УДК 004.738.5

ОБ ИЗМЕРЕНИЯХ ВЕБОМЕТРИЧЕСКИХ ИНДИКАТОРОВ

Печников А.А.

ФГБУН «Институт прикладных математических исследований Карельского научного центра Российской академии наук», Петрозаводск, e-mail: pechnikov@krc.karelia.ru

На примере вебометрического индикатора «размер сайта», измеряемого для множества веб-сайтов Российской академии наук, рассмотрен вопрос об ошибочных измерениях, полученных в результате использования поисковых систем. Предложим подход к «сглаживанию» ошибок поисковых систем, основанный на использовании результатов сканирования сайтов с помощью авторской программы BeeCrawler.

Ключевые слова: вебометрика, веб-сайт, вебометрическое ранжирование, индикаторы

MEASUREMENTS OF WEBOMETRIC INDICATORS

Pechnikov A.A.

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, e-mail: pechnikov@krc.karelia.ru

The issue of erroneous measurements obtained from the use of search engines was considered in the case of webometric indicator «site size» measured for a set of websites of the Russian Academy of Sciences. An approach to «smoothing» the errors of search engines, based on results obtained from scanning of sites using author's program BeeCrawler is offered.

Keywords: webometrics, web site, webometric ranking, indicators

Еще в 2005 году в работе [1] было отмечено, что исследования Веба зачастую основываются на данных, полученных, если можно так выразиться, с помощью самого Веба, и, в частности, с использованием возможностей наиболее распространенных поисковых систем. Однако, поисковые системы, являясь коммерческими проектами, ориентированы на некоего условного «среднего» пользователя (с возможностью адаптации к его запросам), а не на исследователя.

Критические публикации на тему использования поисковых машин в качестве средств измерений появились достаточно давно [2-4]. Например, в работе [2] показано, что для подобранных конкретных примеров Google (по неизвестным причинам) «скрывает» от 48 до 70% проиндексированных им же страниц, содержащих ссылки на заданный сайт. Даже в случае очевидной ошибки в результатах вывода по запросу к поисковой системе мы не получим ответа на вопрос о том, почему эта ошибка произошла. Это, однако, не останавливает исследователей, имеющих в качестве «измерительных устройств» только поисковые системы [6, 7].

В процессе работы по проекту «Вебометрический рейтинг научных учреждений России» [5] в процессе измерений значений вебометрических индикаторов авторам пришлось столкнуться с возникновением ситуаций, которые нельзя назвать иначе как ошибками поисковых систем. Наиболее явно такие ситуации проявились при измерениях с помощью Google индикатора, характеризующего размер сайтов, когда получаемое количество страниц выражалось сотнями тысяч для сайтов весьма скромных размеров (об этом достоверно можно судить из других источников).

Поисковые системы как «измерительные устройства» обладают низкой надежностью, измерения, проведенные в одинаковых условиях, не всегда дают согласующиеся результаты [1]. Это объяснимо, если измерения проводятся через большие интервалы времени (сказывается динамика Веба), но неприемлемо в тех случаях, когда причин для больших расхождений не видно. Измерения индикаторов в рамках проекта [5] проводятся в течение компактного временного интервала (в течение двух недель) и дальнейшая обработка полученных значений не подразумевает новых замеров при обнаружении предполагаемых ошибок, поскольку в этом случае нужно заново измерять все индикаторы.

Однако ни в первом проекте по ранжированию веб-сайтов [8], ни в других проектах [6, 7], вопросу об ошибочных измерениях вебометрических индикаторов должного внимания не уделяется. В данной статье мы предложим подход к «сглаживанию» ошибок поисковых систем при измерениях размеров сайтов.

Вебометрические индикаторы. Общеизвестны основные вебометрические индикаторы, используемые при решении задач ранжирования веб-ресурсов. Можно считать, что 4 индикатора, введенные

ещё в 2004 году в проекте [8], не вызывают принципиальных возражений и в настоящее время:

• размер сайта (*S*, size) – общее количество страниц,

• видимость сайта (*V*, visibility) – количество гипертекстовых ссылок с других веб-ресурсов,

• количество полнотекстовых файлов (*R*, «rich files», т.е. файлов с расширениями doc, pdf, ppt и т.д.),

• научность сайта (*Sc*, «scholar») – количество ссылок на сайт, обнаруживаемых Google Scholar.

Для первых трёх индикаторов следует добавить фразу «обнаруживаемых с помощью Google/Яндекс/ ...». Отметим также, что разработчики проекта [8] недавно анонсировали использование в качестве индикаторов данных, предоставляемых коммерческими системами Ahrefs Pte Ltd (https://ahrefs.com) и Majestic SEO (http://www.majesticseo.com), чем еще больше усугубили ситуацию в плане прозрачности и отсутствия коммерческих влияний [1].

В проекте «Вебометрический рейтинг научных учреждений России» для сбора значений вебометрических индикаторов используются поисковые системы Яндекс и Google и специализированная программа сбора внешних гиперссылок BeeCrawler [9].

Измерения вебометрических индикаторов. В проекте [5] на сегодняшний день проведены измерения индикаторов примерно для 400 сайтов РАН (далее количество сайтов будем обозначать *N*).

Пусть S_Y^i, S_G^i, S_B^i – размеры *i*-го сайта по Яндексу, Google и BeeCrawler, а V_Y^i – видимость *i*-го сайта по Яндексу (она нам понадобится в разделе «Сглаживание ошибок»).

Приемы измерений поясним на примере сайта Российской академии наук (www.ras.ru). в Google на запрос вида «site:www.ras.ru» (текст набирается в поисковой строке, кавычки не нужны) будет выдан ответ «Результатов: примерно 513000». Это значение далее и будет принято в качестве S_G^i для сайта www.ras.ru (более точно – это значение S_G^1, поскольку сайт РАН имеет порядковый номер 1).

В Яндексе для получения количества страниц также можно использовать запрос вида «site:www.ras.ru». Ответ «Нашлось 85 тыс. ответов» округлен, хотя и его можно использовать в качестве приближенного значения S_Y^1. Более удобным представляется использование специального сервиса Яндекс.XML, для чего необходимо зарегистрироваться в системе. Здесь запрос о количестве страниц на www.ras.ru выглядит следующим образом:

http://xmlsearch.yandex.ru/xmlsearch?text=site:www.ras.ru&user=USER&key=KEY,

где USER и KEY – логин и ключ пользователя. В одной из строк развернутого ответа на запрос будет строка <found−docs−human>нашёл 85145 ответов</ found−docs−human>, и 85145 будет являться значением индикатора S_Y^1.

Удивляться существенному расхождению в результатах разных поисковых систем не следует: нам неизвестны правила отбора страниц на сайте, принятые в поисковых системах, и, по-видимому, в Google и Яндексе они различные. При последовательном просмотре результатов вывода Google вскоре выдаст информацию «Мы скрыли некоторые результаты, которые очень похожи на уже представленные выше (683)», то есть проверяемыми являются только 683 результата из 513000 заявленных, а что собой представляют остальные результаты проверить не удастся. Яндекс также выдает не все, а только 1000 ответов по запросу «site:www.ras.ru».

В BeeCrawler реализован порядок обхода страниц «вначале вширь»: сканируется начальная страница нулевого уровня, находятся страницы первого уровня, сканируются страницы первого уровня и находятся страницы второго уровня и т.д. В процессе сканирования создается вспомогательная таблица количества страниц на каждом уровне, т.е. мы имеем

$$S_B^i = S_B^{0,i} + S_B^{1,i} + S_B^{2,i} + ... + S_B^{j,i} + ... + S_B^{M,i},$$

где $S_B^{j,i}$ – количество страниц на *j*-м уровне *i*-го сайта, а *M* – номер наибольшего сканируемого уровня, устанавливаемого при запуске программы (в последней версии сканирования *M*=7).

В табл. 1 приводятся значения индикаторов для нескольких веб-сайтов РАН.

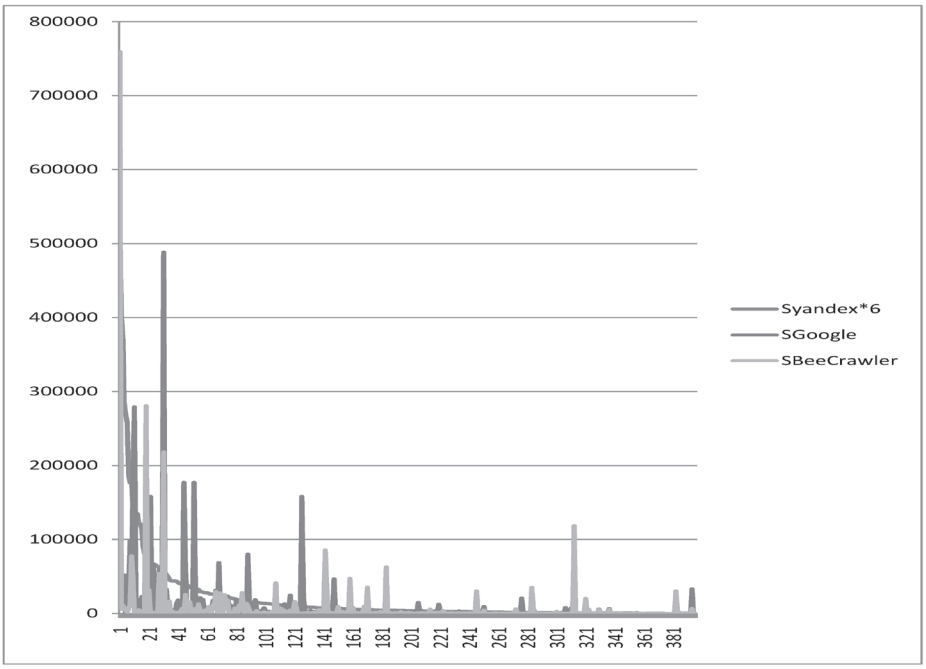
Последние измерения, используемые в статье, проводились в июне-июле 2013 года. Полужирным шрифтом выделены значения S_Y^i и S_G^i, которые представляются ошибочными. В частности, сайт Красноярского научного центра можно оценить визуально в любом браузере.

На рис. 1 приводятся значения размеров сайтов S_Y^i, S_G^i, S_B^i. Для удобства восприятия в едином масштабе значения S_Y^i умножены на коэффициент 6 и упорядочены по убыванию, поэтому соответствующий график выглядит как непрерывная убывающая кривая, хотя и скрытая за пиками на заднем плане.

Таблица 1

Фрагмент таблицы значений вебометрических индикаторов

i	Научное учреждение	Имя сайта	S_Y^i	S_G^i	S_B^i	V_Y^i
1	Российская академия наук	www.ras.ru	83013	606000	599904	20772
3	Отделение физических наук РАН	www.gpad.ac.ru	517	523	460	770
29	Пущинский научный центр РАН	www.psn.ru	5829	153	153	857
34	Красноярский научный центр СО РАН	www.krasn.ru	7291	93	1	500
44	Камчатский научный центр ДВО РАН	www.kscnet.ru	5757	177000	1525	6784
52	Библиотека по естественным наукам РАН	www.benran.ru	6471	177000	864	4154



Значения S_Y^i, S_G^i, S_B^i.

Графики по S_G^i и S_B^i в черно-белом изображении не слишком выразительны и в основном характеризуют пики значений, подозрительных на ошибки. При этом регистрируемые ошибки по S_B^i практически во всех случаях имеют содержательное объяснение: как правило, это так называемые «паучьи ловушки», когда разработчики сайтов (умышленно или неумышленно) создают условия для зацикливания поискового робота [10]. К ним относится организация меню сайта в виде дерева, динамические календари и т.д.

Обозначим $S_Y = (S_Y^1,...,S_Y^N)$, $S_G = (S_G^1,...,S_G^N)$ и $S_B = (S_B^1,...,S_B^N)$. Коэффициенты корреляции по Пирсону имеют следующие значения: $\rho(S_Y, S_G) = 0,639$, $\rho(S_Y, S_B) = 0,532$ и $\rho(S_G, S_B) = 0,890$. Таким образом, можно говорить о высокой степени взаимосвязи между S_G и S_B, и низкой между остальными парами индикаторов.

Функция ранжирования. Функция ранжирования в проекте «Вебометрический рейтинг научных учреждений России» [5] находится в процессе исследования и разработки, но для лучшего понимания процедуры сглаживания ошибок здесь стоит коротко остановиться на основных подходах к ее построению.

Обозначим $Rank(S_Y^i)$ ранг *i*-го сайта по индикатору *S*, измеренному поисковой системой Яндекс. Здесь ранг – это порядковый номер сайта в упорядоченном по убыванию векторе S_Y, то есть сайт с максимальным значением S_Y^i имеет ранг, рав-

ный 1. Для остальных индикаторов ранги определяются аналогично.

Для *i*-го сайта вычисляется интегральный показатель *R(i)* как функция от рангов сайта по каждому индикатору. Далее сайты упорядочиваются по возрастанию значений *R(i)*, сайт с минимальным значением *R(i)* получает вебометрический ранг *WR(i)*, равный 1, и т.д. В настоящее время *R(i)* определяется как сумма рангов *i*-го сайта.

При таком построении функции ранжирования *WR* в случае ошибочных значений вебометрических индикаторов нас интересуют, собственно говоря, не их реальные значения, а ранги по соответствующим индикаторам. Это замечание мы будем иметь в виду в следующем разделе.

Сглаживание ошибок. Сглаживанием ошибок поисковых систем будем называть процедуру вычисления правдоподобных значений индикаторов «размер сайта» вместо измеренных и представляющихся ошибочными значений с использованием данных о количестве страниц сайта, обнаруживаемых BeeCrawler.

Первым шагом такой процедуры является визуальное выявление подозрительных на ошибку значений S_Y^i, S_G^i, S_B^i (с использованием графиков, аналогичных приведенному на рис. 1) и очистка таблицы значений вебометрических индикаторов от строк, соответствующих сайтам, у которых обнаружены такие значения.

Обработка таблицы вебометрических индикаторов, измеренных в июне-июле 2013 года, выявила около 40 таких сайтов. После очистки значения коэффициентов корреляции изменились, теперь $\rho(S_Y, S_B) = 0,648$ и $\rho(S_G, S_B) = 0,929$.

Предположим, что Яндекс формирует значения S_Y^i исходя из следующих правил:

1. на каждом уровне сайта индексируется часть страниц, и чем ниже уровень, тем меньше эта часть;

2. чем больше внешних ссылок сделано на сайт, тем большее количество его страниц индексируется.

Эксперименты, проведенные с результатами работы BeeCrawler с учётом сделанных предположений, приводят нас к построению формулы следующего вида:

$$S_{B\to Y}^i = (S_B^{0,i} + S_B^{1,i} \times d + S_B^{2,i} \times d^2 + ... + S_B^{j,i} \times d^j + ... + S_B^{M,i} \times d^M) \times V_Y^i, \quad (1)$$

где $0<d<1$ – коэффициент затухания, чем ниже уровень сайта, тем меньше страниц индексируется. Вычисления $\rho(S_Y, S_{B\to Y})$ для очищенной таблицы индикаторов показывают, что максимальное значение коэффициента корреляции, равное 0,826, достигается при d=0,075 и M=7 (здесь, как и ранее $S_{B\to Y} = (S_{B\to Y}^1, ..., S_{B\to Y}^N)$).

Учитывая достаточно сильную статистическую зависимость между S_Y^i и $S_{B\to Y}^i$, для сглаживания ошибок Яндекса предлагается использовать соответствующие значения $S_{B\to Y}^i$. Точнее, как было сказано ранее, нас интересуют не сами значения ошибочных S_Y^i, а их $Rank(S_Y^i)$. Поэтому для ошибочного значения с индексом i по формуле (1) вычисляется $S_{B\to Y}^i$, далее определяется его ранг $Rank(S_{B\to Y}^i)$ для вектора $S_{B\to Y}$ и полученное значение присваивается $Rank(S_Y^i)$ как правильное.

Эксперименты показали, что для Google также можно построить формулу следующего вида:

$$S_{B\to G}^i = S_B^{0,i} + S_B^{1,i} \times d + S_B^{2,i} \times d^2 + ... + S_B^{j,i} \times d^j + ... + S_B^{M,i} \times d^M. \quad (2)$$

Обратим внимание на то, что индикатор ссылочной популярности сайта V в (2) не используется. Вычисления $\rho(S_Y S,)_{B\to G}$ для очищенной таблицы индикаторов показывают, что максимальное значение коэффициента корреляции, равное 0,941, достигается при d=0,5 и M=7. Отсюда следует, что формулу (2) можно использовать для сглаживания ошибочных S_G^i по аналогии с формулой (1) для Яндекса (хотя и корреляция между S_G и S_B была уже достаточно большой).

Заключение

В статье предложен подход к сглаживанию ошибок поисковых систем, возникающих при измерениях вебометрического индикатора «размер сайта». Данный подход применяется в проекте «Вебометрический рейтинг научных учреждений России» и в большинстве случаев позволяет использовать точные процедуры исправления очевидных ошибок вместо слабо формализуемых мнений экспертов, в качестве которых пока выступают сами разработчики проекта.

Работа выполняется при поддержке гранта РГНФ № 12-03-12001.

Список литературы

1. Bar-Ilan J. Expectations versus reality – Search engine features needed for Web research at mid / J. Bar-Ilan // International Journal of Scientometrics, Informetrics and Bibliometrics. 2005. Vol. 9. URL:http://www.cybermetrics.info/articles/v9i1p2.pdf.

2. Bar-Ilan J. How much information do search engines disclose on the links to a web page? A longitudinal case study of the 'cybermetrics' home page / J. Bar-Ilan // Journal of Information Science. 2002. Vol. 28, No. 6. P. 455-466.

3. Snyder H. Can search engines be used as tools for web-link analysis? A critical view / H. Snyder, H. Rosenbaum // Journal of documentation. 1999. Vol. 55(4). P. 375-384.

4. Thelwall M. Web impact factors and search engine coverage / M. Thelwall // Journal of Documentation. 2000. Vol. 56(2). P. 185-189.

5. Вебометрический рейтинг научных учреждений России [Электронный ресурс]. Режим доступа: http://webometrics-net.ru (дата обращения 15.07.2013).

6. Рейтинг сайтов научных учреждений СО РАН [Электронный ресурс]. Режим доступа: http://www.ict.nsc.ru/ranking (дата обращения 16.07.2013).

7. Вебометрический индекс российских вузов и НИИ [Электронный ресурс]. Режим доступа: http://ru-webometrics.info (дата обращения 17.07.2013).

8. Ranking Web of World universities [Электронный ресурс]. Режим доступа: http:// www.webometrics.info (дата обращения 17.07.2013).

9. Чернобровкин Д.И., Печников А.А. Свидетельство о гос. регистрации программы для ЭВМ «Программа для поиска и сбора внешних гиперссылок BeeCrawler» Федеральной службы по интеллектуальной собственности, патентам и товарным знакам РФ № 2012619665 от 26 октября 2012 г.

10. Pant G., Srinivasan P., Menczer F. Crawling the Web // In Web Dynamics / Springer. – 2004. Levene M. and Poulovassilis A., eds. P. 153-178.