УЧЕБНЫЙ ЦЕНТР «ТОПЭКСПЕРТ»

Модуль анализа путей роботов по логам

9 поток профессионального курса по SEO

Работу выполнил: Сергей Ворожцов

Дипломный руководитель: Дмитрий Иванов

ТОПЭКСПЕРТ
УЧЕБНЫЙ ЦЕНТР

Модуль анализа путей роботов по логам

1. Описание проекта

Модуль будет анализировать посещения и пути всех роботов по сайту (роботы могут идентифицировать себя, не идентифицировать себя, а также представляться не своими именами). Модуль облегчит краулинг для “белых” роботов и забанит “чёрных” роботов.

2. Задачи, которые решает модуль

1. отслеживание посещений роботами страниц и анализ путей роботов по сайту
2. идентификация “чёрных” роботов
3. идентификация “белых” роботов

3. Входные данные (описание и таблица)

Входные данные -- детальный серверный лог по каждому запросу к серверу. Входными данными также будут классифицированные выходные данные в случае если модуль будет самообучаться.

4. Выходные данные (описание и таблица)

1. для каждого робота выводить список посещённых им страниц (можно отмечать те посещённые роботом страницы, которые уже содержатся в его индексе)
2. для каждого робота выводить цепочки страниц его переходов
3. основные метрики по каждому роботу:
- среднее число запросов в минуту и его сравнение с crawl-delay прописанным в robots.txt
- распределение величины временного интервала между последовательными запросами
- число страниц посещённых за каждый визит
- число различных IP адресов, с которых шли запросы
- статистика по посещению страниц в закрытой от индексации в robots.txt директории
4. другие метрики
5. список роботов, которых предлагается забанить

5. Формулы

Отделяем козлищ от агнцев: роботов от обычных посетителей. Записываем в роботы тех, кто:
- представился как робот
- совершает посещения большого числа страниц в единицу времени (всплеск) или

Автор: Сергей Ворожцов, 9-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

за сутки (кумулятивный эффект)

Считаем такие факторы спамности для каждого робота.

1. само-идентификация робота (RSI = robot self-identification)

RSI = 1, если робот не представился, то есть прикидывается обычным посетителем RSI = 0, если робот представился

2. соблюдение роботом crawl-delay (CDV = crawl-delay violation)
T = средний интервал между последовательными запросами от робота в секундах
C = crawl-delay в секундах
CDV = max (C / T, 1) - 1
Величина от 0 до нескольких десятков. Если CDV = 0, то робот соблюдает crawl-delay. Чем больше значение CDV тем сильнее робот нарушает crawl-delay.

3. игнорирование запрещённых к индексации в robots.txt файлов (IFF = ignoring forbidden for indexation files)
Создаём директорию, запрещаем её индексацию в robots.txt. Считаем N -- число страниц из этой директории, которые посетил робот.
IFF = ln (N + 1)
Если IIF = 0, то робот соблюдает игнорирование запрещённых к индексации в robots.txt файлов. Чем больше IIF, тем сильнее робот нарушает правило запрета.

4. анализируем поведение роботов Яндекс, Google и Bing. Сравниваем их поведение с поведением спамных роботов. Определяем метрики, по которым можно будет отделять спамных роботов.

5. анализируем отклонения между различными посещениями сайта, например, Google'ом. Если одно из посещений сильно отклоняется от стандартных, то это будет сигналом к тому, что возможно другой робот ходит под личиной Google'а. Нужно выделить его в отдельную позицию, например, по IP адресу и далее отслеживать его поведение.

6. вводим фактор затухания прошлых грехов. Если мы проводим анализ каждую неделю, то фактор спамности N-недельной давности делится на 2^N.

На основе вышеуказанных факторов строим эвристику и ранжируем

Автор: Сергей Ворожцов, 9-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

роботов по спамности.

6. Список модулей, с которыми взаимодействует модуль

00. sitemap.xml
20. Модуль проверки robots.txt
21. Модуль поиска битых ссылок и ссылок на редиректы
28. Модуль проверки индексации страниц
69. Поиск висячих узлов
70. Проверка скорости загрузки страницы
100. Модуль защиты от парсинга

7. Описание процессов взаимодействия

00. sitemap.xml

Для страниц посещённых (или проиндексированных) Яндексом и Google'ом, в sitemap.xml ставим <priority> = 0.1, для остальных оставляем <priority> = 1.0

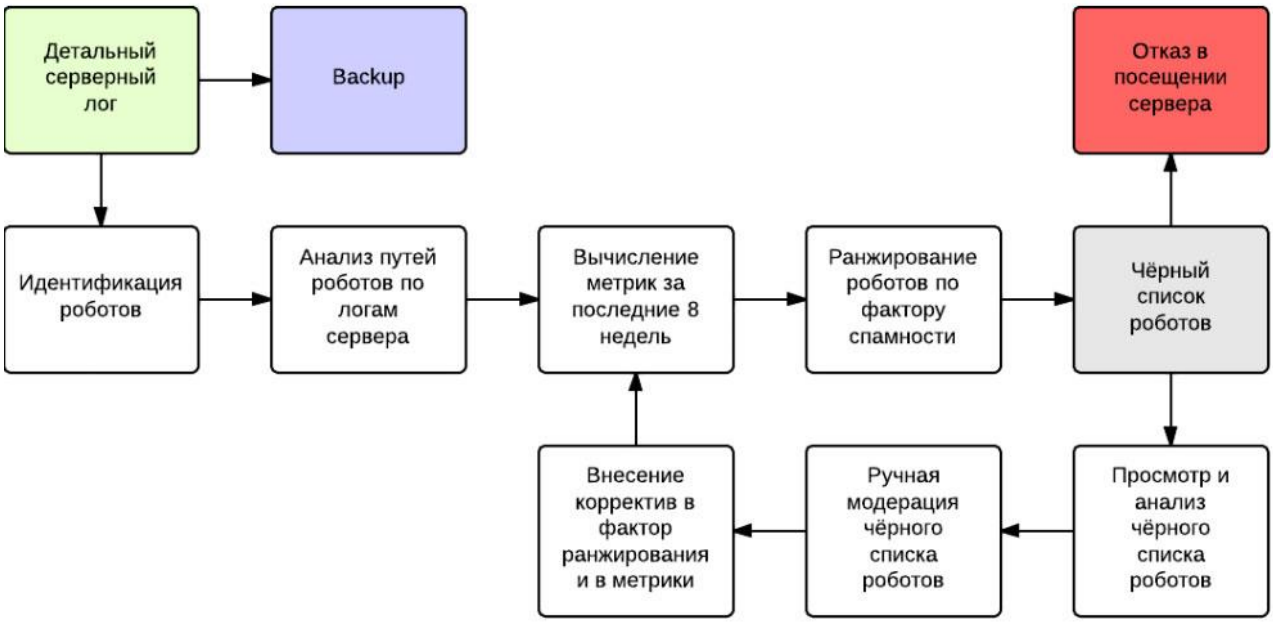
20. Модуль проверки robots.txt

В формулах используется информация из файла robots.txt

21. Модуль поиска битых ссылок и ссылок на редиректы

Проверяем, что ссылки отдают код 200.

8. Карта логических связей выполнения модуля



Детальный серверный лог по каждому запросу к серверу ведётся с момента создания сервера и каждую неделю в 3 часа утра каждую субботу (бэкап данных аккумулированных с субботы по пятницу).
Каждую неделю в 3 часа утра по субботам через cron запускается модуль анализа путей роботов по логам. В соответствии с эвристикой роботы ранжируются по фактору спамности. Определяется чёрный список роботов. Этим роботам будет отказано в обращениях к нашему серверу в ближайшую неделю.
Через user-interface администратор может посмотреть список роботов, метрики для каждого робота и фактор ранжирования.
Администратор может вручную забанить или наоборот разбанить любого робота, а также изменить формулу вычисления фактора спамности. Программист может внести изменения в формулы для метрик.
Администратор также может вручную запустить пересчёт ранжирования роботов.

9. Предполагаемая нагрузка

Нагрузка связана с анализом модулем серверных логов. При большом числе посетителей на сайт (большом размере логов) расчёт метрик и факторов ранжирования может отнимать серверные ресурсы. Рекомендуется работа модуля в режиме низкого приоритета, чтобы не отнимать серверные ресурсы.

10. Особые требования

Записывать backup на backup'ный серверный диск.

11. Процесс запуска модуля

Модуль запускается каждую неделю с помощью cron. Модуль также запускается при обращении к нему администратора.

12. Процесс остановки модуля

Модуль останавливается после расчёта чёрного списка роботов.

13. Формирование бекапов

Нужно сохранять детальную статистику серверных логов за всё время существования сервера.

14. Восстановление бекапов

При потере данных для расчёта метрик за последние 8 недель можно восстановить

Автор: Сергей Ворожцов, 9-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов

эти данные из backup’ов.

15. Предполагаемое расширение модуля

Перевод работы модуля в режим машинного обучения.

16. Возможные причины поломки модуля

Проблема или некорректная работа серверных логов может вызвать ошибку.

17. Работа модуля в случае поломки на каждом участке

Все данные по серверным логам будут в backup’ах и могут быть восстановлены из них. При поломке самих скриптов программист должен исправить ошибки и перезапустить модуль.

Автор: Сергей Ворожцов, 9-ый поток курсов ТопЭксперт

Дипломный руководитель: Дмитрий Иванов