УЧЕБНЫЙ ЦЕНТР «ТОПЭКСПЕРТ»

Модуль «Ловец ботов»

10 поток профессионального курса по SEO

Работу выполнил: Михаил Баринов

Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Модуль: Ловец ботов

Модуль анализирует лог-файлы веб-сервера, выявляет роботов, отслеживает их активность.

1. Задачи, которые решает модуль:

- Отслеживание посещений страниц сайта роботами и анализ путей их перемещения по сайту
- Идентификация «черных» роботов- различных парсеров, спам-ботов.
- Идентификация «белых» роботов- роботов поисковых систем.

2. Входные данные:

- Лог-файлы веб-сервера (/var/log/apache2/*)

Формат файла access_log

IP	Раздел итель	Дата:	Метод передачи данных	Адрес документа	Прото кол	Код ответа сервера	Размер файла в байтах
137.140. 101.25	- -	[08/Sep/2013:03:16:20 +0400]	"GET	/instr/aminet110 .pdf	HTTP/ 1.1"	200	634910
Прим.		*[Число/Месяц/Год:Часы:Минуты:Секунды Чаовой пояс]*	*GET|POST*				

Формат файла error_log

Дата	Тип уведомл ения	При чьем обращении появилась ошибка	Текст ошибки	Адрес документа при обращении к которому произошла ошибка	Адрес документа с которого посетитель попал на страницу с ошибкой
[Sun Sep 08 06:11:50 2013]	[error]	[client 178.93.118.54]	File does not exist:	/var/www/site/phpbb,	referer: http://www.site.ru/phpbb/index.php

- Таблица соответствий IP-USER-AGENT в БД

ID	IP	USER-AGENT
34	87.250.255.243	User-Agent Mozilla/5.0 (compatible; Yandex...)

3. Выходные данные:

1. Список посещенных каждым роботом страниц. Выделить страницы, которые уже содержатся с индексе поисковой системы.

URL документа	Адрес откуда прише л	Адрес куда ушел	Дата прихода	Дата ухода	Время на страниц е	Тип робота	Страниц а в индексе
/kuz_remont /	/	/kuz_remont/pokraska /	10.09.201 3 22:59:11	10.09.201 3 22:59:22	11 сек.	Yandex (87.250.255.243)	Да

2. Пути переходов по страницам сайта для каждого робота.

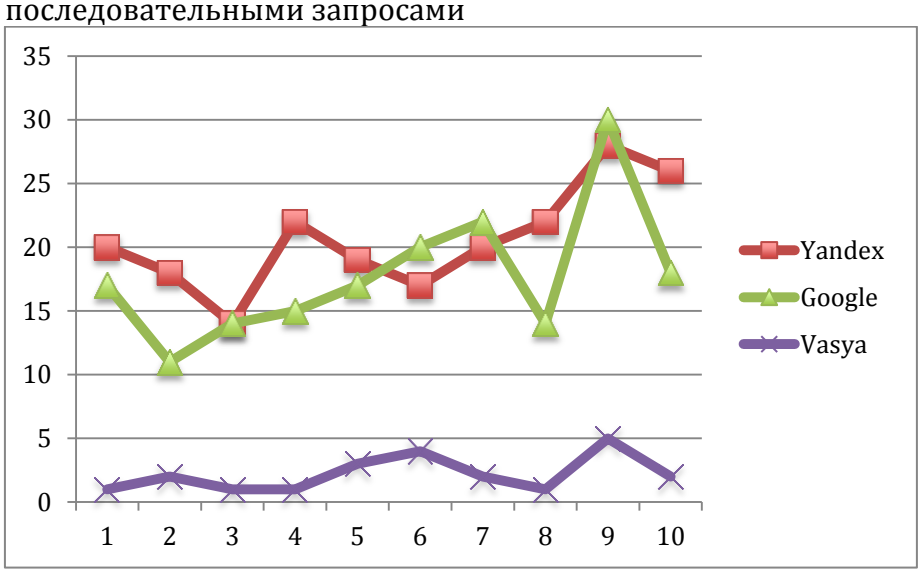
Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

GoogleBot (66.249.65.141)	
Дата	**Адрес страницы**
2013.09.10 22:07:14	/
2013.09.10 22:07:44	/news

3. Основные метрики:
   a. Среднее число запросов в минуту и его с crawl-delay в robots.txt

Робот	**Ср. число запросов в мин.**	**Отклонение от crawl-delay**
Google	20	0%
Yandex	18	10%
Vasya	5	75%

   b. Распределение величины временного интервала между последовательными запросами



   c. Число уникальных IP адресов, с которых шли запросы.

Уникальных IP адресов:	9625540

   d. Статистика посещений страниц закрытых в robots.txt.

Робот	**Дата**	**Страница**
Yandex (87.250.255.243)	08.09.2013 14:11:56	/logs
176.9.157.230	09.09.2013 11:58:12	/secretzone

   e. Частота заходов каждого робота на сайт

Yandex (87.250.255.243)	
Дата захода	**Время от предыдущего захода**
01.08.2013 12:30:30	25 часов 10 минут 20 секунд
03.08.2013 13:40:20	25 часов 09 минут 50 секунд
05.08.2013 15:50:30	26 часов 10 минут 10 секунд
Средняя частота заходов робота на сайт	**1 раз в 25 часов 30 минут 7 секунд**

   f. Число уходов робота из-за ошибки

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

Робот	Число уходов из-за ошибки
Yandex	4
Google	6
Всего:	**10**

4. Список роботов, которых необходимо забанить

Робот	Дата последнего захода
VasyaBot	14.08.2013 12.00.11
PetyaBot	18.08.2013.14:34:19

5. Список URL переход на которые повлек появление ошибки

Робот	**Дата**	**Адрес переход на который повлек ошибку**	**Адрес при переходе с которого произошла ошибка**	**Ошибка**
Yandex	07.09.2013 11:01	/page/visuzel/	/page/contacts/	File does not exist
Google	09.09.2013 09:10	/page/img/bg.png	/page/about/	File does not exist

4. Формулы

1. Самоидентификация робота (RSI- Robot Self-Identification)
   RSI=1, если робот не представился, то есть представился как обычный пользователь (USER-AGENT браузера)
   RSI=0, если робот представился.
2. Соблюдение Crawl-Delay (CDV- Crawl Delay Violation)
   CDV=max(C/T)-1,
   где C- Crawl-Delay в секундах,
   T- интервал между последовательными переходами.

   Если CDV=0, робот соблюдает Crawl-Delay. Чем больше CDV,
   Тем больше робот нарушает Crawl-Delay.

   Отрицательное значение указывает на то, что робот переходит со страницы на страницу с интервалом выше указанного в Crawl-Delay и данном случае может быть приравнен нулю (CDW=0)
3. Игнорирование инструкций в robots.txt на индексирование файлов (IFF- Ignoring Forbidden for Indexation files)
   Создать директорию, запретить ее индексацию в robots.txt.
   Посчитать число страниц, посещенных роботом, в данной директории.

   IFF=ln(N+1),
   где N- число посещённых страниц в запрещённой для индексации директории.

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

Если IFF=0, робот соблюдает инструкции в robots.txt.
Чем выше значение IFF, тем больше робот нарушает правила запрета на индексацию.

4. Анализ поведения роботов Яндекс, Google, Bing. Сравнение поведения с поведением спамных роботов. Определение метрик по которым можно будет выделять спамных роботов. Например,

   Минимальное время между переходами с страницы на страницу $T_{ср}$
   $T_{мин}$=МИН(($T_{мин}$(Я),$T_{мин}$(G),Тмин(B))
   где $T_{мин}$(Я) – минимальное время между переходами со страницы на страницу робота Яндекс, $T_{мин}$(G) – минимальное время между переходами со страницы на страницу робота Google, $T_{мин}$(B) – минимальное время между переходами со страницы на страницу робота Bing.

   Максимальное время между переходами со страницы на страницу
   $T_{макс}$=МАКС($T_{макс}$(Я) ,$T_{макс}$(G),$T_{макс}$(B))
   где $T_{ср}$(Я) – максимальное время между переходами со страницы на страницу робота Яндекс, $T_{ср}$(G) – максимальное время между переходами со страницы на страницу робота Google, $T_{ср}$(B) – максимальное время между переходами со страницы на страницу робота Bing.

   Отклонение от T(Rob) - DFMT (Deviation From Mean Time)
   Если выполняется условие:
   $T_{мин}$>T(Rob)< $T_{макс}$ => DFMT=0, иначе DFMT=1, означает, что на робота надо обратить внимание.
   T(Rob)- среднее время между переходами со страницы на страницу исследуемого робота.

5. Анализ отклонений между различными посещениями сайта, например, роботом Google. Если одно из посещений сильно отличается от средних значений, то возможно, какой-то робот выдает себя на робота Google. Необходимо выделить этих роботов по IP и далее отслеживать их.

6. Фактор затухания прошлых грехов. Если мы проводим анализ каждую неделю, то фактор спамности N-недельной давности делится на 2^N
7. Время бана BT (Ban Time), BT=60*N минут, гдеN – фактор спамности.
8. Вероятность того, что робот является «Черным» BR (Black Robot)
   BR=0 все нормально, пропустить,
   BR=33 подозрительно проследить
   BR=66 очень подозрительно, выдать капчу
   BR=99 «черный» робот, забанить на время TB=60*N, где N– фактор спамности.

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

6. Список модулей, с которыми взаимодействует модуль

Происходит взаимодействие со следующими модулями:

- Модуль выявления соответствий IP-USER-AGENT
- Модуль блокиратор «черных» роботов, парсеров, спам-ботов

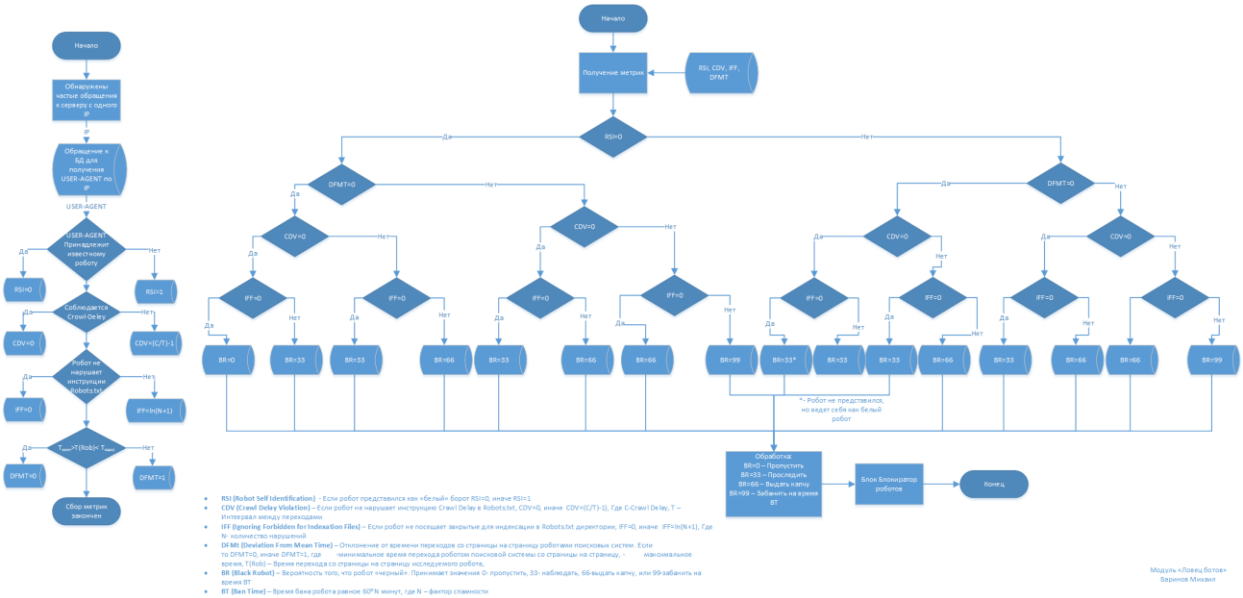
7. Описание процессов взаимодействия

- Модуль выявления соответствий IP-USER-AGENT
    - Получение данных из Таблицы базы данных, в которой содержится информация о соответствиях.

8. Карта логических связей выполнения модуля

Сбор метрик

Анализ BR – Вероятность, что робот «черный»



9. Предполагаемая нагрузка

- Обработка больших объем информации может повлечь снижение производительности системы.

10. Особые требования

- Модуль запускается автоматически по CRON с периодичностью указанной в CRONTAB
- Время запуска модуля должно соответствовать времени наименьшей загрузки сервера
- Если сервер имеет равномерную загрузку в течении суток, перенести обработку на отдельный сервер.

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

- Модуль не должен обрабатывать уже обработанные данные повторно. Результаты проверки и время последней обработанной строки должны быть сохранены в БД

11. Процесс остановки модуля

- Модуль останавливается автоматически по завершению проверки

12. Процесс запуска модуля

- Модуль запускается автоматически согласно расписанию прописанному в CRONTAB.
- Модуль запускается повторно через интервал времени не менее среднего времени анализа лог-файла, при условии, что предыдущий анализ уже закончен и загрузка сервера не превышает на текущей момент обозначенного максимума.
- В случае не соблюдения указанных выше условий, робот повторяет попытку запуска через заранее определенный интервал времени.

13. Формирование резервных копий

- Необходимо один раз в сутки делать резервную копию базы данных модуля
- Резервное копирование каталога, где расположены файлы модуля, один раз в сутки

14. Восстановление из резервной копии

- Восстановление путём развёртывания последней резервной копии БД с результатами анализа
- Восстановление путем развёртывания последней резервной копии каталога с файлами модуля.

15. Предполагаемые расширения модуля

Доработать модуль до блокиратора «черных» роботов, парсеров, спам-ботов

16. Возможные причины поломки модуля

- Нарушение работоспособности БД
- Аппаратные неисправности сервера
- Невозможность продолжения анализа лог-файла, так как файл отсутствует на сервере (например файл уже заархивирован, а вместо него создан новый)

17. Работа модуля в случае поломки

- Нарушение работоспособности БД.
  Модуль останавливает свою работу.
  - Сделать откат к рабочей версии БД
  - Запустить модуль вручную

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов

- Аппаратные неисправности сервера.
  Модуль прекращает работу.
    - Восстановить работоспособность сервера
    - Запустить модуль вручную
- Невозможность продолжения анализа лог-файла, так как файл отсутствует на сервере (например файл уже заархивирован, а вместо него создан новый)
    - Модуль начинает анализировать новый лог-файл сначала.
    - Необходимо подобрать такое расписание запуска модуля, чтобы модуль успевал анализировать логи до их архивации.

Автор: Михаил Баринов, 10-ый поток курсов ТопЭксперт
Дипломный руководитель: Дмитрий Иванов