Канн Сергей Константинович (ГПНТБ СО РАН)

Достоверность библиотечной веб-статистики

Получение достоверной статистики библиотечных сайтов является далеко не тривиальной задачей. На ухудшение качества веб-статистики влияет масса факторов. В первую очередь назовем прогрессирующее усложнение Интернет-технологий: использование разного рода автоматических средств индексирования и слежения за ресурсами, программ проверки ссылок, перехвата трафика, распространения вирусных сетей («ботнетов»). Качество веб-статистики также в значительной степени зависит от индивидуальных особенностей сайта, используемых технологий представления ресурсов и условий доступа к ним, файловой структуры и установленных защит (фильтров), ограждающих сервер от неправомерной накрутки счетчиков.

Сегодня существует много систем, предлагающих регистрацию пользователей или введение особых проверочных кодов (CAPTCHA), реализующих тест Тьюринга по автоматическому различению компьютеров и людей. Хорошо оснащенные сайты опираются на соответствующий уровень программирования, позволяющий получать относительно достоверную веб-статистику. На этих сайтах применяются идентификационные технологии JavaScript, средства activeX и явa-апплетов для установления более точных характеристик их посетителей. Но, к сожалению, большинство сайтов библиотек из-за недостатков своего развития защищены слабо и не могут «очистить» свою статистику от деятельности ботов. Чаще всего, библиотеки и не ставят такой задачи, исходя из принципов открытости и публичности библиотечной информации. В итоге, они становятся объектами не вполне правомерных действий ботов, имеющих далеко не библиотечные цели, а иногда и вовсе криминальные (например, при DDoS-атаках). Кроме технических неполадок эти действия, как правило, ведут к значительному искажению статистики. Например, за пять суток в конце мая – начале июня 2014 г. с одного из петербургских IP-адресов к серверу Отделения ГПНТБ СО РАН было сделано до 140 тыс. обращений на скачивание файла PDF в 56 мегабайт весом. Эти атаки, которые нельзя расценить иначе, прекратились только тогда, когда были забанены сисадмином. Но ведь они могли быть и не выявлены и тогда цифры, созданные ботом, без корректировки вошли бы в ежемесячную статистику. Аналогичным образом в июне месяце были запрещены свободные доступы к сервису лог-анализа AWStats (он был запаролен), после чего количество обращений к www.prometeus.nsc.ru резко снизилось – в среднем на 58%. Эти примеры показывают, в каких масштабах может искажаться статистика посещений, если не препятствовать «информационному шуму», создаваемому ботами.

С другой стороны, и отказаться от сбора и анализа веб-статистики не представляется возможным. Ее значимость определяется не столько возможностью участвовать в разного рода рейтингах, сколько необходимостью рационального управления – в интересах мониторинга использования ресурсов, определения веб-аудитории сайта, изучения запросов пользователей, оценки эффективности и принятий решений по развитию функционала. Волей-неволей приходится совершенствовать инструменты сбора и анализа веб-статистики, разрабатывать критерии повышения достоверности статистической информации и отсеивания искажений. Не последнюю роль в улучшении показателей могут сыграть системы Google Analytics, LiveInternet, Я-Метрика и др., продвигающие веб-анализ деятельности сайтов на основе системы поведенческих факторов. Хотя и они далеко не идеальны и было бы ошибкой доверять им в полной мере. Важное значение имеет анализ лог-файла сервера (журнала доступов и ошибок), позволяющий извлечь сведения о некорректных реферерах, запросах с лишними «слэшами», непонятных юзер-агентах и прочих деталях трафика, идентифицирующих ботов. В любом случае, проблему анализа веб-статистики библиотечных сайтов придется решать очень долго, тщательно и индивидуально для разных сайтов и ресурсов. Наконец, бесспорно и то, что проблему достоверности можно решать лишь комплексно, с использованием всех возможных средств сбора и анализа статистики.