

Оценка работы библиотеки с помощью web-метрик
Assessment of Work of Library Using Webmetrics
Оцінка роботи бібліотеки за допомогою web-метрик

Е. В. Ковязина

*Институт вычислительного моделирования СО РАН,
Красноярск, Россия*

Elena Kovyazina

*Institute of Computational Modelling
Siberian Division of the Russian Academy of Sciences,
Krasnoyarsk, Russia*

О. В. Ковязіна

*Інститут обчислювального моделювання СВ РАН,
Красноярськ, Росія*

В докладе представлен опыт применения web-метрик в практике работы библиотек Красноярского научного центра Сибирского отделения РАН. Определены некоторые правила применения анализа log-файлов для оценки популярности электронных ресурсов. Обозначены рекомендации для статистической оценки собственных информационных ресурсов и лицензионных онлайн-баз данных внешних поставщиков.

The paper describes the experience of application of webmetrics in practice of the libraries of the Krasnoyarsk Scientific Center of the Siberian Branch of the RAS. Some of the rules of analysis of log-files for assessment of popularity of electronic resources are defined. Recommendations for statistical evaluation of internal information resources and licensed online databases of external providers are identified.

У доповіді представлено досвід застосування web-метрик в практиці роботи бібліотек Красноярського наукового центру Сибірського відділення РАН. Визначено деякі правила застосування аналізу log-файлів для оцінки популярності електронних ресурсів. Подані рекомендації для статистичної оцінки власних інформаційних ресурсів та ліцензійних онлайн-баз даних зовнішніх постачальників.

Приметой времени является неуклонный и всё ускоряющийся переход библиотек в электронную среду. Если сплошную оцифровку фондов библиотек тормозит, в определенной степени, законодательство о защите авторских прав, то ничто не мешает библиотекам перемещать в Интернет собственные ресурсы и услуги – от электронного каталога до виртуальных библиографических справок, от методических рекомендаций до семинаров повышения квалификации. Онлайн-доступ к электронным журналам и книгам безгранично расширил возможности комплектования, а растущая компьютерная грамотность читателей увеличивает их потребность в удаленном обслуживании. Вся эта работа требует значительных трудовых усилий специалистов различного профиля, а также дополнительных финансовых затрат. С развитием автоматизации и выходом библиотек в Интернет эти затраты растут, а количество персонала не только не сокращается, напротив, увеличивается, к тому же значительно выросли требования к его квалификации. Являются ли эти усилия оправданными? Насколько востребованы читателями предлагаемые библиотеками услуги? Как оценить работу библиотеки в электронной среде в общепринятых терминах посещаемости и обращаемости фонда, если читатели отделены от библиотеки километрами расстояний, проводов и компьютерной техники?

Для оценки эффективности Интернет-сайтов всемирная паутина использует web-метрики, или э-метрики. Существует масса руководств по улучшению организации и дизайна web-сайта, повышению продаж в Интернет магазине на основе анализа web-метрик, например, [1]. Применение web-метрик для оценки использования онлайн-ресурсов в зарубежных библиотеках наиболее полно освещены в монографии Э. Уайта и Э. Д. Камаля [2]. Можно ли оценить качество и востребованность собственных электронных ресурсов библиотеки, руководствуясь рекомендациями авторов и имея минимум специальной подготовки. Такие попытки были предприняты в сети

библиотек Красноярского научного центра Сибирского отделения РАН (КНЦ СО РАН) при подготовке годового отчета за 2011 год, а также в дальнейшей работе.

Каждая библиотека сети КНЦ СО РАН располагает следующими электронными ресурсами, подлежащими анализу:

1. Основной web-сайт библиотеки, как правило, многостраничный, обслуживаемый библиотекой самостоятельно и размещенный на собственной компьютерной базе.

2. Web-страница электронного каталога и библиографических баз данных библиотеки, содержащая Web-ИРБИС и размещенная на одном из двух согласованно работающих серверов. Статистические данные по каждой такой странице предоставляет библиотека-владелец сервера.

На основных сайтах большинства библиотек установлены простые бесплатные счетчики посещений, такие, например, как Openstat (ранее, Spylog), позволяющие вести подсчет хитов и визитов и получать отчет по ним за определенный период времени. Для годовых цифровых отчетов библиотеки этих данных вполне достаточно, к тому же такие счетчики позволяют участвовать в различных рейтингах Интернет. Однако, для того, чтобы составить полную картину использования сайта, включая «портрет» читателя, этих данных недостаточно. Интернет предлагает готовые инструменты анализа сайта, наиболее известные из которых Google Analytics и Яндекс. Метрика. Оба этих инструмента имеют свои достоинства и недостатки [2], дают полную информацию о визитах, просмотрах и посетителях, времени, проведенном на странице и времени, проведенном на сайте, отказах, и даже поминутном трафике на сайте. С помощью Google Analytics также можно определить перечень страниц и каталогов сайта, индексируемых либо неиндексируемых роботом поисковика. Однако эти инструменты предназначены, в основном для коммерческого использования, результаты анализа направлены на привлечение на сайт потенциальных покупателей и техническую поддержку сайта. Инструменты требуют вмешательства в коды всех страниц сайта, регистрации и работают с отчуждаемыми данными. А Google Analytics предполагает при этом еще и значительную специальную подготовку, хорошее понимание используемых в web-аналитике терминов. Проверка организации сайта библиотеки ИВМ СО РАН с помощью Google Analytics дала хорошие результаты, которые будут таковыми всегда, когда выполняются следующие правила организации сайта, вполне приемлемые для библиотеки:

а) количество переходов (щелчков) для достижения любого раздела сайта минимально – нет малоинформативных и повторяющихся разделов;

б) любая страница (или связанный набор фреймов) размером не более величины экрана, информационная насыщенность страницы регулируется количеством фреймов;

в) все графические элементы имеют встроенные комментарии;

г) нет никаких ограничений на доступ к разделам и документам.

При выполнении этих правил **посещение** и **просмотр** страницы не различаются, статистика несколько упрощается. **Хиты** включают все графические элементы страницы (части логотипов, рисунки, баннеры, графические разделители элементов страницы и т.п.). Их анализ не имеет особого смысла для содержательного наполнения сайта и из статистики исключается. Если нет ограничений на просмотр каких-либо частей сайта, и нет ошибок в кодах страницы, то в статистике отсутствуют **отказы**. Таким образом, для оптимизации структуры сайта, устранения ошибок в кодах и стандартизации сайтов инструменты Google Analytics очень эффективны.

Более подробную информацию о качестве использовании сайта, составе визитеров и т.д. можно получить из анализа файлов регистрации web-сервера или log-файлов. Log-файлы формируются программным обеспечением web-сервера для собственных технических нужд, и фиксируют все транзакции на сайте. Для анализа регистрационных файлов в Интернете предлагается множество готового программного обеспечения, как платного, так и свободно распространяемого. Большая часть бесплатного программного обеспечения не имеет инструкций по использованию и не обладает удобным пользовательским интерфейсом. После ряда неудачных опытов со свободно распространяемыми программами для детального анализа log-файла web-сервера Apache сайта электронного каталога библиотеки ИВМ СО РАН была использована платная программа Web Log Storming Professional (версия с ограниченным временем жизни). С помощью этого инструмента был проведен анализ log-файлов первого квартала 2012 года. Анализ был проведен для того, чтобы составить представление о возможностях программного обеспечения, составе трафика и визитерах. Наряду с

графиком тенденций в посещениях (рис. 1), хитах и пропускной способности сайта, программный продукт предоставляет проранжированные списки предпочитаемых страниц для входа и выхода визитеров, наиболее предпочтительных документов, ошибок и не найденных документов и т.д. Кроме того, формируются прорисованные в виде диаграмм распределения визитеров по странам, городам и регионам мира (рис. 2).

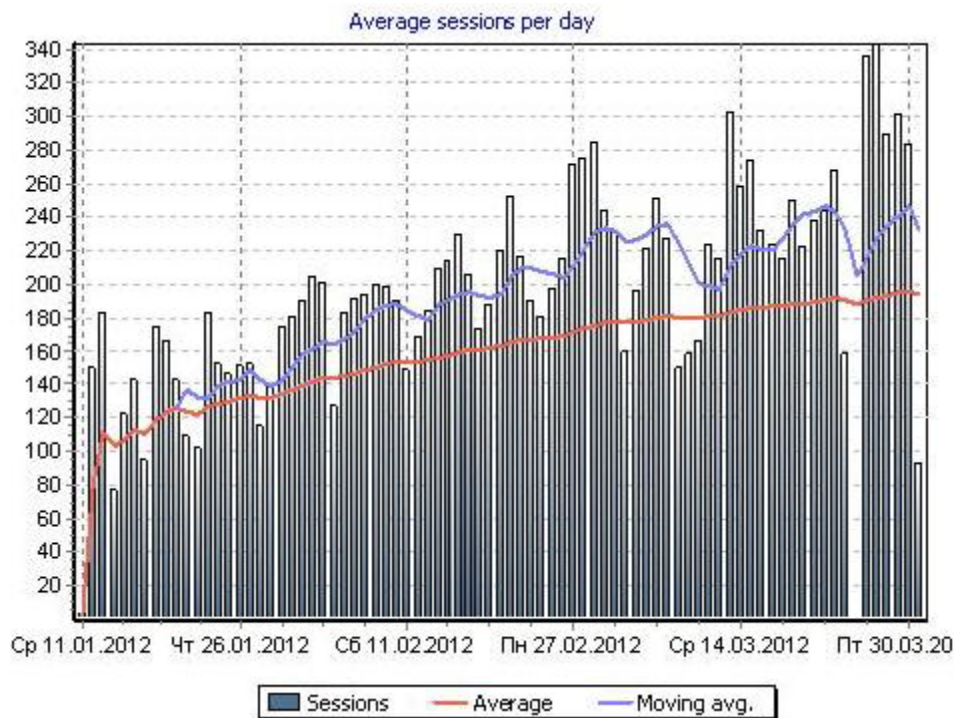


Рисунок 1. Динамика посещений сайта электронного каталога библиотеки ИВМ СО РАН

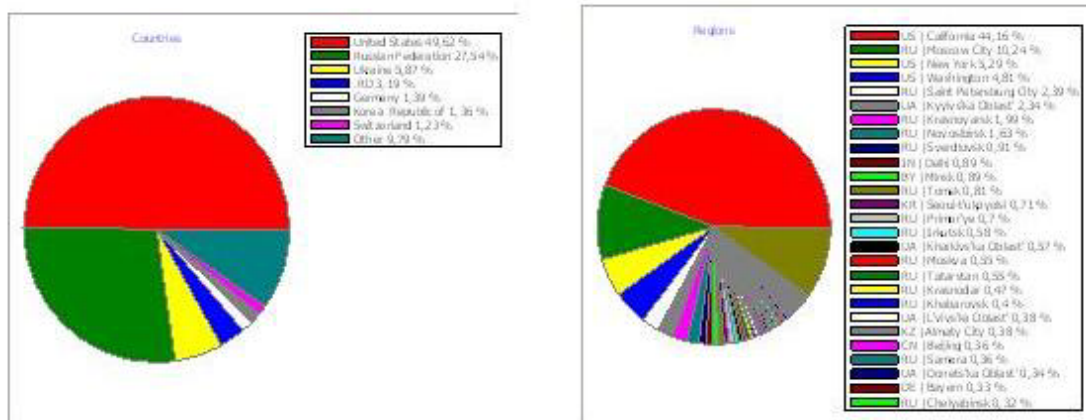


Рисунок 2. Распределение посещений по городам и странам мира

Круговые диаграммы наглядно показывают долю в посещениях роботов поисковых систем – большая часть диаграммы, окрашенная красным, и ИРБИС-корпорации – участок зеленого цвета.

Информация об адресах, с которых приходили визитеры, а также сравнение статистик посещений основного сайта, поисковой системы электронного каталога и данных провайдеров онлайн-новых лицензионных ресурсов позволили сделать вывод о том, что web-сайт библиотеки не является наиболее популярным отправным пунктом для доступа не только к удаленным онлайн-ресурсам, но и к электронному каталогу библиотеки. Пользователи предпочитают поиск с помощью крупных Интернет-поисковиков, подобных Google или Yandex. Как следствие, использование технологии переходов для сбора статистики обращений к ресурсам, детально излагаемой в [2], представляется нецелесообразным.

Для более подробного анализа использования собственных электронных ресурсов библиотеки был проведен детальный анализ лог-файлов Web-ИРБИС за 2011 год и базы данных статистики ИРБИС за 1 квартал 2012 года. Данные посещений, извлеченные из лог-файлов, приведены в таблице 1.

Таблица 1.

Структура запросов к базам данных Web-ИРБИС

Библиотека сети	Запросов, всего, тыс.	Роботы-индексаторы, тыс.	ИРБИС-корпорация, тыс.	Внешние IP-адреса, тыс.	Из них. локальная сеть, тыс.
ИВМ СО РАН	3634	553	3070	10,6	0,3
ИЛ СО РАН	4008	1583	2401	24,9	1,9
ЦНБ КНЦ СО РАН	1496	0	1491	5,3	0,2
ИФ СО РАН	1903	0	1889	13,9	9,8
ИБФ СО РАН	1079	0	1075	3,2	0,1
ИХиХТ СО РАН	1214	0	1208	6,6	2,2

Содержимое таблицы показывает, что сервера сети КНЦ СО РАН, поддерживающие электронные ресурсы библиотек, проводят различную политику по отношению к роботам-индексаторам поисковых систем Интернет. Один сервер регистрирует все вхождения поисковых роботов, на другом – индексация роботами запрещена. Были предприняты попытки отследить особенности работы с отключенными роботами и их влияние на статистику посещений библиотечных ресурсов. Проведена серия несложных поисковых опытов в Google и Yandex с последующим просмотром того как результаты запросов регистрируются в лог-файлах и базе данных статистики ИРБИС. Запрос на список трудов сотрудника в библиотеке по месту его работы регистрируется как поступивший с IP поисковика, если результирующий список уместается в страницу, и пользователь не производит с ней никаких действий, таких как прокрутка или листание. Если такие действия производятся, регистрируется IP-адрес пользователя. Список трудов вообще невозможно получить, если каталоги библиотеки не индексируются поисковой машиной. Отключение индексирующих роботов по их уникальным IP-адресам затруднительно, так как поисковые системы все время меняют не только конкретные адреса роботов в пределах отведённой таким системам сетки адресов, но и всю сетку, например, 95.198.245.* – Яндекс, 74.125.*, 66.249.*, 209.85.* – Google, 208.115.113.* – по видимому, Yahoo. Можно запретить индексацию для всей сетки, теряя при этом некоторую часть посетителей, пришедших из поисковых систем. В то же время адреса поисковиков легко отслеживаются при анализе посещений в базе данных статистики Web-ИРБИС, и могут быть легко удалены из общей статистики.

Таблица 2.

Обращения к базам данных Web-ИРБИС на сайте ИЛ СО РАН

Название БД	Содержание БД	Кол-во запросов, в тыс.	В процентах от общего числа запросов к БД	Запросы с внешних IP в тыс.
FORAD	Авторефераты диссертаций	330,19	16,19	3,38
FORDI	Диссертации	313,53	15,38	3,21
FORBK	Книги русскоязычные	751,8	36,87	7,69
FORIN	Книги иностранные	415,03	20,36	4,25
FORPR	Труды сотрудников	40,705	2,00	0,42
FORJ	Иностранные журналы	11,69	0,57	0,12
PRIR	Тематическая	35,42	1,74	0,36
FORCC	Реферативная	33,95	1,67	0,35
NAUKA	Тематическая	18,06	0,89	0,18
SEVER	Тематическая	34,23	1,68	0,35
FORSF	РСФ труды	23,555	1,16	0,24
RUSJ	Отечественные Журналы	11,025	0,54	0,11
ECON	Тематическая	19,67	0,96	0,20

По-видимому, если позволяют мощности сервера, индексацию каталогов не стоит отключать абсолютно. Можно отключать их периодически при большой загрузке сервера. Постоянно следует закрывать для индексирования лишь отдельные информационные ресурсы библиотеки, имеющие локальную ценность. В качестве примера приведем статистику обращений к информационным ресурсам библиотеки Института леса им. В.Н.Сукачева СО РАН (таблица 2). Даже такая простая статистика показывает наиболее востребованные информационные ресурсы, а также ресурсы, имеющие ограниченную ценность (только для специалистов института), индексацию которых можно запретить.

ИРБИС-корпорация также дает существенную часть статистики посещений, из которой визуально не выделяются робот и запросы реальных пользователей. Поиск «на лету», принятый в корпорации по умолчанию, приводит к тому, что каждый поисковый запрос трансформируется в несколько запросов (от 2 до 15 в зависимости от длины термина). Следует задуматься о необходимости такого сервиса, вес которого в статистике будет только расти с расширением числа участников корпорации.

В качестве заключения отметим, что все методы сбора данных дают только приблизительные статистические оценки. Точных методов оценки эффективности работы в Интернет, по-видимому, не существует. Отчасти это объясняется недостаточной строгостью определений web-метрик. Например, такой термин как «сессия», интуитивно понятный, в глобальной паутине не может быть определен строго. И в различных аналитических инструментах под сессией понимают разные интервалы времени пребывания посетителя на сайте – примерно от 10 до 30 минут. Показатели числа сессий при этом могут сильно различаться. Такое различие толкований допускают едва ли не все web-метрики. Дополнительный смысл статистические данные принимают при исследовании сравнительных характеристик в «сообществах» Интернет-сайтов, например, на множестве сайтов академических институтов или университетов. Но эти вопросы являются темой других исследований [4].

Литература

1. Кошик, Авинаш. Веб-аналитика: анализ информации о посетителях веб-сайтов = Web Analytics: An Hour a Day./ А.Кошик. – М.:Диалектика, 2008. – 464 с.
2. Уайт, Эндрю. Статистические методы работы с электронными документами в библиотечной сфере, или Э-метрики: как использовать данные для управления и оценки электронных ресурсов и фондов : монография / Э. Уайт, Э.Д. Камаль. – М.: Омега-Л, 2006. – 393 с.
3. Гутникова А. Веб-аналитика в сравнении: Google Analytics и Яндекс.Метрика / А.Гутникова // Практика Интернет маркетинга. № 6. – 2009.
4. Печников, А.А. Структурные исследования научного веба [Электронный ресурс] / А.А.Печников, Н.Б.Луговая // XVI всероссийская научно-методическая конференция «Телематика-2009», 22–25 июня 2009 г., г. Санкт-Петербург: материалы семинара. – СПб., 2009. – URL: <http://tm.ifmo.ru/tm2009/src/018c.pdf>