных в WWW материалов.

Но экспансия GA в мире web-аналитики объясняется не только его доступностью с точки зрения приобретения, внедрения и освоения. Другим важным свойством GA являются широкие возможности адаптации GA под конкретные задачи web-аналитики. Отчеты из стандартного набора отчетов GA дают только общее представление о востребованности интернет-ресурса. Поэтому GA содержит целый арсенал средств настройки и конфигурации, дающих пользователю возможность создать собственный сетевой сервис для анализа эффективности интернет-ресурса с точки зрения соответствия исходным требованиям и задачам.

Любое приложение web-аналитики, созданное с использованием GA, можно представить как иерархию вложенных объектов трех типов: **аккаунт** – **ресурс** – **профиль** (представление).

Аккаунт является контейнером верхнего уровня, объединяющим в одну группу схожие по своему назначению интернет-ресурсы.

Ресурс определяет интернет-ресурс или web-приложение, эффективность которого будет оцениваться средствами GA.

Представление (view) управляет доступом к статистическим данным и аналитическим инструментам.

Отметим дополнительно, что один аккаунт может объединять до 50 ресурсов, а для каждого ресурса может быть создано до 25 представлений.

Каждый из классов объектов GA имеет свой набор параметров конфигурации, их область действия определяется положением объекта в иерархии.

Чтобы оценить возможности средств настройки GA, рассмотрим основные инструменты, используемые для создания и конфигурирования собственного сервиса web-аналитики на основе GA.

4.1.1 Фильтры и сегменты GA

Одной из первичных задач web-аналитики является разбиение по определенному критерию всего объема исходных данных на отдельные

подмножества. Это позволяет в дальнейшем как сократить объем рассматриваемых данных, так и анализировать выделенные подмножества данных независимо друг от друга. Для такого разбиения исходных данных в GA доступны два инструмента – фильтры и сегменты. Отчасти пересекающиеся по своей функциональности, эти инструменты, тем не менее, существенно отличаются друг от друга.

Фильтры используются, как правило, для исключения по определенным критериям части данных из дальнейшего анализа. (Таким критерием может быть, например, принадлежность IP-адреса источника запроса определенному сегменту сети.) Они применяются ко всему потоку входных данных, и изменения, вносимые фильтрами, являются необратимыми. Все данные, отсеченные в результате применения фильтра, являются безвозвратно потерянными, и, наоборот, действие фильтра не распространяется на данные, полученные ранее момента его создания.

Сегменты создаются пользователем GA для выделения по определенному критерию из всего объема исходной информации о трафике отдельных наборов данных, которые затем будут рассматриваться средствами GA независимо друг от друга. Критерий отбора для сегмента состоит из одного или нескольких правил, определяющих условие вхождения данных в сегмент с учетом свойств пользовательской сессии, атрибутов пользователя и атрибутов запросов. Критерий отбора применяется ко всему доступному объему данных о трафике, включая данные, полученные до момента создания сегмента. Создание или удаление сегмента не изменяет сам массив исходных данных и не влияет на использовании других сегментов в представлении.

Создать сегмент пользователь GA может, используя интерактивный интерфейс, представленный на странице отчетов. Для определения критерия отбора есть как набор уже готовых правил отбора данных, так и возможность создавать свои собственные правила. При этом результаты сегментирования отображаются на представлении отчетов на странице сразу после определения нового критерия отбора.

4.1.2 Цели и пути

Стандартные параметры и метрики предназначены лишь для общей оценки эффективности использования интернет-ресурса. Поэтому GA дает возможность определять для любого интернет-ресурса собственные показатели эффективности, вводя понятие цели. Под целями понимаются те желаемые действия (просмотр определенной страницы, загрузка документа и т.п.),

которые мы ожидаем от каждого посетителя сайта. В качестве цели в GA можно определить событие одного из следующих типов:

- открытие пользователем определенной страницы, заданной своим URL;
- время, проведенное пользователем при просмотре ресурса: считается, что цель достигнута, если пользователь оставался на интернет-ресурсе дольше некоторого заданного значения;
- число страниц, просмотренных пользователем при посещении: цель считается достигнутой, если число просмотренных страниц превысило заданную пороговую величину;
- любое другое действие пользователя (нажатие кнопки, загрузка файла), если оно фиксируется для данного интернет-ресурса в GA.

Как только набор целей для интернет-ресурса определен, GA автоматически начинает формировать набор отчетов, представляющих информацию о числе достигнутых целей за интервал времени, распределении числа достигнутых целей по источникам трафика, группам посетителей и многое другое. Это, в свою очередь, позволяет оценить эффективность ресурса, используя такой КПЭ, как "**конверсия сайта**" – процент посетителей сайта, выполнивших целевое действие.

Для каждой цели мы можем задать предполагаемую последовательность действий (путь), которую посетитель сайта проходит перед тем, как цель считается достигнутой. Определив такую последовательность, можно средствами GA проследить прохождение посетителей по заданному пути и подсчитать число отказов на каждом этапе. Таким образом, мы не только получаем информацию о том, какая доля посетителей сайта выполняет нужное нам действие, но и можем предположить, почему остальные посетители этого не сделали.

4.1.3 Отчеты и средства их настройки

Как уже было отмечено ранее, в GA доступ к полученным статистическим данным определяется на уровне представления. Используя представление, можно указать набор формируемых отчетов, способ отображения отчетов на странице, то, как данные в отчетах будут сгруппированы и какие данные из рассмотрения исключаются. Для одного ресурса в GA допускается создание нескольких представлений.

В представлении можно использовать как стандартные типы отчетов, создаваемые GA по умолчанию для любого ресурса, так и конструировать, на основе набора доступных метрик и параметров, свои собственные. Допускается

объединять несколько отчетов на одной странице GA, формируя сводки отчетов.

Интерактивные средства настройки отчетов на странице GA дают пользователю возможность в реальном времени менять формат отображения данных, фильтровать данные по заданным условиям или группировать данные по определенному признаку.

4.2 Средства web-аналитики с открытым кодом, основанные на анализе log-файлов

Несмотря на то что средства web-аналитики, основанные на анализе данных log-файла web-сервера, утрачивают свою популярность и вытесняются сетевыми сервисами, подобными Google Analytics, некоторые из появившихся еще в 1990-х годах продуктов до сих пор активно используются для оценки эффективности использования интернет-ресурсов. Они обладают достаточно развитым функционалом и вполне могут быть использованы для решения базовых задач web-аналитики там, где обращение к сетевым сервисам неудобно по тем или иным причинам.

Средства web-аналитики этого типа, несмотря на все различия в программной реализации, основаны на достаточно схожей схеме обработки данных:

- содержимое log-файла web-сервера периодически сканируется для получения новых данных о посещаемости сайта;
- полученные данные могут быть использованы непосредственно для формирования отчетов или сохраняются во внутреннем промежуточном формате для оптимизации создания отчетов в будущем;
- на основе полученных из log-файла данных создаются отчеты в формате, определенном пользователем.

Сравнить характеристики наиболее известных программ web-аналитики, основанных на анализе log-файлов, можно, обратившись к таблице, приведенной в [6]:

Таблица 1

	AWstats	Analog	Webalizer	W3Perl	Sawmill Analytics
Версия	7.5 декабрь 2016	6.0.12 C:Amie Edition* июнь 2015	2.23-08 август 2013	3.20 июнь 2015	8.8.0 январь 2019
Язык программирования	Perl	C	C	Perl	C/Salang
Тип лицензии	GPL	GPL	GPL	GPL	Коммерческая
Поддержка формата log-файла Apache	Да	Да	Да	Да	Да
Поддержка формата log-файла Microsoft IIS	Да	Да	Да**	Да	Да
Обновление статистики/ создание отчетов	Командная строка и/или web-интерфейс	Командная строка и/или web-интерфейс	Командная строка	Командная строка CGI/web-интерфейс	Командная строка CGI/web-интерфейс
Подсчет числа пользовательских посещений	Да	Нет	Да	Да	Да
Подсчет числа уникальных посетителей	Да	Нет	Нет	Нет	Да
Число посещений по странам	Да	Да	Да	Да	Да
Число посещений по хостам	Да	Да	Да	Да	Да

Отчет по запросам роботов (число определяе-мых роботов)	Да (793)	Да (8)	Нет	Да	Да (250)
Отчет по стартовым страницам	Да	Нет	Да	Нет	Да
Отчет по страницам выхода	Да	Нет	Да	Нет	Да
Распределе-ние запросов по ОС (число распознавае-мых ОС)	Да (84)	Да (29)	Нет	Да	Да
Распреде-ление запросов по браузерам (число распознавае-мых браузеров)	Да (210)	Да (9)	Да (4)	Да	Да (20)
Отчет об ошибках HTTP	Да	Да	Да	Да	Да
Экспорт в формате pdf	Да	Нет	Нет	Да	Экспорт в форматы PDF/ HTML/ CSV

Примечания:

* – Analog C:Amie Edition является независимой веткой исходного проекта Analog

** – Необходимо установить дополнение

5. Web-аналитика для сайта ИПМ им. М.В. Келдыша РАН

5.1 Методы web-аналитики для сайтов научных организаций

Конечно, не является большим секретом то, что основное свое применение web-аналитика находит в сфере электронной торговли, для оценки эффективности web-ресурсов как средства продвижения и прямой продажи товаров и услуг. Но не стоит забывать и тот факт, что Всемирная паутина, возникнув в 1990-е как средство публикации и распространения гипертекстовой научной документации, продолжает развиваться как крупнейшая платформа для распространения и обмена научной информацией. Более того, современные интернет-технологии позволяют использовать WWW и как инструмент оценки уровня полученных научных результатов, их востребованности в профессиональном сообществе.

Целью создания сайта научной организации является информирование о ведущихся в организации научных проектах и представление имеющегося научного потенциала. Поэтому стандартные методы и критерии, применяемые в web-аналитике для оценки эффективности коммерческих сайтов, требуют в данном случае серьезного пересмотра и адаптации. Основная причина для такого пересмотра очевидна: при определении эффективности научного web-ресурса мы не отталкиваемся от цифр стоимости проданных товаров или услуг. Другая особенность сайтов научных организаций заключается в преобладающей анонимности пользователей: сайты ориентированы на свободное распространение информации и, как правило, не требуют предварительной регистрации пользователей.

Если придерживаться принятой в web-аналитике классификации интернет-ресурсов [1], то сайты научных организаций по своему предназначению и функциональности обладают свойствами, характерными для трех типов сайтов. Как уже было отмечено выше, главная задача сайта научной организации – это распространение информации о научном потенциале организации и выполняемых исследованиях. Поэтому любой сайт научной организации является в первую очередь сайтом информационным (content/media). Используя методы web-аналитики для анализа эффективности информационных сайтов, мы предполагаем получить ответ на два основных вопроса: какие материалы вызывают наибольший интерес и насколько их представление на сайте удобно пользователям. Чтобы ответить на эти вопросы, стандартные метрики – **просмотры страниц**, **уникальные посетители**, **визиты посетителей** – дополняются КПЭ, позволяющими оценить отношение посетителей к информационному наполнению сайта. К таким КПЭ относятся: **длина сессии**

(visit depth), **показатель возврата посетителей**, **показатель привлечения посетителей** и **показатель повторного обращения к странице.** Длина сессии определяет среднее число страниц, просмотренных за сессию посетителем сайта. Чем больше это среднее значение, тем дольше посетители задерживаются на страницах сайта, что является несомненным признаком их интереса к представленным на сайте материалам. В отличие от простой метрики подсчета повторных посетителей, показатель возврата посетителей вычисляется как отношение числа уникальных посетителей к общему количеству сессий и является показателем уровня лояльности посетителей сайта. Чем ниже значение этого показателя, тем больше постоянных посетителей среди аудитории сайта. Показатель привлечения пользователей определяется как отношение числа новых посетителей к общему числу уникальных посетителей и характеризует потенциал расширения аудитории сайта. Для оценки интереса пользователей к определенной странице сайта можно использовать показатель повторного обращения к странице, вычисляемый как отношение общего числа запросов страницы к числу уникальных посетителей, открывших ее.

Дополнительный интерес представляет информация, характеризующая аудиторию сайта – ее географическое распределение, распределение активности пользователей по времени суток и дням недели, используемые браузеры и операционные системы.

Хотя большинство сайтов научных организаций не требуют от своих посетителей предварительной авторизации, довольно часто на их страницах можно встретить формы регистрации пользователей – это могут быть формы регистрации участников конференций или подписка на новостные рассылки. В таком случае web-сайт выступает как сайт, предназначенный для сбора персональной информации пользователей (lead generation site). При этом одни КПЭ становятся неактуальными – например, для пользователей, целенаправленно обратившихся к сайту для ввода своих персональных данных, длина сессии всегда будет равна 1-2 страницам, и такой результат не может рассматриваться как плохой. Более важную роль начинают играть КПЭ, позволяющие оценить удобство реализации самой процедуры регистрации пользователя. Для такой оценки в web-аналитике часто используют КПЭ, называемый **коэффициентом конверсии** (conversion rate). Он будет определяться в данном случае как отношение числа всех посетителей страницы регистрации к числу посетителей, прошедших процедуру регистрации до конца.

Как правило, сайты научных организаций содержат справочные разделы. Это информация о сотрудниках организации, базы научных данных, каталоги электронных изданий. Для таких разделов важно получить оценку того, насколько быстро и удобно пользователь интернет-ресурса находит интересующую его информацию, с минимальным количеством переходов по страницам раздела. В этом случае для оценки можно использовать набор КПЭ, включающий показатели, уже рассмотренные ранее: **длина сессии**, **показатель возврата посетителей**, **показатель повторного обращения к странице**. Хотя при этом интерпретироваться данные показатели могут совершенно по-другому. Так, слишком частое обращение пользователя к одной и той же странице является скорее следствием его затруднений или непонимания интерфейса доступа к информации, а большая длина сессии свидетельствует о том, что пользователю приходится пройти слишком сложный путь в поисках нужной информации.

Достоверная картина популярности размещенных на сайте научной организации материалов должна основываться на данных о запросах, исходящих от «живой», реальной аудитории сети Интернет. Поэтому крайне важно уметь выделять и исключать из дальнейшего рассмотрения среди всех запросов, выполненных web-сервером, запросы, порожденные сетевыми приложениями, автоматически сканирующими страницы сайта. Такие сетевые приложения принято называть программами-роботами. Предварительная фильтрация запросов, исходящих от программ-роботов, необходима еще и потому, что сайты научных организаций в первую очередь представляют интерес для достаточно небольшого круга специалистов в той или иной области научных исследований. Как правило, число запросов от реальных пользователей к страницам таких сайтов относительно невелико. Любая же программа-робот, запущенная для сканирования сайтов, способна исказить статистику обращений к сайту до неузнаваемости.

Чтобы выделить автоматически генерируемые запросы программ-роботов из общего трафика, необходимо разобраться: кто и зачем использует программы-роботы. В первую очередь программы-роботы используются всеми крупнейшими поисковыми системами сети Интернет (Google, Yandex и др.) для извлечения информации со страниц сайтов и последующего индексирования в своих базах данных. Такие программы-роботы подчиняются при сканировании сайтов определенным соглашениям и явно обозначают себя в заголовке запроса к сайту. Фильтрация запросов, исходящих от программ-роботов, в этом случае не составляет большего труда. Существует еще множество подобных более

специализированных сетевых приложений, используемых для сбора коммерческой и социальной информации в интернете. То, насколько они в своей реализации соответствуют общепринятым стандартам, полностью зависит от доброй воли их разработчиков. Но программы-роботы могут использоваться и как вредоносное сетевое приложение. Один из самых распространенных сценариев использования их в этом качестве – это атака сервера сайта многочисленными запросами, ставящая своей целью вызвать отказ сервера в обслуживании дальнейших обращений (так называемая DDoS-атака). Другим примером недобросовестного использования программ-роботов является технология рассылки нежелательной рекламной информации, получившая название "referrer spam". Программы-роботы используются для автоматической генерации и рассылки http-запросов, в которых в качестве адреса страницы – источника запроса сознательно указывается URL-адрес рекламируемого сайта. Записи о таких запросах, попадая в log-файл атакованного сервера, впоследствии обрабатываются средствами web-аналитики, и ссылки на рекламируемый сайт автоматически включаются в состав отчетов как информация о внешних страницах, с которых были выполнены переходы на страницы сайта. Если же отчеты публикуются в формате web-страниц, открытых для публичного доступа, то они, наряду с остальными страницами сайта, просматриваются роботами поисковых систем. Встреченные при этом ссылки на рекламируемый сайт учитываются поисковыми системами и повышают рейтинг сайта при выдаче результатов поиска. Как правило, недобросовестные программы-роботы игнорируют правила, ограничивающие сканирование документов web-ресурса, и маскируют свою активность, фальсифицируя в заголовке http-запроса имя программы – отправителя запроса. Чтобы выявить подобные запросы в log-файле, приходится применять более сложные подходы и программные алгоритмы.

Отметим, что web-ресурсы научных организаций являются не единственными жертвами программ-роботов. С того момента как объем трафика, приходящего на сайт, превратился в товар со своей вполне конкретной стоимостью, появилось множество технологий манипулирования этой величиной. И для реализации этих технологий создаются все новые программы-роботы, увеличивающие и без того существенную долю «искусственного» трафика в сети Интернет.

5.2 Сайт ИПМ им. М.В. Келдыша РАН

Сайт Федерального исследовательского центра "Институт прикладной математики им. М.В.Келдыша Российской академии наук" является по своему

статусу некоммерческим WWW-сайтом государственного научного учреждения. Целями его создания являются:

- ознакомление мировой общественности с научными достижениями института и его историей;
- презентация научных проектов;
- публикация в электронном виде научных трудов сотрудников института;
- информирование о текущих научных событиях: семинарах, конференциях.

Официальная информация об институте и его истории представлены в следующих разделах сайта https://www.keldysh.ru.

- Мемориальная страница М.В.Келдыша.
- Официальная информация об Институте: полное наименование и устав, история создания и научные достижения.
- Страницы дирекции и ученого совета с информацией о персональном составе каждого органа.
- Направления исследований – перечень основных направлений научных исследований, выполняемых в Институте. Этот раздел включает гиперссылки на web-страницы, посвященные научным работам Института в той или иной области исследований.
- Проекты – список наиболее крупных научных проектов, реализуемых в Институте, в том числе проектов, выполняемых по грантам РФФИ, РНФ и Минобрнауки. Включает список ссылок на web-страницы, посвященные отдельным проектам.
- Диссертационные советы – информация о работающих в Институте диссертационных советах: их специализация, персональный состав и контактные данные. Приводится список прошедших и планируемых защит; в свободном доступе опубликованы все представленные к каждой защите материалы.
- Аспирантура – содержит информацию о научных специальностях, по которым проводится обучение, правила приема и сдачи экзаменов, программы учебных курсов. Публикуется текущая информация для аспирантов.
- Научная библиотека содержит информацию о периодических изданиях, учредителем которых является Институт, ссылки на каталог электронной онлайновой библиотеки препринтов и трудов сотрудников Института и каталог традиционной научной библиотеки печатных изданий.

- Страницы памяти посвящены выдающимся деятелям советской и российской науки, в разное время работавшим в стенах Института.

Все эти разделы сайта разрабатываются централизованно и поддерживаются группой сопровождения сайта.

Разделы сайта с официальной информацией дополняются разделами, посвященными конкретным научным проектам и направлениям исследований. Данные разделы представляют собой фактически независимые мини-сайты, со своим собственным дизайном и структурой; они создаются и поддерживаются самими участниками этих работ.

Таким образом, наполнение сайта можно условно разделить на три составляющие:

- разделы сайта, содержащие официальную информацию об институте, сотрудниках и руководстве института;
- электронная онлайновая библиотека научных работ сотрудников, включая полнотекстовую коллекцию препринтов ИПМ РАН, начиная с 2001 года;
- разделы сайта, посвященные отдельным научным проектам и направлениям исследований.

Web-сайт ИПМ РАН реализован на основе http-сервера Apache версии 2.2, работающего под управлением ОС Linux.

5.3 Выбор программного решения для системы web-статистики

Приведенная выше краткая характеристика сайта ИПМ РАН определяет те особенности, которые приходится учитывать при выборе программного обеспечения для построения системы web-аналитики сайта.

В первую очередь необходимо считаться с тем фактом, что разделы сайта ИПМ РАН создаются и развиваются децентрализовано, усилиями независимых групп разработчиков. Это затрудняет использование средств web-аналитики, требующих для сбора первичных данных включения специального программного кода в состав каждой страницы web-сайта. Расширяется и круг потенциальных потребителей данных web-аналитики. Возникает проблема не только организации совместного доступа к создаваемым отчетам, но и необходимость настройки средств web-аналитики с учетом потребностей различных пользователей.

Любой анализ востребованности интернет-ресурса средствами web-аналитики будет неполным, если мы ничего не знаем о происхождении запросов. Для каждого запроса нужно уметь определять IP-адрес, с которого был отправлен запрос; характеристики компьютера и программы-клиента,

которые были использованы для создания запроса; и многое другое. Без этого невозможно достоверно оценить популярность размещенных на страницах ресурса материалов, предварительно исключив из рассмотрения запросы от программ-роботов. Невозможно и получить представление об аудитории интернет-ресурса: географическое расположение, используемые типы компьютеров и браузеры. Сам http-запрос, к сожалению, содержит минимум информации об источнике запроса – это его IP-адрес и строка описания клиентского приложения. Поэтому для более полного и достоверного определения происхождения каждого запроса приходится дополнять данные log-файла информацией из внешних источников: справочников распределения IP-адресов по доменам, баз данных идентификаторов, используемых для обозначения клиентского приложения в http-запросе. Информация, представленная в таких справочниках и базах данных, достаточно быстро меняется и нуждается в постоянном обновлении. Способность достаточно полно определять свойства источника запроса на основе данных из log-файла, дополненных актуальной информацией из внешних источников, является одной из ключевых характеристик, рассматриваемых при выборе программного решения web-аналитики.

Учитывая все вышесказанное, представляется обоснованным остановить свой выбор на одном из тех программных решений web-аналитики, которое удовлетворяло бы следующему набору требований:

- использует в качестве первичного источника данных log-файл сервера и не требует для сбора исходных данных редактирования текста страниц сайта;
- реализовано для ОС Linux и совместимо с http-сервером Apache;
- создает основные виды отчетов для стандартных метрик и параметров;
- содержит средства фильтрации трафика, источником которого являются программы-роботы;
- поддерживает вывод отчетов в формате html-страниц;
- допускает возможность запуска как web-приложение;
- является некоммерческим, свободно-распространяемым продуктом;
- поддерживается своими разработчиками в актуальном состоянии.

5.4 Анализатор log-файлов AWStats – краткая характеристика

В качестве основы для создания системы web-аналитики сайта ИПМ РАН был выбран пакет AWStats [7] как готовое программное решение для построения отчетов web-статистики на основе разбора и анализа log-файла http-

сервера Apache. Созданный еще в 2000 году энтузиастом разработки свободного программного обеспечения Лораном Дестайер (Laurent Destailleur), AWStats за прошедшие 20 лет превратился в широко используемый программный пакет, включаемый в состав многих дистрибутивов ОС Linux. Написанный на языке Perl, AWStats предназначен для решения широкого круга задач – создания статистических отчетов на основе анализа log-файлов серверов наиболее популярных интернет-служб: web, ftp и электронной почты. Но эта универсальность не сужает возможности его применения как инструмента web-аналитики; список основных видов отчетов, которые можно получить для интернет-ресурса с помощью AWStats, включает:

- число сессий и уникальных посетителей;
- продолжительность сессий;
- распределение запросов по дням недели и времени суток;
- распределение источников запросов по национальным доменам;
- распределение запросов по страницам сайта;
- источники запросов (hosts);
- используемые браузеры и ОС;
- запросы роботов;
- распределение запросов по кодам завершения, включая коды ошибок.

Для построения отчетов AWStats может быть запущен на выполнение, используя интерфейс командной строки, или вызван непосредственно со страниц сайта как CGI-скрипт. В последнем случае отчеты могут создаваться в реальном времени по текущему содержимому log-файла.

Чтобы сократить время построения отчетов по данным, содержащимся в log-файле http-сервера, эта процедура выполняется за два шага. Сначала на основе записей log-файла создаются промежуточные базы данных, содержащие данные web-статистики в собственном внутреннем представлении. Затем на основе этих баз данных уже создается представление отчетов в формате html-страниц. Предусмотрена возможность инкрементального обновления баз данных для включения информации о новых записях, появившихся в log-файле.

Параметры запуска и настройки AWStats определяются одним или несколькими конфигурационными файлами. Конфигурационные файлы являются текстовыми файлами и включают директивы определения параметров конфигурации в привычном для среды Unix формате: имя параметра и его значение. Таких конфигурационных файлов для каждой установки AWStats может быть создано сколь угодно много; каждый из них определяет свой, независимый, набор параметров для использования программы. Директивы

конфигурации разбиты на несколько групп. Основная группа включает директивы, устанавливающие значения для параметров, без определения которых запуск AWStats невозможен: это имя и расположение log-файла, его формат, расположение рабочих каталогов программы. Остальные директивы являются необязательными и используются для фильтрации исходных данных, задания набора генерируемых отчетов, определения уровня детализации аналитических данных в отчетах и управления внешним представлением отчетов при их выводе в html-формате. Отдельная группа директив дает пользователю AWStats возможности, хотя и несколько ограниченные, для создания собственных видов отчетов. Используя директивы этой группы, пользователь для своего отчета может:

- Задать правила отбора исходных данных, включаемых в отчет – допускается создание правил фильтрации по значению полей записи в log-файле.
- Определить одно из полей записи log-файла в качестве параметра, используемого в отчете. Значения параметра определяются как результат применения заданного регулярного выражения к содержимому поля.
- Определить, какие из стандартных метрик будут вычисляться для каждого значения параметра.
- Допускается вычисление среднего значения или общей суммы для используемых метрик. Суммарное или среднее значение метрики выводится отдельной строкой в конце отчета.

Функциональность AWStats может быть расширена за счет использования дополнительных модулей, создаваемых сторонними разработчиками. Примером такого модуля является модуль GeoIPfree, позволяющий на основе информации из справочной базы распределения IP-адресов по национальным доменам более точно привязать адрес источника запроса к его географическому местоположению.

5.5 Система web-статистики сайта ИПМ РАН

Система web-статистики сайта ИПМ РАН использует AWStats как основное средство сбора, анализа и представления статистических данных. Полученные отчеты представляют посещаемость сайта в двух основных проекциях:

- статистические данные о посещаемости за текущие сутки, получаемые в реальном времени;
- статистические данные о посещаемости сайта в текущем году с разделением по месяцам и дням текущего месяца.

Дополнительно сохраняются архивные отчеты о посещаемости сайта за прошедшие года, начиная с 2015 года.

Для реализации подобной схемы сбора, анализа и представления данных о посещаемости сайта исходный программный пакет AWStats был дополнен скриптами обновления баз статистических данных и построения новых отчетов, запускаемых автоматически по расписанию, раз в сутки. Архивация отчетов прошедшего года также выполняется в автоматическом режиме.

5.5.1 Статистика посещаемости за текущие сутки

Отчеты о посещаемости сайта за текущие сутки создаются в режиме реального времени при переходе по ссылке «Сегодня» раздела «Статистика доступа к web-серверу» сайта ИПМ РАН.

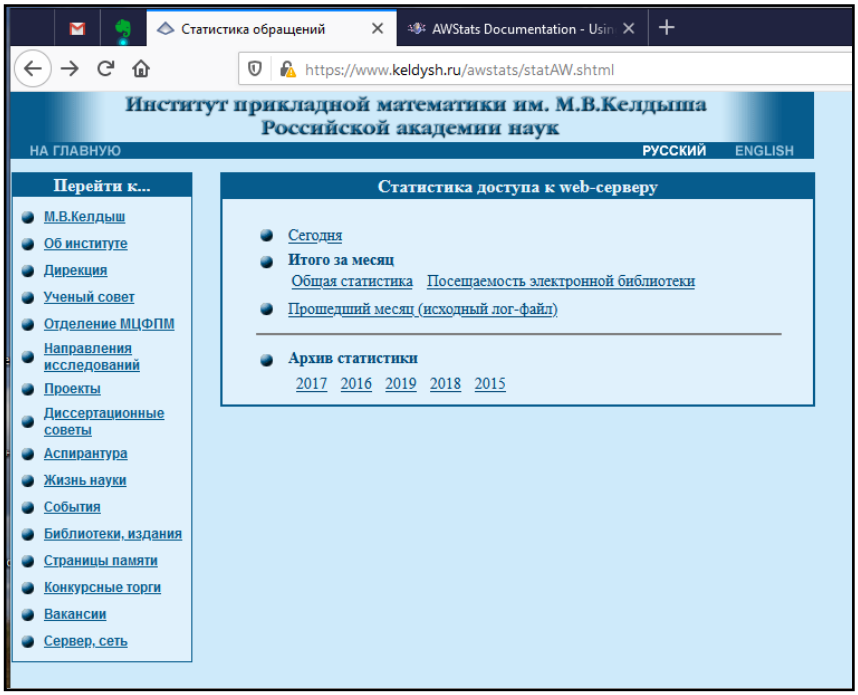


Рис. 1. Страница раздела «Статистика доступа к web-серверу»

При этом происходит обращение к программе AWStats для одновременного обновления собственной базы статистических данных и построения отчетов, представляемых пользователю в формате html-страниц. Полученные данные отражают статистику посещаемости сайта с начала текущих суток до момента запуска AWStats при переходе по ссылке. После завершения работы программы пользователю становится доступен для просмотра и анализа набор отчетов, полный перечень которых приведен ниже.

Общая статистика посещений сайта за текущие сутки. Отчет содержит общую статистику обращений к сайту за текущие сутки и включает следующие

метрики: общее число запросов и число просмотров страниц сайта, число уникальных посетителей и число сессий, общий объем переданного трафика. Эта статистика приводится раздельно для запросов, исходящих от реальных посетителей сайта, и запросов, сгенерированных прочими источниками (программами-роботами, вирусами). В отчетах, рассматриваемых далее, учитываются только запросы, исходящие от реальных посетителей.

Распределение запросов к сайту по времени суток. Представлено распределение числа запросов, числа просмотров страниц сайта и объем переданного трафика по часовым интервалам текущих суток (от 0 часов до 23).

Распределение запросов к сайту по национальным доменам. Показывает распределение по национальным доменам общего числа запросов, числа запросов страниц сайта и объем переданного трафика. Для более точного определения принадлежности IP-адреса национальному домену используется дополнительный модуль GeoIPfree и внешняя база данных.

Распределение запросов к сайту по источникам трафика. Для каждого источника трафика определяется общее число запросов, число запросов страниц сайта, объем переданного трафика и время последнего обращения. Чтобы сократить время генерации отчетов, преобразование IP-адресов источников запросов в их доменные имена не выполняется. Отчет представлен в двух версиях: краткой и полной. В краткой версии приводится статистика для десяти наиболее активных источников трафика, полная содержит данные для всех источников. Полная версия отчета содержит возможность динамической фильтрации IP-адресов источников запросов, для задания ограничений используется синтаксис регулярных выражений.

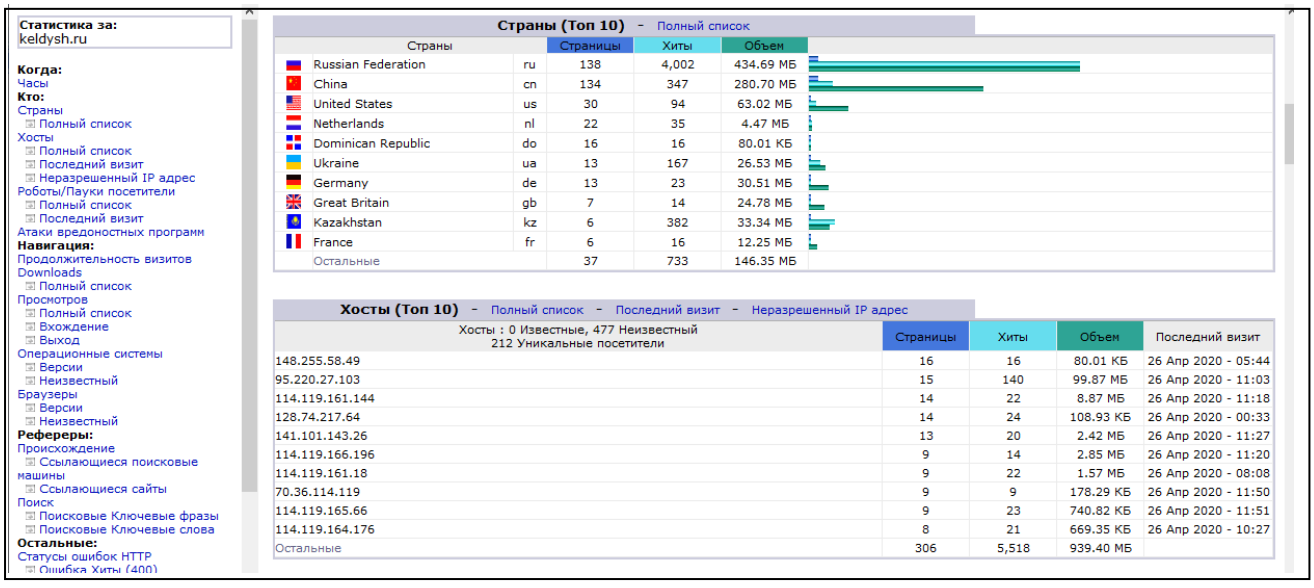


Рис. 2. Отчеты о распределении запросов по странам и источникам трафика

Распределение запросов к сайту по программам-роботам. Это единственный отчет, отражающий активность обращений к сайту программ-роботов. Для каждой программы-робота определяется число запросов, объем переданного трафика и время последнего обращения. Идентификация программ-роботов основана только на сопоставлении имени программы – источника запроса со списком регулярных выражений, представляющих все возможные варианты именования клиентских приложений программ-роботов, используемые в http-запросе. Доступны две версии отчета: краткая и полная.

Продолжительность визитов. Приводится распределение количества сессий по их продолжительности.

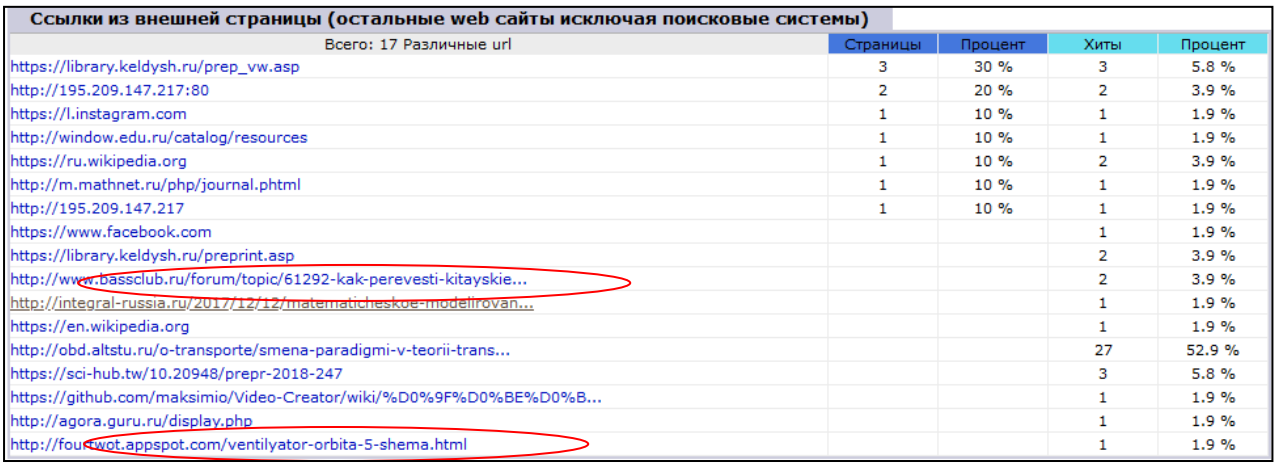
Загружаемые файлы. Показано распределение числа запросов на загрузку среди доступных для загрузки файлов на сайте. Для каждого файла, к которому обращались хотя бы один раз за текущие сутки, приводятся общее число запросов на загрузку, число запросов на частичную загрузку, общий объем переданных данных и объем данных, приходящийся на один запрос в среднем. Имеются полная и краткая версии отчета.

Распределение запросов по страницам сайта. Приводится распределение запросов по страницам сайта. Для тех страниц сайта, к которым обращались хотя бы один раз за рассматриваемый период времени, вычисляется общее число ее просмотров и средний объем переданных при этом данных. Дополнительно определяется число визитов, для которых данная страница была начальной при посещении сайта, и число визитов, для которых на данной странице завершился просмотр сайта. Полная версия отчета дает возможность устанавливать правила включения или исключения страниц из отчета по значению URL-адреса страницы.

Распределение запросов по семействам ОС. Распределение запросов по семействам браузеров. Отчеты характеризуют программное обеспечение, используемое для обращения к сайту. Для каждого семейства ОС и браузеров определяется число выполненных запросов и число просмотров страниц, вычисляется процентная доля этих запросов в общем числе запросов. Полные версии отчетов содержат детализацию этих данных с учетом различных версий программного обеспечения каждого семейства.

Происхождение запросов. Представляет распределение запросов по URL ссылающихся страниц – web-страниц, с которых был сделан переход на страницы сайта. Для каждого URL определяется число запросов и число просмотров страниц сайта, вычисляется процентная доля этих запросов в общем числе запросов. Эта статистика формируется раздельно для страниц

поисковых систем и страниц остальных сайтов. Именно отчеты этого типа используются для распространения рекламных ссылок по технологии "referrer spam".



Ссылки из внешней страницы (остальные web сайты исключая поисковые системы)

Всего: 17 Различные url	Страницы	Процент	Хиты	Процент
https://library.keldysh.ru/prep_vw.asp	3	30 %	3	5.8 %
http://195.209.147.217:80	2	20 %	2	3.9 %
https://l.instagram.com	1	10 %	1	1.9 %
http://window.edu.ru/catalog/resources	1	10 %	1	1.9 %
https://ru.wikipedia.org	1	10 %	2	3.9 %
http://m.mathnet.ru/php/journal.phtml	1	10 %	1	1.9 %
http://195.209.147.217	1	10 %	1	1.9 %
https://www.facebook.com			1	1.9 %
https://library.keldysh.ru/preprint.asp			2	3.9 %
http://www.bassclub.ru/forum/topic/61292-kak-perevesti-kitayskie...			2	3.9 %
http://integral-russia.ru/2017/12/12/matematicheskoe-modelirovan...			1	1.9 %
https://en.wikipedia.org			1	1.9 %
http://obd.altstu.ru/o-transporte/smena-paradigmi-v-teorii-trans...			27	52.9 %
https://sci-hub.tw/10.20948/prepr-2018-247			3	5.8 %
https://github.com/maksimio/Video-Creator/wiki/%D0%9F%D0%BE%D0%B...			1	1.9 %
http://agora.guru.ru/display.php			1	1.9 %
http://fourtwot.appspot.com/ventilyator-orbita-5-shema.html			1	1.9 %

Рис. 3. Фрагмент отчета "Ссылки из внешней страницы".

Красным цветом выделены URL, попавшие в отчет как результат технологии "referrer spam".

Ключевые поисковые фразы. Ключевые поисковые слова. Приводится список ключевых слов и ключевых фраз, которые наиболее часто использовались в поисковых запросах, приведших пользователей на страницы сайта. Для каждого ключевого слова или фразы указывается число поисковых запросов, в которых они встречались, и процентная доля таких запросов в общем числе поисковых запросов.

Коды ошибочного завершения запросов к сайту. Определяется распределение http-запросов по кодам ошибочного завершения. Для каждого кода ошибочного завершения приводятся: общее число запросов, завершившихся с данным кодом, процентная доля таких запросов в общем числе запросов, объем трафика, переданного при выполнении этих запросов.

Статистика доступа к электронной библиотеке. Специально созданный отчет, показывающий распределение обращений к полным текстам публикаций в электронной библиотеке ИПМ РАН. В отчет включаются данные о запросах файлов с полными текстами препринтов. Для каждого такого файла определяется: суммарное число запросов, процентная доля таких запросов в общем количестве запросов к препринтам и дата последнего обращения к файлу.

5.5.2 Статистика посещаемости в текущем году

Практика показала, что процесс формирования отчетов web-статистики для сайта ИПМ РАН за большие интервалы времени (от нескольких дней) занимает достаточно продолжительное время. Поэтому построение отчетов о посещаемости сайта за все прошедшие дни текущего года выполняется в фоновом режиме, и программа AWStasts запускается для этого автоматически, по заранее заданному расписанию, раз в сутки. Набор формируемых при этом отчетов включает все типы отчетов, создаваемых в режиме реального времени для текущего дня месяца, и дополнительно содержит еще два вида отчета:

Распределение запросов по дням месяца. Представляет распределение запросов к сайту по дням месяца. Для каждого дня месяца определяется число обработанных запросов и число просмотров страниц, количество состоявшихся сессий пользователей и объем переданных данных.

Распределение запросов по дням недели. Представляет распределение запросов по дням недели (понедельник-воскресенье) месяца. Для каждого дня недели определяется: суммарное число обработанных запросов и просмотров страниц, общий объем переданных данных.

Отчеты создаются на основании данных обо всех запросах, обработанных сервером сайта с 1 января текущего года до полуночи прошедших суток. По умолчанию отчеты представляют информацию о посещаемости сайта за текущий месяц. Используя web-интерфейс страницы отчетов, этот временной интервал можно изменить для просмотра статистики посещаемости сайта за любой день текущего месяца или прошедший месяц года.

5.5.3 Статистика посещаемости электронной библиотеки

Отдельная группа отчетов отражает статистику обращений пользователей за текущий год к полным текстам публикаций, хранящимся в онлайновой электронной библиотеке ИПМ РАН. Данная группа отчетов по своему составу полностью совпадает с отчетами, отражающими посещаемость сайта в текущем году, но учитывает только данные о запросах пользователями файлов с полными текстами электронных публикаций. Отчеты о посещаемости электронной библиотеки также формируются раз в сутки, в фоновом режиме.

5.5.4 Архив отчетов посещаемости сайта за прошлые года

По завершении календарного года все отчеты о посещаемости сайта в прошедшем году удаляются. Сохраняется только общий отчет с итоговыми данными о посещаемости сайта за год, без возможности детализации по дням и месяцам прошедшего года. Архивные отчеты за 2015-19 годы доступны на

сайте ИПМ РАН в виде статических html-страниц в разделе «Статистика доступа к web-серверу» (рис. 1).

6. Оценка востребованности публикаций электронной библиотеки

6.1 Выбор критерия востребованности публикаций

Отчеты, создаваемые системой web-статистики сайта ИПМ РАН, адресованы все-таки в первую очередь тем, кто занят развитием и сопровождением сайта. Но для авторов публикаций, размещенных в онлайновой библиотеке, наибольший интерес представляют данные о востребованности посетителями сайта их собственных произведений. Основным показателем такой востребованности является, безусловно, число запросов полного текста электронной публикации, при условии, что эти обращения исходят от реальных читателей онлайновой библиотеки. Такая статистика может быть в простой и сжатой форме включена непосредственно в электронную карточку самой публикации в онлайновой библиотеке.

Сама по себе задача подсчета числа обращений к текстам публикаций на основании записей о http-запросах в log-файле web-сервера в общем проста и не требует подробного обсуждения. Но такой подсчет будет бесполезным без предварительной фильтрации запросов от программ-роботов – ведь автора публикации в первую очередь интересуют «живые» читатели его произведения.

И если при создании общей системы web-статистики сайта кажется более рациональным использовать как основу готовое программное решение, то данная задача требует собственной реализации как процедуры получения и обработки статистических данных, так и способа представления результатов.

6.2 Задача фильтрации запросов роботов в log-файле web-сервера

Необходимость предварительной фильтрации запросов, исходящих от программ-роботов, при применении методов web-аналитики уже достаточно подробно обсуждалась ранее. Особенно это актуально для публикаций электронной библиотеки: как показывает практика, число запросов за сутки к тексту научной публикации редко превышает два-три десятка, при этом согласно статистике, приведенной в [9], около трети из них может исходить от программ-роботов.

Если мы используем как источник данных web-статистики записи log-файла сервера, то фильтрация запросов от программ-роботов выполняется на этапе первичного разбора записей log-файла. Используемые для фильтрации методики очень разнообразны и могут варьироваться от простой проверки полей запроса на совпадение со списком заранее заданных строк до сложных

аналитических алгоритмов, использующих вероятностные модели и машинное обучение. Классификация таких методик подробно рассмотрена в [8].

Самый очевидный способ обнаружения запросов, принадлежащих программам-роботам, основан на простом сравнении отдельных полей записи log-файла с заранее известными списками ключевых значений. Так мы можем обнаружить запрос от программы-робота по идентификатору программного приложения, указанному в строке описания источника запроса. Или сравнить IP-адрес источника запроса с базой данных адресов, активно используемых программами-роботами. Такой способ обнаружения запросов от программ-роботов прост и однозначен, но его применение основано на двух существенных допущениях:

- поля запроса содержат достоверную информацию;
- заранее известен набор ключевых значений, который можно использовать для определения программы-робота по содержимому полей записи log-файла.

Отметим также, что отсутствие или недопустимое значение для некоторых полей записи log-файла также могут рассматриваться как признак принадлежности запроса программе-роботу. Так, «анонимные» запросы с неопределенным полем описания источника запроса, как правило, исходят от программ-роботов, нарушающих сетевые протоколы и соглашения.

Другие методики обнаружения запросов от программ-роботов основаны на анализе общих характеристик последовательности запросов, относящихся к одному визиту пользователя. К числу таких характеристик относятся: распределение запросов по используемым методам запроса и типам запрашиваемых web-ресурсов, интенсивность запросов в единицу времени и объем переданных при этом данных. Процедура выделения запросов от программ-роботов распадается на два этапа: сначала для всех запросов одного визита вычисляются необходимые метрики, затем их значения используются для классификации источника запроса. Причем такая классификация основана на критериях, которые чаще всего определяются эмпирическим путем: известно, например, что программы-роботы во время просмотра сайта чаще запрашивают только атрибуты документов и, как правило, не загружают файлы с изображениями, размещенными на web-странице. Окончательное решение о характере источника запросов принимается при условии выполнения нескольких таких критериев одновременно. Необходимо подчеркнуть, что принятое решение верно только для рассмотренной последовательности запросов. Механизм динамического выделения IP-адресов не позволяет

распространить сделанный выбор на все запросы, источником которых является данный IP-адрес.

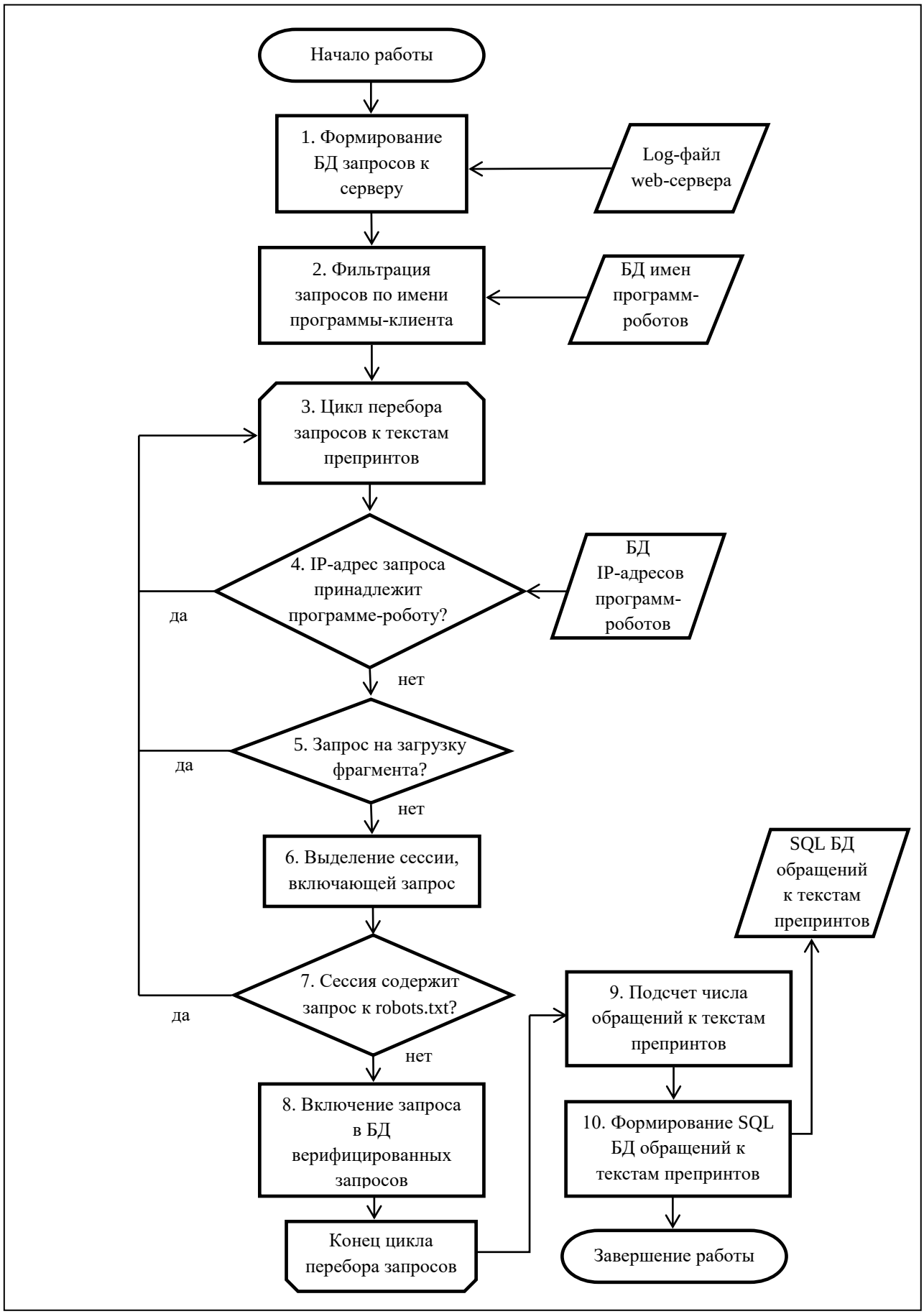
Развитием методов предыдущей группы являются алгоритмы, основанные на машинном обучении и применении вероятностных моделей. Как и в предыдущем подходе, на первом этапе определяются последовательности запросов, относящиеся к одному визиту пользователя, и для каждого отдельного визита вычисляется набор характеристик. Затем, используя заранее построенную вероятностную модель, принимается решение о характере источника запросов каждого визита («живой» посетитель или программа-робот). Мы не будем останавливаться подробно на методах этой группы, так как они не используются в рассматриваемой реализации счетчика обращений.

6.3 Счетчик числа обращений к текстам публикаций – особенности реализации

6.3.1 Краткая характеристика программы

Процедура подсчета числа обращений к полным текстам публикаций электронной библиотеки реализована как независимая программная утилита, ее текущая версия учитывает обращения пользователей только к полным текстам препринтов ИПМ РАН. Для фильтрации запросов применяются как методы, использующие синтаксический анализ полей запроса, так и методы, основанные на анализе характеристик визитов пользователей. Программа не содержит средств формирования статистических отчетов, а только передает вычисленные метрики для последующей обработки программному обеспечению электронной библиотеки. После обработки данные о числе запросов полного текста препринта представляются непосредственно в карточке самого препринта, размещенного в онлайновой библиотеке. Программа реализована на языке Ruby и использует дополнительные библиотеки для разбора записей log-файла web-сервера, доступа к базам данных и внешним источникам данных в формате JSON.

6.3.2 Общая блок-схема алгоритма



6.3.3 Описание работы программы

Рассмотрим шаги выполнения алгоритма программы в той последовательности, в которой они приведены на общей блок-схеме.

Шаг 1. Формирование базы данных (БД) запросов к web-серверу. Выполняется последовательное считывание всех записей, содержащихся в log-файле web-сервера. Для каждой записи log-файла создается запись БД, содержащая следующую информацию об исходном http-запросе:

- IP-адрес запроса;
- тип http-запроса;
- время создания запроса;
- код завершения запроса;
- URL-адрес запрашиваемого ресурса;
- строка описания клиентского приложения, создавшего запрос, включая идентификацию программы-клиента.

Для хранения информации об обрабатываемых запросах используется СУБД MongoDB – документо-ориентированная, нереляционная СУБД, позволяющая хранить в одной базе данных документы различной структуры. Программирование запросов к объектам, хранимым в базе данных, во многом облегчается применением Mongoid, инструментария для языка Ruby, реализующего объектно-ориентированный интерфейс доступа к СУБД MongoDB.

Шаг 2. Фильтрация запросов по имени программного приложения. Записи о запросах, размещенные в БД на предыдущем шаге, проверяются на совпадение имени приложения – источника запроса – со списком имен приложений, используемых программами-роботами. Имена приложений программ-роботов хранятся в отдельной базе данных, которая формируется на основе свободно распространяемого списка регулярных выражений, представленного в формате JSON [10]. Каждое регулярное выражение определяет все вариации включения и написания имени программы-робота в строке описания источника http-запроса. Список содержит более 400 регулярных выражений, соответствующих идентификаторам наиболее распространенных программ-роботов, и периодически обновляется. Запросы, которые определяются как исходящие от программы-робота по имени программы-клиента, из дальнейшего рассмотрения исключаются.

Шаг 3. Среди запросов, прошедших проверку на предыдущем шаге, для дальнейшей обработки выбираются успешные http-запросы к файлам, содержащим тексты препринтов. Отобрать такие запросы несложно, поскольку

имена файлов с полными текстами препринтов формируются единообразно, в соответствии с соглашениями, принятыми в онлайновой библиотеке ИПМ РАН.

Шаг 4. Проверка IP-адреса источника запроса. Для каждого запроса к тексту препринта проверяется принадлежность IP-адреса источника к подмножествам адресов, которые используются программами-роботами. Диапазоны адресов хранятся в отдельной базе данных, формируемой на основе текстового списка, получаемого из открытого источника [11].

Шаг 5. Проверка на запрос фрагмента файла. Спецификация http-протокола позволяет сформировать запрос для передачи только части документа, указав в заголовке запроса диапазон байт запрашиваемого фрагмента. Таким образом можно разбить загрузку большого по объему файла на несколько частей и использовать для загрузки каждой части свой запрос. Этим свойством http-протокола широко пользуются современные браузеры для оптимизации и распараллеливания процесса загрузки. Но с точки зрения web-статистики последовательность запросов на загрузку фрагментов файла должна расцениваться как одно обращение пользователя к документу. Четких формальных критериев для выделения такой последовательности запросов нет – каждый браузер реализует собственный алгоритм разбиения загрузки большого документа на несколько запросов. Поэтому одно из решений этой проблемы основывается на допущении, что все запросы на полную или частичную загрузку файла, исходящие от определенного IP-адреса в течение достаточно малого интервала времени, являются следствием одного обращения пользователя к документу. Из выделенной таким образом последовательности запросов далее рассматривается только первый по времени запрос, остальные игнорируются.

Шаг 6. Выделение сессии запросов. На следующем этапе отбора запросов рассматриваются общие характеристики сессии пользователя, включающей текущий запрос к тексту препринта. Для этого необходимо предварительно выделить все запросы, относящиеся к данной сессии. При этом считается, что в одну сессию с текущим запросом входят все запросы, которые удовлетворяют следующим условиям:

- отправлены с того же IP-адреса, что и текущий запрос;
- созданы одним и тем же клиентским приложением (строки описания источника запроса совпадают);
- временной интервал между последовательными запросами в сессии не превышает заранее определенного порогового значения.

Шаг 7. Проверка на загрузку файла robots.txt во время сессии. Файл robots.txt был разработан консорциумом W3C для управления поведением программ-роботов при обходе сайта. Этот текстовый файл, размещаемый в корневом каталоге сайта, содержит инструкции, которые регламентируют доступ программам-роботам к отдельным файлам или каталогам сайта. Следование инструкциям, приведенным в файле robots.txt, является для программ-роботов добровольным, однако большинство программ-роботов поисковых систем придерживается их и загружает файл robots.txt в начале каждого сеанса просмотра сайта. Таким образом, факт загрузки файла robots.txt во время сеанса является явным признаком того, что все запросы сессии исходят от программы-робота и их можно исключить из дальнейшего рассмотрения.

Шаг 8. Запросы, прошедшие описанные выше этапы отбора, рассматриваются как верифицированные запросы к текстам препринтов от реальных посетителей сайта.

Шаги 9-10. Подсчитав количество верифицированных запросов, относящихся к URL файлов с текстами препринтов, мы получаем, таким образом, общее число обращений посетителей к полному тексту каждого препринта. Эти данные заносятся в служебную базу данных SQL-сервера электронной библиотеки. Данные о числе обращений для каждого препринта представлены в базе данных записями следующего формата:

- дата обращения;
- URL файла полного текста препринта;
- число обращений к файлу.

Сервер баз данных электронной библиотеки реализован на платформе Microsoft SQL Server; для доступа к служебной базе данных используется объектно-ориентированный интерфейс доступа к SQL-базам данных, реализованный в пакете Sequel.

6.3.4 Представление статистики обращений

Записанные в служебную SQL базу данных результаты работы программы затем обрабатываются средствами программного обеспечения электронной библиотеки и используются для актуализации статистики обращений в основной базе данных электронных изданий. Суммарная статистика обращений к препринтам ИПМ представляется на карточках электронных изданий в следующем формате (см. рис.4):

- общее число обращений от дня публикации препринта или от дня запуска программы подсчета обращений к препринтам (09.01.2020);

- число обращений за последние 30 дней;
- разница в числе обращений по сравнению с предыдущим 30-дневным интервалом сбора статистики.

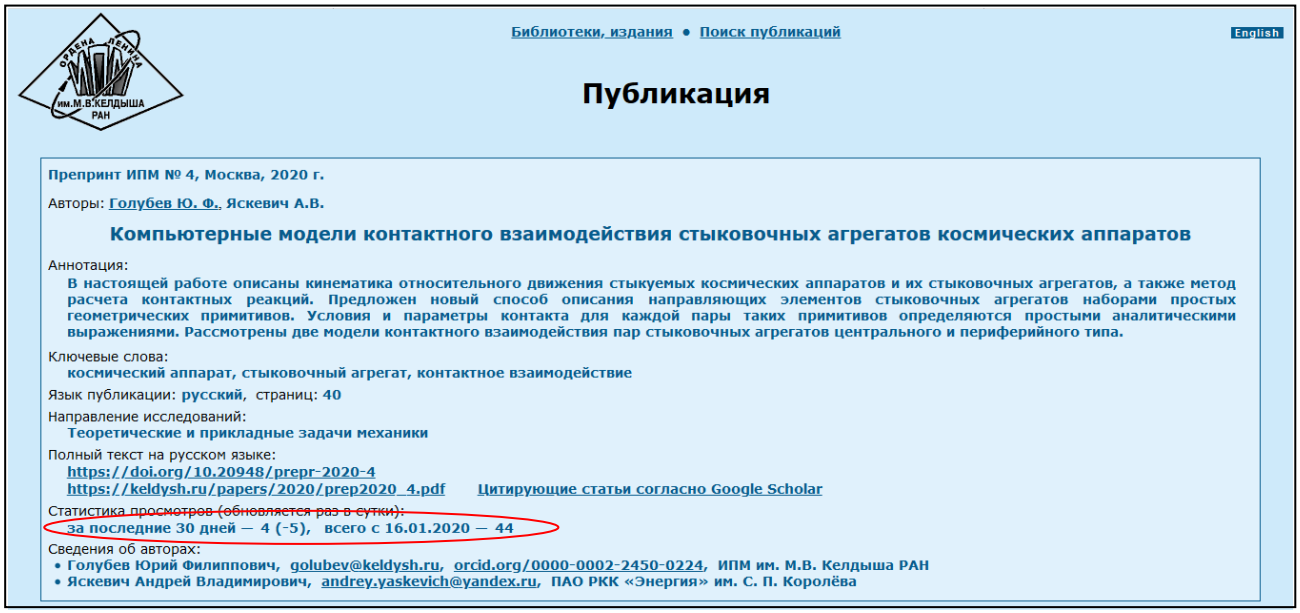


Рис. 4. Карточка электронного издания с представленной информацией о статистике обращений (выделена красным цветом)

Программа подсчета числа обращений к текстам препринтов запускается автоматически в фоновом режиме раз в сутки. Соответственно, раз в сутки обновляется и информация о числе обращений, представленная на карточке электронного издания.

7. Заключение

Цель настоящей работы – дать представление о том, как стандартные методы web-аналитики могут быть использованы для оценки эффективности сайтов научных организаций как инструмента распространения научной информации. Успех применения средств web-аналитики целиком зависит от того, насколько правильно выбраны критерии оценки реализации сайта и его использования. Некоторые из стандартных критериев web-аналитики могут быть напрямую использованы для анализа представления информации на научных сайтах, применение других требует их адаптации и переосмысления. Используя средства web-аналитики, можно не только определить показатели загруженности сайта запросами, но и, что значительно более важно, оценить эффективность использования сайта научной организации как средства

распространения научной информации, наметить пути его развития и совершенствования.

Список литературы

1. Danielle Booth, Bernard J. Jansen. A review of methodologies for analyzing websites//Handbook of Research on Web Log Analysis, IGI Global, 2009, pp.143-164. https://doi.org/10.4018/9781605669823.ch009

2. Guangzhi Zheng, Svetlana Peltsverger. Web Analytics Overview //Encyclopedia of Information Science and Technology, Third Edition, IGI Global, 2015. https://doi.org/10.4018/978-1-4666-5888-2.ch756

3. Авинаш Кошик. Веб-аналитика. Анализ информации о посетителях веб-сайтов. М.: Издательский дом "Вильямс", 2009

4. Trevor Collins. Evaluating the impact of research online with Google Analytics.//PERO Project Working Papers.

URL: https://warwick.ac.uk/fac/soc/sociology/staff/jensen/ericjensen/pero/pero-google-analytics-guide-v4.pdf

5. Jody Condit Fagan. The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment // The Journal of Academic Librarianship, 2014, v.40, pp. 25–34

6. Log analyzers Comparisons // AWStats logfile analyzer 7.8 Documentation URL: https://awstats.sourceforge.io/docs/awstats_compare.html

7. AWStats official web site. URL: https://awstats.sourceforge.io

8. Derek Doran, Swapna S. Gokhale. Web robot detection techniques: overview and limitations// Data Mining and Knowledge Discovery, June, 2011. https://doi.org/10.1007/s10618-010-0180-z

9. Huntington, P., Nicholas, D., & Jamali, H.R. Web robot detection in the scholarly information environment // Journal of Information Science, 34(5), 2008б pp.726–741

10. Syntactic patterns of HTTP user-agents used by bots / robots / crawlers / scrapers / spiders. URL: https://github.com/monperrus/crawler-user-agents

11. Free IP2Location Firewall List by Search Engine. URL: https://www.ip2location.com/free/robot-whitelist