Система «Statistics AS»

Сапожников И.В.
Научный руководитель: к.п.н., доцент Слинкин Д.А.
ФГБОУ ВПО «Шадринский государственный педагогический институт»,
г. Шадринск

На сегодняшний день очень сложно представить сайт без собственной статистики, включающей в себя не только посещения пользователем страниц сайта, но и данные о его перемещениях по сайту, об операционной системе, веб-браузере, доменных зонах и странах, количество уникальных посетителей и другие функции.

Интернет-статистика - это набор сервисов, позволяющих собирать и анализировать различную информацию о посетителях сайта. Вся информация выбирается из данных, предоставляемых сетевыми протоколами. Для получения переменных протокола из запроса пользователя требуется, чтобы вместе со страницей сайта выполнился серверный скрипт системы интернет-статистики, в котором обрабатывается и формируется информация о посещениях.

Статистика посещений сайта считается основным показателем популярности сайта. Она помогает анализировать результаты посещений разделов сайта и определять менее востребованные части тем самым модернизировать их содержимое или отказаться от размещения данной информации. Дает возможность создавать отчеты, содержащие актуальные данные о необходимости того или иного раздела сайта.

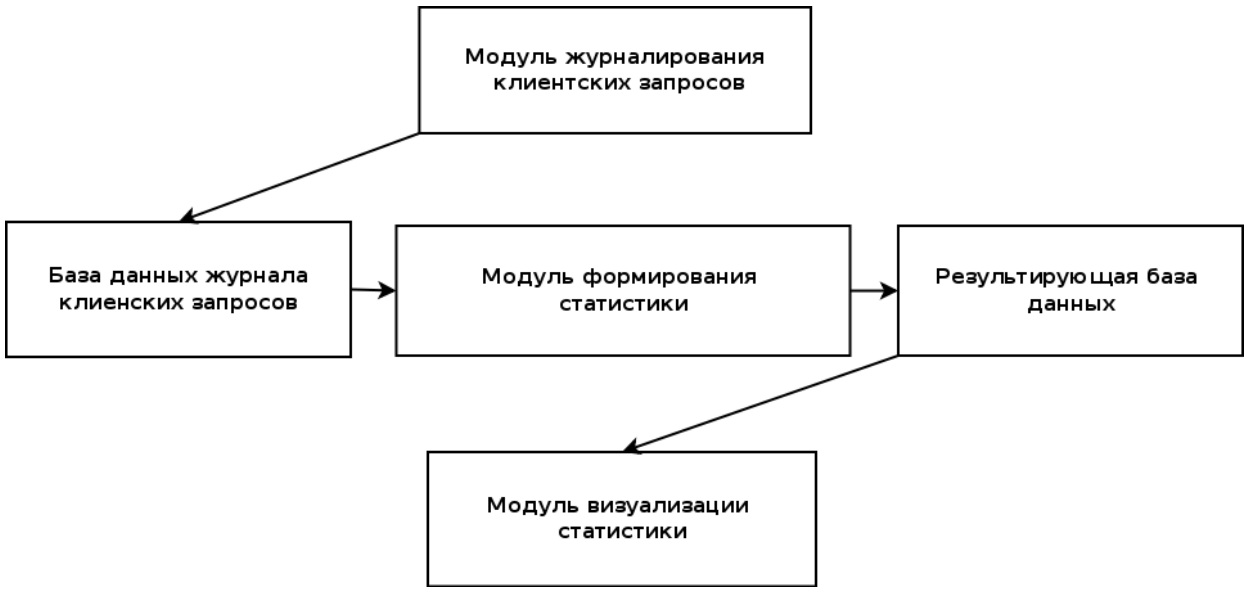
Большинство систем ведущих статистическую информацию о посещениях сайта используют анализаторы логов, которые не позволяют получать данные в режиме реального времени, или же являются коммерческими и требуют существенных денежных затрат.

Представляемый проект - это система статистики посещаемости сайта. Целью создания данной системы было - обеспечение в режиме реального

времени анализа посещаемости отдельных разделов сайта, определения их востребованности, реализации возможности генерации отчетов по различным критериям. Исходя из цели были поставлены следующие задачи:

1. Получать статистику в реальном времени.
2. Отслеживать посещения пользователей различных частей сайта.
3. Учитывать подробную информацию о пользователях(название операционной системы, название браузера, количество принятого и отправленого трафика и др.).
4. Обеспечить кроссплатформенную работу системы статистики.
5. Обеспечить кроссбраузерную визуализацию статистических результатов.
6. Обеспечить возможность создания графиков и печати статистических данных.

Система потенциально разрабатывалась как кроссплатформенная. В текущей реализации она использует операционную систему Linux, базу данный Mysql 5 и веб-сервер Apache2. Структуру работы системы можно представить в виде следующей диаграммы:



Вся система состоит из трех модулей и двух логических компонентов. Рассмотрим их более подробно.

База данных журнала клиентских запросов

Она включает в себя множество таблиц одинаковой структуры. Каждая из которых представляет результат посещаемости сайта за промежуток времени равный одному дню. Данный механизм позволяет наиболее эффективно располагать данные в MYISAM таблицах без их повреждения из-за большого объема данных. Структура данных таблиц имеет следующий вид:

remote_host	IP-адрес пользователя.
date_time	Время и дата посещения страницы сайта.
method	Тип запроса(Get, Post).
server_name	Имя хоста.
port	Номер порта.
directory	Путь к директории в которую произошел запрос
file_name	Имя файла к которуму произошло обращение
query	Строка параметров запроса.
status	Код http-запроса.
recv_seconds	Время получения данных.
recv_bytes	Количество полученных байт.
user_agent	Информация о клиентском приложении.

Результирующая база данных

Данный компонент включает таблицы использующиеся в визуализации статистической информации, а также при формировании статистики. Он содержит следующие таблицы: stats_days_shgpi, stats_months_shgpi, stats_browsers_shgpi, stats_os_shgpi, stats_months_os, stats_months_browsers, stats_criterious, stats_parts, stats_parts_dirs, stats_typeofstats, stats_user_agents, stats_unique.

Структура таблицы stats_days_shgpi:

cnt_bot	Количество запросов ботов
cnt_nbot	Количество запросов людей
cnt_in_bot	Количество запросов ботов из локальной сети
cnt_out_bot	Количество запросов ботов из глобальной сети
cnt_in_nbot	Количество запросов людей из локальной сети
cnt_out_nbot	Количество запросов людей из глобальной сети
cnt_ue_bot	Количество уникальных запросов ботов
cnt_ue_nbot	Количество уникальных запросов людей
toDirectory	Ссылка на путь к директории
date_time	Время последнего обновления

Данная таблица предназначена для хранения данных клиентских запросов в конкретную директорию за каждый день. В ней содержится статистическая информация сформированная по нескольким критериям: уникальности, посещениям из локальной сети, посещениям из глобальной сети и обоих сетей вместе.

Структура таблицы stats_months_shgpi имеют такой же вид как и stats_days_shgpi с одним исключением, что данные в ней хранятся просуммированные за месяц. Такой механизм позволяет избавится от необходимости подсчета посещений за длительный период, и тем самым уменьшает время по получению информации о клиентских запросах для визуализации.

Структура таблицы stats_browsers_shgpi:

cnt	Количество запросов людей
in_cnt	Количество запросов людей из локальной сети
out_cnt	Количество запросов людей из глобальной сети
ue_cnt	Количество уникальных запросов

	людей
browser	Название веб-браузера
toDirectory	Ссылка на путь к директории
date_time	Время последнего обновления

Структура таблицы stats_os_shgpi:

cnt	Количество запросов людей
in_cnt	Количество запросов людей из локальной сети
out_cnt	Количество запросов людей из глобальной сети
ue_cnt	Количество уникальных запросов людей
os	Название операционной системы
toDirectory	Ссылка на путь к директории
date_time	Время последнего обновления

Данные таблицы хранят информацию о веб-браузерах и операционных системах, которые используют пользователи осуществляющие запросы на сайт.

Таблицы stats_months_os и stats_months_browsers имеют одинаковые структуры с таблицами stats_os_shgpi и stats_browsers_shgpi с одним исключением, что данные в них содержатся просуммированными за месяц.

Структура таблицы stats_unique:

user_agent	Информация о клиентском приложении
remote_host	IP-адрес
directory	Путь к директории

В этой таблице хранится информация об уникальных запросах в конкретную директорию. Уникальность определяется различными сочетаниями 3 указанных выше параметров.

Таблица stats_criterious содержит информацию о названиях критериев отбора необходимых при визуализации.

Структура таблицы stats_parts:

toParent	Поддиректория

name	Название раздела сайта
link	Ссылка на раздел сайта

Данная таблица хранит информацию об директориях сайта, для которых формируется статистическая информация. В ней присутствует поле toParent с помощью которого можно создавать логические разделы сайта.

Таблица stats_typeofstats содержит информацию о видах статистических данных.

Структура таблицы stats_parts_dirs:

toPart	Ссылка на раздел сайта
directory	Путь к директории

Данная таблица содержит информацию о директориях, которые являются текущих или их синонимами, так называемая alias-таблица.

Модуль журналирования клиентских запросов

При выборе средств организации журнала были поставлены исходные требования:

1. Возможность получать статистику в режиме реального времени.
2. Возможность быстро формировать статистику по разным параметрам.
3. Независимость от ограничений формата журнала Apache2 или других приложений.

Все анализаторы журналов основанные на выгрузке данных непосредственно из журналов веб-сервера не подходили по пункту 1 и 2. Поэтому был использован модуль mod_log_config веб-сервера Apache2. Он использует в качестве приемника данных программу, которой на стандартный вход посылаются строки запросов в определенном формате. На основе этого модуля была написана программа на кроссплатформенном языке программирования Free Pascal, которая получала статистические данные с веб-сервера и отправляла их в базу данных журнала клиентских запросов в специально сформированную таблицу которая была описана выше.

Модуль формирования статистики

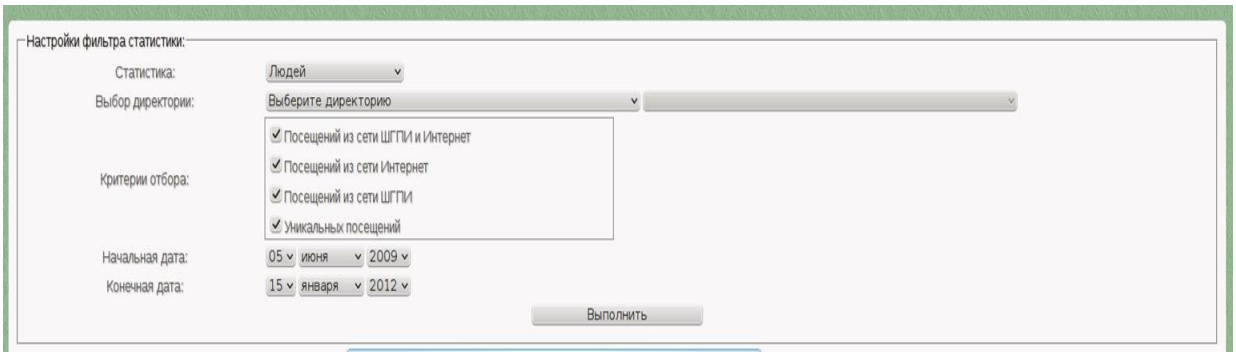
Данный модуль реализован в виде двух скриптов. Первый из них stats_hits реализован на языке sql и является хранимой процедурой внутри базы данных. Он выполняет задачи по извлечению из базы данных журнала клиентских запросов статистической информации, формированию конечных данных по различным параметрам и добавлению их в результирующую базу данных. Второй скрипт stats_update написанный на языке программирования php выполняет работу по запуску описанной выше хранимой процедуры с заданными параметрами. Он запускается программой cron (демон-планировщик задач в UNIX-подобных операционных системах) раз в одну минуту. С таким интервалом и происходит обновление статистических данный. Выбор такого поведения связан с экономией ресурсов сервера, для более мощных компьютеров этот промежуток можно уменьшить. Помимо этого скрипт позволяет восстанавливать результирующую базу данных в исходное состоянии при ее потере, проверять наличие новых данных и дозаписывать их в базу, а также добавлять статистические данные за промежуток времени в течении которого он не выполнялся. Параметры системы находятся в файле config.inc.php. Они включают:

1. host - начальное значение IP-адресов являющихся внутреней сетью организации.
2. bots - Список ключевых слов определяющих поисковых роботов маскирующихся под обычные браузеры.
3. user_agents — список валидных user-агентов используемых пользователями.
4. browsers — список веб-браузеров.
5. br_def — описание каждого веб-браузера на языке sql для распределения его между группами.
6. os — список операционных систем.
7. os_def — описание каждой операционной системы на языке sql для распределения ее между группами.

Модуль визуализации статистики

Данный модуль реализован в виде класса на языке php. Он делает выборку статистической информации из результирующей базы данных и формирует ее визуальное представление с возможностью выбора различных параметров просмотра. Также существует возможность просматривать графики посещения по различным критериям, популярности браузеров, популярности операционных систем. Для их построения используется бесплатная свободно распространяемая библиотека Libchart.

Для выбора параметров визуализации используется простой и понятный фильтр:



В него входит:

1. Выбор вида статистики(например, статистика браузеров).
2. Выбор директории и поддиректории(например, библиотека).
3. Выбор критериев отбора(например, Из сети Интернет).
4. Выбор временного интервала(например, с 01 июня 2010 по 02 июня 2010).

Модуль визуализации реализован с поддержкой всех современных браузеров, а также позволяет создавать персональные отчеты по различным параметрам.

Ознакомится с работой системы в реальных условиях и с данными о посещениях почти за 3 года можно по адресу — http://shgpi.edu.ru/stats/