Всевидящее ОКО

Анализ посещаемости сайта

Специалисты, обеспечивающие техническую поддержку и информационное сопровождение веб-ресурсов, должны уделять значительное внимание анализу лог-файлов журнала доступа веб-сервера.

Даже при сравнительно небольшой (около сотни уникальных хостов в день) посещаемости сайта счет суточного объема логов идет на мегабайты, поэтому о непосредственном анализе таких массивов данных не может быть и речи. В целях автоматизации процесса были разработаны многочисленные программы — анализаторы лог-файлов, к примеру Webalizer, Analog, WebTrends Log Analyzer и т. д.

Несмотря на все многообразие отчетов, предоставляемых каждым из подобных средств, их число в любом случае является ограниченным. К тому же отмеченные анализаторы строят свои отчеты лишь с некоторой заданной периодичностью (как правило, один раз в сутки в часы наименьшей активности посетителей), что не дает возможности наблюдать за посещаемостью в режиме реального времени.

Скрипт ALLA (Artemy Lomov's Log Analyzer) — вполне законченное веб-приложение, осуществляющее расширенный поиск записей в лог-файлах, представленных в комбинированном формате NCSA, по значениям отдельных полей, допуская возможность использования регулярных выражений. Полная версия программы (Perl-код объемом чуть более 20 кбайт) размещена на Chip CD.

Говоря математическим языком, упомянутые выше статистические анализаторы логов (Webalizer и т. п.) предоставляют интегральные отчеты, рассматривающие все факторы активности посетителей в совокупности, тогда как ALLA формирует дифференциальные отчеты о состоянии тех или иных конкретных факторов в данный момент времени. ALLA не претендует таким образом на то, чтобы заменить собой статистические анализаторы; рассматриваемый в этой статье скрипт призван лишь гармонично дополнить их функциональность.

Установка и настройка

Вопреки обыкновению не будем обсуждать код скрипта, поскольку он довольно велик. В то же время хотелось бы отметить, что текст программы снабжен достаточно полными комментариями, так что веб-мастерам, имеющим опыт разработки приложений на Perl, не составит труда разобраться в коде и при необходимости модифицировать его с учетом своих потребностей.

Текущая версия поддерживает только комбинированный (combined) формат лог-файлов сервера, включающий в себя девять полей:

- доменный адрес хоста или IP-адрес клиента;
- информация сервера идентификации identd;
- имя пользователя, примененное при авторизации для доступа в закрытую область узла;
- дата и время обработки запроса;
- имя запрошенного ресурса, HTTP-метод и версия протокола;

»

- код ответа сервера;
- размер ответа сервера;
- реферал (ресурс, с которого пришел посетитель);
- информация об агенте (браузере) пользователя.

Этот формат используется подавляющее большинство виртуальных хостингов. В то же время в Apache по умолчанию принят формат, не содержащий последних двух полей, поэтому обладателям выделенных серверов придется слегка видоизменить конфигурацию.

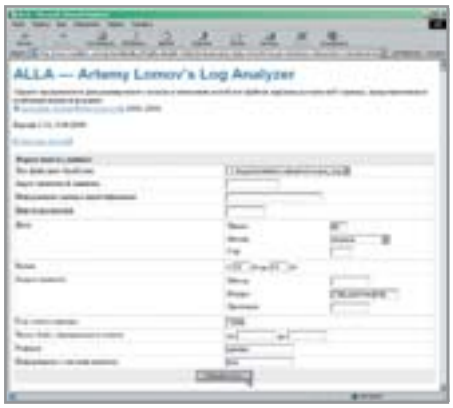
Файл скрипта alla.pl нужно разместить в соответствующей директории сервера (по умолчанию — cgi-bin) и наделить правами доступа 755. В том же самом каталоге должен размещаться текстовый файл allaconf.txt. В нем задается относительный путь к лог-файлу журнала доступа:

```
logfile ../logs/access.log
```

В ряде случаев (к примеру, если в рамках одного аккаунта обслуживается несколько виртуальных серверов) бывает необходимо анализировать множество лог-файлов. ALLA поддерживает и такую возможность — файл allaconf.txt может содержать несколько строк, подобных приведенной выше. При этом имена лог-файлов выбираются пользователем при работе со скриптом в выпадающем списке HTML-формы.

По умолчанию найденные записи выводятся в неформатированном строковом виде. Но можно получить и красивый табличный отчет, если добавить в allaconf.txt строку output extended. При этом следует помнить, что загрузка отчета с расширенным форматированием отнимает больше времени: в особенности это заметно при удаленном доступе при помощи модема.

При выводе отчетов все шестнадцатеричные последовательности, встречающиеся в полях, соответствующих запрошенному ресурсу и рефералу, декодируются до читабельного вида. В длинных URL при табличном оформлении отчетов символ амперсанда («&») отбивается пробелами, чтобы обеспечить нормальный перенос строк в ячейках.



▲ HTML-форма скрипта ALLA. Цвета интерфейса можно изменить

▲ По умолчанию записи выводятся в неформатированном виде

Примеры использования

Основной элемент пользовательского интерфейса приложения ALLA — веб-форма, при помощи которой можно задать один или несколько критериев поиска записей (используется логика «И»). Ниже приведены несколько примеров заполнения полей формы.

Отчет будет отражать HTTP-запросы всех документов, название которых содержит подстроку «index»:

Поле формы	Значение
Ресурс	index

Успешные запросы заглавной и внутренних статичных страниц сайта («/» и файлы *.htm, *.html, *.xhtml, *.shtml):

Поле формы	Значение
Ресурс	(^/$\|\.(s\|x)?htm(l)?$)
Код ответа	200

Успешные запросы динамических (*.cgi, *.pl, *.php) страниц сайта в рабочее время суток:

Поле формы	Значение
Ресурс	\.(cgi|pl|php)
Время	С 9:00 по 17:59
Код ответа	200

Успешные переходы на индексную и внутренние статические страницы сайта со страниц результатов поиска внешних поисковых машин; в отчете будут видны строки поисковых запросов:

Поле формы	Значение
Ресурс	(^/$\|\.(s\|x)?htm(l)?$)
Реферал	%
Код ответа	200

Символ % указывает на наличие в URL закодированных символьных последовательностей, которые с высокой вероятностью являются поисковыми запросами.

Мониторинг скачивания больших документов (объемом от 100 кбайт) в форматах Word, Excel, PDF и ZIP:

Поле формы	Значение
Ресурс	\.(doc|pdf|xls|zip)$
Число байт, переданных в ответе	от 100 000

Все запросы, приведшие к ошибке «404 Not Found» (отчет незаменим для обнаружения «битых» ссылок, в том числе с других сайтов):

Поле формы	Значение
Код ответа	404

Запросы, породившие любую из ошибок стороны клиента или сервера (коды ответа 4xx и 5xx):

Поле формы	Значение
Код ответа	^(4|5)\d\d$

Запросы от имени пользователя admin, породившие ошибку «401 Authorization Required» (отчет позволяет обнаружить попытки проникновения в закрытые области узла):

Поле формы	Значение
Имя пользователя	admin
Код ответа	401

Благодаря широким возможностям регулярных выражений ALLA может генерировать и многие другие весьма специфичные отчеты, недоступные статистическим анализаторам лог-файлов, позволяя веб-мастеру эффективно анализировать активность посетителей, вовремя отслеживать ошибки в работе сайта и осуществлять мониторинг возможных атак.

■ ■ ■ **Артемий Ломов**