БИБЛИОСФЕРА, 2009, № 3, с. 56–58

Информатика

УДК 025.4.026 : 004 + 002.513.5 : 004
ББК 78.37 + 73 + 78.30

ОСОБЕННОСТИ ИНДЕКСИРОВАНИЯ РЕСУРСОВ БИБЛИОТЕЧНОГО САЙТА РОБОТАМИ ПОИСКОВЫХ МАШИН

© С. К. Канн, 2009

Государственная публичная научно-техническая библиотека Сибирского отделения Российской академии наук
630200, г. Новосибирск, ул. Восход, 15

Изучение особенностей процесса индексирования документов крупнейшими поисковыми машинами позволяет точнее определить место ресурсов библиотечного сайта в общем веб-пространстве. Автор анализирует доступ роботов к двум типам ресурсов, обновляемых на постоянной и нерегулярной основе. Результаты тестирования сайта www.prometeus.nsc.ru показывают значительные расхождения в поведении отечественных и зарубежных поисковых машин.

Ключевые слова: Интернет, библиотечный сайт, веб-документы, поисковые машины, Гугл, Яндекс, индексирование библиотечных ресурсов.

The study of special features of the indexing process by the largest search engines allows to determine the global position for library resources in the general www-space. The author analyses the robots' access to two types of website documents with constant or irregular renovations. The result of testing, based on the materials of www.prometeus.nsc.ru, reveals essential difference between the main home and foreign search engines.

Key words: Internet, library website, web documents, search engines, Google, Yandex, indexing of library resources.

Основной поток посетителей сайта отделения Государственной публичной научно-технической библиотеки Сибирского отделения Российской академии наук (до 80–90%) формируется за счет обращения к глобальным поисковым машинам Google, Yahoo, MSN, Рэмблер. Эта четверка лидирует и в мировом информационном пространстве, далеко опережая своих конкурентов. Вместе с тем по мировой паутине бродит масса других программ-роботов, взаимодействие с которыми требует непрерывного совершенствования процессов поисковой оптимизации веб-сайтов (*search engine optimization – SEO*).

За девять месяцев 2008 г. число роботов, приходы которых отметил веб-сервер отделения, превысило полсотни, однако эта цифра далеко не окончательная, так как почти 2/3 роботов посещают ресурсы библиотеки «нелегально», обходя стандартную процедуру обращения к файлу *robots.txt*. Так называемый «неотображаемый трафик», сгенерированный роботами и ответами сервера со специальным http-кодом, за указанный период достиг 39 гигабайт. Масштабы аккумулирования информации огромны. Только две крупнейшие поисковые системы Google и Yahoo сделали к сайту отделения почти по миллиону доступов каждая и суммарно скачали свыше 22 гигабайт информации, что в 47 раз превышает весь объем ресурсов, накопленный на сайте *www.prometeus.nsc.ru* за одиннадцать лет работы.

Для того чтобы повысить эффективность отдачи от этих ресурсов и продолжить дальнейшее расширение аудитории пользователей, в первом полугодии 2008 г. было проведено изучение особенностей индексирования сайта роботами основных поисковых машин (таблица, с. 57). Учитывая сезонную «волнообразность» притока посетителей, изучаемый период охватил как «восходящую» линию обращений (с января по май), так и ее нисходящий тренд (с мая по июль). На это важно обратить внимание в связи с тем, что, по данным статистики сервера, приход лета означает почти четырехкратное падение посещаемости как по числу посетителей, так и по запросу страниц. В этот период выставляется меньше новой информации, реже редактируются старые документы, ослабевает пользовательская нагрузка на сервер. Между тем для индексирования страниц роботами такое затишье, наоборот, позволяет извлечь некоторые преимущества, поэтому в их отношении летнего спада посещений не наблюдается (таблица, с. 57).

Специфика сайта *www.prometeus.nsc.ru* заключается в его продолжительном существовании в сети и длительной раскрутке ресурсов. Нам показалось

Индексирование сайта отделения ГПНТБ СО РАН роботами основных поисковых машин (количество доступов за 9 месяцев 2008 г.)*

Роботы	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Итого за 9 мес.
Googlebot	125 810	70 591	79 735	73 431	115 534	135 235	144 958	116 370	169 107	1 030 771
Yahoo Slurp	74 760	55 423	74 006	133 754	133 433	118 899	115 915	87 238	61 508	854 936
StackRambler	35 080	19 725	25 623	26 797	25 788	24 907	27 616	33 619	31 043	25 0198
MSNBot	30 452	26 158	26 378	33 437	55 244	51 860	71 336	38 036	65 349	398 250
Yandex bot	10 725	20 770	32 633	15 697	21 686	33 199	59 748	61 215	16 2738	418 411
Turn It In	1 574	–	3 492	1 676	2 637	3 233	4 452	1 352	6 288	24 704
Alexa (IA Archiver)	13 86	886	851	934	1 207	1 210	155	217	1 290	8 136
BaiDuSpider	1 279	1 401	1 515	1 500	1 542	1 462	1 682	1 516	1 464	13 361
AskJeeves	729	1546	1 826	964	1 120	1 265	1 050	621	257	9 378

* По данным статистической системы AWStats веб-сервера отделения ГПНТБ СО РАН.

интересным проследить, как реагируют роботы на два типа обновлений: 1) ресурсов, актуализируемых постоянно, и 2) ресурсов, возникающих или пополняемых нерегулярно. В первом случае речь шла о документах еженедельной выставки новых поступлений, существующей с 23 октября 1997 г., о материалах дайджеста «Российская наука и мир» (выставляется с февраля 1998 г.), новостях библиотеки и ее подразделений, проектах, получивших грантовую поддержку, и о ряде других устойчиво развиваемых ресурсов. Во втором – об эпизодически возникающих библиографических списках и указателях, оглавлениях книг, трудах сотрудников и партнеров библиотеки. Кроме того, изучалось обращение роботов к абсолютно новым (внезапно появившимся) комплексам документов, например посвященным созданию Клуба библиотекарей, посещению Кемеровской областной библиотеки 29 мая 2008 г. в рамках Всероссийского дня библиотек или ряду других инициатив, пока не ставших традицией.

Первые два десятка поисковых машин индексируют сайт почти непрерывно – об этом свидетельствуют ежедневные визиты их разведчиков-«ботов». Помимо уже называвшейся четверки машин наибольшую активность проявляют Turn It In, BaiDuSpider, AskJeeves, Alexa (IA Archiver), Lycos. Но все они на порядок уступают лидерам индексирования и поиска. Так, роботы Гугла (Googlebot 2.1) отличаются тем, что ведут «плотный» мониторинг давно существующих ресурсов, отслеживая формальные и содержательные обновления, вливание новой информации, появление новых документов, расширяющих рамки ресурса. Проведенные тесты показали, что Google является пионером в индексировании новых файлов выставки новых поступлений и, вообще, является единственным, кто индексирует эту выставку «неделя в неделю». У остальных поисковых систем задержка индексирования достигает трех и более недель. Как правило, Google хранит и самые свежие копии прежних выставок. Yahoo, немного уступающая Гуглу по скорости отражения новой информации, к сожалению, не предлагает, как ее коллега, сервиса кэширования (сохраненных копий).

Вместе с тем роботы Yahoo (Slurp и Slurp 3.0) обнаружили завидную мобильность в выявлении новых, нерегулярно возникающих документов. Они умудрялись индексировать неожиданно выставляемые документы (в том числе в абсолютно новых, недавно созданных директориях) уже в день их появления на сервере или на следующие сутки. Частота дальнейших визитов роботов Yahoo в разы превышала показатели всех остальных конкурентов. В отношении «нерегулярных» обновлений Yahoo опережала Google примерно на сутки. Еще одни сутки уступали Yahoo роботы Рэмблера (StackRambler 2.0). Их не выручало даже то, что на страницах сайта отделения установлен код баннерного проекта Rambler's Top 100 (id=474349). Казалось бы, Рэмблеру стоило использовать это преимущество для своевременного индексирования страниц, но этого почему-то не происходило. В отношении же MSN (Live Search) можно заметить, что задержка прибытия роботов этой поисковой системы по сравнению с «пионерами индексирования» (Yahoo, Google и Рэмблером) временами достигала целого месяца.

Особый интерес имело изучение взаимодействия поисковых машин с библиографическими указателями по актуальным проблемам естествознания, техники, технологии, экологии и пр., составленными партнером и почетным читателем библиотеки А. П. Зарубиным. Первый подобный указатель появился на сайте еще 9 февраля 1999 г. С тех пор два десятка работ (примерно 90 вебстраниц) аккумулировали до 12 тыс. библиографических записей. В 2008 г. был подготовлен и выставлен очередной указатель, посвященный современным подходам к Периодической системе Д. И. Менделеева. Всего через 12 часов после установки на сервер оба текстовых файла указателя проиндексировали роботы Гугла (в ночь на 4 марта), а через сутки – Yahoo. Робот Рэмблера пришел только 13 марта. До конца июня указатель посещали роботы Yahoo (118 раз), Google (74), MSN (31) и Рэмблера (20). С 17 марта документы указателя стали присутствовать в поисковых выдачах Гугла, число которых к концу июня достигло 204. Из числа остальных 120 выдач на долю Рэмблера пришлось 42 и MSN – 23 (остальные поиски велись другими поисковиками – nigma.ru, elementy.ru, etc.). За четыре месяца на указатель было сделано не менее 57 закладок в браузерах (подсчитано по вызову файла *favicon.ico*).

Крайне неожиданными оказались результаты тестирования поисковой машины Яндекса (на середину июля 2008 г.). Выяснилось, что за все первое полугодие 2008 г. роботы Яндекса проиндексировали не более 15% новых документов, созданных с января по июнь включительно. Последнюю индексацию свежих документов Яндекс провел в начале мая, а все остальное время его роботы многократно «перелопачивали» давно известные, ранее созданные страницы. При этом совершенно игнорировались целые массивы новой информации, такие как ВНП за последние 14 недель (до середины июля), файлы книжных оглавлений, дайджест «Российская наука и мир» (с января по апрель – более свежие еще не выставлены на сайт), новые документы проекта «Научные школы Новосибирского научного центра» (материалы об академиках В. В. Болдыреве и В. Н. Пармоне) и т. д.

Задержки в индексировании новых ресурсов Яндексом заметили не только мы. Под другим углом зрения об этом говорится в статье сотрудников Института вычислительных технологий СО РАН [1, с. 129]. Можно сделать вывод, что слоган корпорации Яндекс («*со временем найдется всё*») слишком вольно трактует цену времени. По нашему мнению, в данный момент база Яндекса не может считаться самым актуализируемым отечественным массивом данных, каким он был еще совсем недавно. Возможно, разработчики меняют поисковый алгоритм или слишком увлеклись созданием новых сервисов, но в сети шутят, что «на Яндексе в сохраненных копиях болтаются версии страниц, написанных еще с ятями» [2].

Подводя итог всему сказанному, нужно отметить, что изучение специфических особенностей процесса индексирования веб-документов роботами крупнейших поисковых машин дает возможность точнее определить место ресурсов отделения ГПНТБ СО РАН в общем www-пространстве и продолжить целенаправленную работу по переходу на новую технологию сайтостроения, подразумевающую диверсификацию сайта библиотеки.

Список литературы

1. Рейтинг сайтов научных организаций СО РАН / Ю. И. Шокин [и др.] // Вычисл. технологии. – 2008. – Т. 13, № 3. – С.128–135.
2. Борьба с роботами [Электронный ресурс]. – URL : http://www.klim.by/Borba-s-robotami.87.0.html. – 17.07.2008 г.

Материал поступил в редакцию 10.04.2009 г.

Сведения об авторе: *Канн Сергей Константинович – старший научный сотрудник лаборатории информационных ресурсов отделения ГПНТБ СО РАН, тел.: (383) 330-61-60, e-mail: serge@prometeus.nsc.ru*