

Алексей НАБЕРЕЖНЫЙ,

Юрий КОБЕОСКИЙ

Интернетом правят роботы

Многие начинающие владельцы веб-сайтов мечтают увидеть ссылки к своим ресурсам на самых высоких местах в рейтингах поисковых систем. Но большинство, как правило, не оказывается даже в первой двадцатке, не то что среди лидеров. А между тем веб-мастеру следует подумать, сделал ли он все возможное, чтобы его сайт стал лидером.

Давайте рассмотрим типичный случай раскрутки и последующего процветания тематического веб-сайта. Вы покупаете домен, находите доступный хостинг и настраиваете его под ваш сайт. Регистрируете веб-ресурс на поисковых машинах типа Google, Yahoo, Altavista, Yandex, Rambler и т.д. Чтобы сориентироваться в выборе лучшего автоматического помощника для эффективного продвижения на всех мировых поисковиках, просмотрите статью «Прописка в Интернете» (<http://www.osp.ru/pcworld/2002/12/060.htm>). Вот, собственно, и все. Что теперь? Можно, конечно, сидеть и ждать, пока поисковики направят на ваш сайт роботов — специальные программы-агенты. Они постранично исследуют ваш ресурс, проиндексируют все страницы, проверят ссылки и сохранят информацию в БД поисковой машины. К сожалению, ускорить данный процесс вы не способны. Но даже когда ваш сайт наконец-то посетит робот (он же паук, он же спайдер, он же веб-агент), то это не значит, что ваш ресурс тут же будет добавлен в БД поисковой машины. Иногда ваша заявка будет рассматриваться работниками компании несколько недель. Но не ждите у моря погоды и потратьте это время на подготовку визита долгожданных поисковых роботов на ваш многообещающий сайт. Что мы имеем в виду?

Несомненно, визит роботов — вещь желательная и полезная, поскольку они улучшают видимость ресурса и повышают трафик. Но если ваша главная входная страница содержит рекламные ролики, апплеты, Java-скрипты и т.д., то велика вероятность, что поисковик так никогда и не узнает о содержании сайта ничего полезного, кроме информации о первой страничке. Как правило, роботы не способны выудить навигационные ссылки из различных динамических сценариев, хотя любому пользователю позволяется свободно передвигаться по сайту, переходя от одной страницы к другой. Поисковый агент сможет «увидеть» нужную информацию с помощью одного специального файла robots.txt, размещенного на сайте.

Веб-ресурс содержит не только данные, предназначенные для всеобщего обозрения, но и личные сведения, информацию для служебного использования внутри фирмы, страницы на разных языках, не поддерживаемые отдельными поисковиками, закрытые мультимедийные файлы, адреса электронной почты — одним словом, информацию «не для посторонних». Естественно, нужно ограничить допуск к данным такого рода, а тем более индексацию этих документов поисковыми роботами. Далее, например, поисковик Google отслеживает ваши странствия по Сети через встраиваемую в Internet Explorer панель инструментов и сообщает на сайт «хозяина» о тех страницах, что вы посетили. И если вы набрали в адресной строке браузера изолированный адрес www.YourSite.com/MyNudePhotos20034, то будьте уверены, что на этот каталог сайта вскоре будет послан поисковый робот, который его проиндексирует, даже когда отсутствуют внешние ссылки. Неприятно? Почитайте обсуждение этой темы на форуме профессионалов (<http://www.webmasterworld.com/forum80/39.htm>) и тогда поймете, откуда GoogleBot знает о существовании файла, к которому нет и не может быть никаких внешних ссылок. Именно так пару лет назад интернет-коллекцию наших фотографий пополнили некоторые тайные «фотошедевры» российских нудистов. Таким же образом поисковые машины «сдают» пиратские сайты с серийными ключиками, вследствие чего те немедленно закрываются под давлением авторов программ.

Но не так уж все безнадежно. Опытные мастера способны не только выдавать нужные инструкции к действиям роботов через файл robots.txt, но и проверять эффективность своей работы и определять популярность сайта.

Еще в 1994 г. (!) группа интернет-пользователей под руководством Мартина Костера разработала протокол исключения роботов. Он быстро был принят веб-мастерами и изготовителями поисковых программ-роботов, чтобы упорядочить процесс индексации и контролировать его. Когда вы создадите стандартный файл robots.txt, то получите возможность запретить одному, нескольким или даже всем роботам доступ к каталогу, файлу или определенному типу файлов. Вам необходимо лишь правильно отформатировать текстовый файл и поместить его в корень веб-сервера с именем robots.txt (помните, что все буквы должны быть строчными!). При посещении веб-ресурса цивилизованные пауки в первую очередь запрашивают данный файл, а затем покорно следуют

его указаниям. Но, как показывают исследования, не все роботы столь «законопослушны». Чтобы убедиться в этом, нужно открыть лог-файл, где серверный компьютер фиксирует данные о посещениях веб-сайта. В каждой записи такого журнала приведены сведения о времени посещения, запрошенной странице, пользовательском агенте посетителя (программе, примененной для открытия страницы) и IP-адресе, с которого пришел запрос.

Для многострадальной раскрутки веб-сайта мы рекомендуем специальное ПО, значительно облегчающее его продвижение и настройку, а также анализирующее эффективность прилагаемых усилий. Здесь рассмотрены две программы, предусматривающие как создание файлов robots.txt, так и контроль над следованием данным там инструкциям, и описаны еще четыре, исполняющие лишь одну из указанных функций. Часть упоминаемых программ представляют собой прекрасный инструмент для изучения популярности сайта и для определения характеристик его аудитории. Почти все программы анализа служебного лог-файла могут легко соперничать по популярности с самыми распространенными онлайн-сервисами статистики посещений в России. Но обо всем по порядку...

RoboGen 1.52

Эта небольшая программа имеет ограниченные возможности и английский интерфейс, однако анонсирована разработчиками как современная и мощная. Предназначена для создания файла robots.txt и загрузки его на сайт по FTP, с чем легко справляется, имеет базу данных с информацией о 178 известных роботах (это довольно много). Кроме того, есть дополнительная бесплатная надстройка в виде утилиты RoboTag 1.5 для создания метатегов, встраиваемых в html-текст веб-страниц.

RoboGen 1.52 совместим с Windows 95/98/Me/2000/XP. Программа появилась на свет в сентябре 1999 г., а ее последняя версия, 1.52, была размещена на сайте компании в июле 2002 г. Стоимость программы — 24,95 долл., сайт — <http://www.rietta.com/robogen/>.

Недостатки:

- не предусмотрен анализ лог-файлов;
- невозможна работа через прокси-сервер.

Web Log Explorer — Log Analyzer 1.8

Очень мощная программа для проведения многостороннего анализа посещаемости веб-сайта, написанная Сергеем Кармановым из Кемерово. Она также генерирует 23 основных и вспомогательных отчета в табличной форме и диаграммах. Просмотреть статистику посещаемости веб-ресурса допускается практически под любым углом и в любой форме, включая Search Engines Report и SE Robots Report (именно здесь отображаются данные о поисковых роботах, побывавших на сайте). Кстати, по отчетам удастся легко понять логику поведения и раскрыть некоторые секреты работы поисковых машин с сайтом.

Программа распознает 23 основных поисковых робота-паука.

Web Log Explorer предназначена, скорее, для ведения общей статистики популярности веб-сайта. Вы можете просматривать графики посещений по дням недели, по времени суток, по пользовательским агентам и даже по операционным системам пользователей. В общем, берегитесь, специальные интернет-сервисы SpyLog и HotLog: эта программа удобней и полезней ваших платных услуг! Тем более что лог-файлы посещения могут быть расписаны программой по 53 тыс. городов мира, включая более 500 российских населенных пунктов. Жаль только, что интерфейс программы не поддерживает русский язык (правда, автор обещал в самом ближайшем будущем сие исправить), но даже начинающие веб-мастера сумеют разобраться с данным приложением на английском. А если вы заглянете в файл помощи программы, то найдете информацию, подсказывающую, как выйти практически из любой затруднительной ситуации.

Web Log Explorer правильно обрабатывает самые распространенные форматы лог-файлов: Apache Extended, Apache Common, IIS Log Formats 4.0, Standard Common Log File Format, Standard Combined Log File Format, NCSA Common Log File Format, NCSA Combined Log File Format, CERN Common Log File Format и CERN Combined Log File Format.

Программа Web Log Explorer — Log Analyzer 1.8 стоит 49,90 долл., но в течение первых 30 дней работает бесплатно.

На сайте разработчиков лежит программа-близнец Web Log Suite — Log Analyzer 1.8, генерирующая статистику посещаемости в виде готовых HTML-страниц. Стоимость этого анализатора — 54,90 долл.

Недостатки:

- ограниченный анализ по 23 роботам;
- отсутствие редактора файлов robots.txt (ожидается в следующей версии);
- отсутствие модуля, работающего на стороне сервера (приходится загружать на ПК большие лог-файлы).

Сайт программы — <http://www.exacttrend.com/WebLogExplorer/>.

Advanced Log Analyzer 1.4

«Навороченная» программа для всестороннего анализа log-файлов сайтов. Может работать в режиме графического интерфейса, а также в качестве консольной программы. Имеет серверный вариант как CGI-приложение для работы на Microsoft Internet Information Server (IIS) и Apache-серверах. Предоставляет широчайшую свободу настроек и традиционных отчетов по сотням показателей (наиболее посещаемые страницы, статистика посещений по дням, неделям и месяцам), и нетрадиционных (время, ушедшее на загрузку файлов и страниц). Работает с девятью основными форматами лог-файлов, в том числе и с архивированными логами — .gz-файлами. Имеет такую интересную функцию настройки, как возрастающий анализ (incremental analysis), предусматривающий кэширование предварительно проанализированных лог-файлов. Таким образом, при работе с файлом больших размеров программа не анализирует его целиком, а использует сохраненные в кэше данные, разбирая только свежие записи. В ней есть также и планировщик заданий, автоматически просматривающий лог-файлы с заданной периодичностью. Думаем, что эту программу можно было бы считать лучшей для анализа лог-файлов, если бы не одно маленькое «но» — хотя она и имеет наряду с англоязычным интерфейсом еще пять дополнительных, среди них нет русского, так что программа не умеет правильно обрабатывать запросы российских поисковых машин. Если ваш сайт рассчитан на аудиторию за пределами России, обязательно познакомьтесь с ней (один месяц ею можно пользоваться бесплатно).

Теперь о грустном. При простом на первый взгляд принципе организации основного окна программы, где имеется пять панелей, переключаться между которыми помогают навигационные кнопки, бросается в глаза некая хаотичность их расположения. Иногда программа поступает не так, как вы ожидаете, — при нажатии отдельных кнопок, предназначенных, казалось бы, понятно для чего, вас забрасывает не на ту панель, куда требовалось. Сложно было разобраться и в настройках отчетов, и в фильтрах результатов.

Продукт Advanced Log Analyzer 1.4 для использования дома стоит 49,99 долл., а его профессиональная версия — 129,99 долл.

Недостатки:

- сложная организация интерфейса;
- отсутствие БД роботов и редактора файлов robots.txt.

Сайт программы — <http://www.vknoware.com/ala/>.

123LogAnalyzer 3.10

Программа 123LogAnalyzer — удобный инструмент, превращающий «сырые» лог-файлы в упорядоченные результаты всестороннего анализа аудитории и перспектив вашего ресурса. 123LogAnalyzer описывает, каким образом посетители просматривают ваш веб-сайт, какие страницы читают (или игнорируют), как долго проводят время на вашем веб-участке, откуда прибывают, по каким поисковым словам находят вас и т.д.

Прогрессивные фильтры 123LogAnalyzer позволяют изучать только тех посетителей, которые попали на определенные страницы, или тех, кто загрузил определенные файлы, например мультимедиа. Вы можете также сосредоточиться на тех, кто прибывает со специфического URL, из определенной страны или в конкретные часы.

Программа-анализатор функционирует как на различных ОС клиентского ПК, так и на стороне сервера вашего сайта (Windows NT/2000 server, Linux/FreeBSD/ Sun Solaris/Mac OS X servers). Она умеет хорошо работать с большими лог-файлами и способна ежесекундно анализировать около 94 тыс строк, т. е. за минуту справляется с 1-Гбайт файлом.

Англоязычный 123LogAnalyzer 3.10 с однопользовательской лицензией стоит 129,95 долл., а с десятью корпоративными лицензиями — 149,95 долл. Более того, если вы захотите добавить географические отчеты по городам, то вам придется выложить еще 99,99 долл. Но первые 30 дней программа 123LogAnalyzer будет работать бесплатно. Пределом запросов владельцев 123LogAnalyzer 3.10 стали 699,95 долл. за версию, которую можно устанавливать не более чем на 50 компьютерах. А у пиратов мы видели такую же всего за 2 долл...

По количеству и возможностям настроек отчетов мы назвали бы эту программу лучшей среди рассмотренных здесь, однако ее цена, а также то, что она не способна работать с файлом robots.txt, заставили нас отдать предпочтение другому ПО.

Недостатки:

- запутанная система отчетов;
- отсутствие БД поисковых роботов и редактора файлов robots.txt.

Сайт программы — <http://www.123loganalyzer.com/>.

Robots.txt Editor and Log Analyzer 1.1

Данная программа — новичок на мировом рынке, но уже успела себя отлично зарекомендовать. Когда мы пользовались Robots.txt Editor and Log Analyzer, у нас сложилось впечатление, что разработчики предусмотрели все, хотя такого конечно же не бывает. База данных о пользовательских агентах роботов включает 182 записи, которые можно пополнять своими, регулярные обновления, русскоязычный интерфейс, детальный и содержательный файл помощи.

Программа позволяет запрещать доступ не только к каталогам сайта, но и просто к файлам с определенным расширением. Раньше мы не встречали такой функции, а между тем, согласитесь, это весьма полезная возможность, особенно если вы храните на сайте мультимедийные файлы, с особым упорством и регулярностью копируемые роботами Google, что создает огромный и бесполезный для вас трафик. Мы сталкивались с этим на сайте www.russia.digdot.com, где хранятся несколько наших мультимедийных файлов о России.

Лог-анализатор автоматически определяет формат лог-файла (поддерживаются серверы Apache, Microsoft IIS и «маргинальный» Zeus) и позволяет работать с архивными лог-файлами формата .gz. Даже если программа затруднится самостоятельно определить формат (мы скармливали ей все возможные типы, но поставить ее в тупик так и не смогли), все равно будет доступна функция настройки формата. Лог-анализатор дает возможность импортировать лог-файлы, извлекать оттуда данные и экспортировать их в форматы Microsoft Excel CSV, XML или в один из серверных лог-форматов для дальнейшей обработки и сохранения. Особое внимание следует обратить на сохранение и пополнение БД посещений роботов, чтобы позднее можно было отслеживать тенденции.

Многоязычный интерфейс программы Robots.txt Editor and Log Analyzer удобен в навигации. Разобраться в основных функциях довольно легко, даже не обращая ни к каким справкам, хотя кнопка их вызова есть в каждом окне.

Панель Disallow (Запретить) позволяет проводить навигацию по папкам и файлам сайта, сохраненным на локальной машине или на FTP-сервере. В зависимости от поставленных целей и веб-дизайнерского опыта вы можете выбрать путь к отдельному файлу, запретить доступ к нему всем или одному роботу, закрыть доступ ко всему сайту или снять ограничения для отдельных каталогов или файлов. Здесь же разрешается запретить доступ паукам к файлам по расширению (например, .mp3 или .exe).

Пиктограммы разного цвета, обозначающие успешные или неуспешные запросы роботов, существенно облегчают восприятие информации. Особым значком маркируются роботы, не запросившие файл robots.txt и, следовательно, представляющие собой потенциальную угрозу безопасности сайта.

Robots.txt Editor and Log Analyzer 1.1 доступен в двух ценовых вариантах: Professional edition за 49,90 долл. и Enterprise edition за 99,90 долл. Различия между версиями небольшие, однако в Professional нельзя настраивать отчеты.

Недостатки:

- при загрузке 100-Мбайт лог-файла программа Robots.txt Editor and Log Analyzer прикинулась зависнувшим детищем Microsoft, но через некоторое время все-таки ожила.

Сайт программы — <http://net-promoter.com/ru/>.

Robot-Manager Professional Edition 3.1

Эта полнофункциональная утилита состоит из четырех модулей: для работы со списком пауков-роботов и структурой сайта, для визуального создания и редактирования управляющего файла robots.txt и для анализа лог-файлов.

Очень удобна структура организации навигации. Переключаться между модулями вы можете с помощью боковой панели управления. Панель выбора роботов представляет собой таблицу с именами роботов, информацией о них и их пользовательских агентах. Роботов допускается группировать по стране изготовления и языку. Панель запрета, где вы выбираете файлы и каталоги, которые хотите сделать недоступными, очень наглядно и просто позволяет указать путь к нужному каталогу или файлу. Для создания robots.txt необязательно выходить в Сеть, можно работать и с локальной копией сайта.

Программа Robot-Manager позволяет классифицировать роботов по категориям: поисковые сайты, музыкальные сайты, новости.

База данных программы содержит информацию о 71 пауке, что намного скромнее, чем у Robots.txt Editor and Log Analyzer.

Красивый дизайн и простая навигация делают программу Robot-Manager Professional Edition 3.1 пригодной для начинающих веб-мастеров.

Хорошее впечатление оставляет ненавязчивая забота о пользователе — при открытии каждого окна справа появляется краткая инструкция по работе с продуктом (правда, на английском языке).

Robot-Manager поставляется в двух вариантах: Standard Edition за 59,95 долл. и в Professional Edition за 129,95 долл. Немного дороговато, но программа создана в США, где и запросы, и зарплата программистов довольно высоки.

Недостатки:

- англоязычный интерфейс и американская поддержка;
- отсутствие БД роботов большинства российских поисковиков.

Сайт программы — <http://www.websitemanagementtools.com/products.php>.

* * *

Можно было бы рассказывать и о других полезных утилитах, работающих с форматом robots.txt и замечательно анализирующих «сырые» логи, благо выбор огромен. Но когда мы познакомились со множеством сходных программ, то пришли к выводу, что нашего внимания заслуживают лишь Robot-Manager Professional Edition и Robots.txt Editor and Log Analyzer. Все остальные программы требуют либо установки дополнительного ПО, либо проведения части работ вручную.

В профессиональные услуги хороших веб-мастеров и специалистов по продвижению продукта всегда входит работа с файлами robots.txt и анализаторами посещений, что стоит довольно дорого. Поэтому вы не встретите толковых бесплатных утилит, подобных рассмотренным. Но если вы только учитесь на веб-мастера либо категорически не желаете работать с платным ПО, постарайтесь использовать на все 100% первые бесплатные месяцы работы, предоставляемые каждой программой. Нам удавалось ежемесячно переустанавливать ПО и работать заново. Правда, приходилось делать образ системы с помощью программы Drive Image перед началом каждой инсталляции нужного программного инструмента, но это будет темой уже другой статьи.