ПЕЧНИКОВ Андрей Анатольевич
Институт прикладных математических
исследований Карельского научного центра РАН,
ведущий научный сотрудник,
доктор технических наук, доцент
pechnikov@krc.karelia.ru

РАЗМЫШЛЕНИЯ о ВЕБОМЕТРИЧЕСКОМ РЕЙТИНГЕ

Вебометрический рейтинг испанской исследовательской группы Cybermetrics Lab достаточно хорошо известен в России, возможно, и по той причине, что несколько лет назад его результаты использовались в Национальном рейтинге классических университетов (кстати, при поддержке Минобразования РФ). Автор занимается задачами вебометрики (вебометрика – гораздо шире, чем вебометрические рейтинги) с 2006 года, и естественно, что вебометрические рейтинги попали в сферу его интересов, хотя вопросов собралось больше, чем ответов. Но опыт накоплен, поэтому захотелось поделиться с читателями журнала некоторыми соображениями о вебометрических рейтингах.

Интернет настолько решительно вошел в нашу жизнь, что кажется, будто бы он был всегда. А ведь многие из нас хорошо помнят те времена, когда брошюра (а еще лучше целая книга), посвященная университету, научному центру, институту (можно продолжать: факультету, отделу, кафедре), издавалась лишь по случаю юбилея, а потом вручалась наиболее важным гостям и коллегам. Эти брошюры прочитывались очень внимательно, сотрудники не без удовлетворения находили свои имена, фамилии, степени и звания, огорчались, не обнаружив себя на фотографиях, и иронизировали по поводу вкравшихся опечаток.

Появились Интернет и Веб.

Использование автором термина «Веб» в русской транскрипции является в какой-то мере противопоставлением понятию «Интернет», которое сегодня в обиходе часто используется и для обозначения Веба. На самом деле Интернет – это глобальная телекоммуникационная сеть информационных и вычислительных ресурсов, а Веб – глобальное информационного пространство, основанное на физической инфраструктуре Интернета, специальном протоколе передачи данных и особом языке, поэтому термины «Веб» и «Интернет» не являются синонимами.

Вскоре появились веб-сайты университетов, научных центров, институтов, а также кафедр, отделов и так далее (и даже персональные страницы – но они все-таки выпадают из контекста этой статьи). Куда же исчез наш интерес к этим «виртуальным брошюрам»? Почему мы спокойно взираем (или вовсе не взираем) на прошлогодние «новости», устаревшие объявления, фотографии сотрудников, которые давным-давно не работают, и отсутствие фотографий работающих? Почему нас не волнует то, что список наших научных работ «за последние 5 лет» заканчивается 2009 годом?

Следует оговориться, что эта статья все-таки имеет большее отношение к сайтам научных учреждений, чем университетов, поскольку, во-первых, именно научным учреждениям посвящен проект, выполняемый под руководством автора в настоящее время, а во-вторых, на официальных сайтах университетов России в последнее наведен определенный порядок.

Столь эмоциональная преамбула понадобилась автору для того, чтобы высказать следующую крамольную мысль: мы занимаемся ранжированием сайтов не для того, чтобы назвать лучшие и худшие (или не только для этого). А для того, чтобы предоставить возможность заинтересованным лицам хотя бы чуть-чуть проанализировать

> ***Все западные поисковые системы существенно занижают значения индикаторов для российских сайтов по сравнению с Яндексом***

ситуацию и попробовать ответить на вопрос типа: почему сайт моего института имеет рейтинг ниже, чем сайт института, где работает мой коллега (директор, заведующий лабораторией, научный сотрудник, веб-мастер и т.д.)? И сделать свой сайт еще лучше.

Так вот, для анализа ситуации нужны (объективные) показатели, значения которых можно измерить для конкретного сайта. Так уж сложилось в вебометрическом ранжировании, что основными показателями (или индикаторами, по терминологии испанской группы Cybermetrics Lab, запустившей, по видимому, первый проект по вебометрическому ранжированию еще в 2004 году – http://www.webometrics.info) являлись следующие:

1) размер сайта (S – size) – общее количество страниц;

2) видимость сайта (V – visibility) – количество уникальных гипертекстовых ссылок на сайт с других веб-ресурсов;

3) количество полнотекстовых файлов (R – rich files) – суммарное количество файлов с расширениями PDF, DOC, PS, PPT и т.д.;

4) научность сайта (Sc – scholar) – индикатор, вычисляемый Google Scholar.

Для первых трех индикаторов инструментами измерений долгое время являлись известные поисковые системы Google, Yahoo, Bing и др. В испанском рейтинге поисковая система Яндекс не используется.

Далее строится функция ранжирования с учетом введенных весовых коэффициентов ценности того или иного индикатора.

Пару лет назад испанцы несколько поменяли свои подходы, в принципе, сохранив перечисленные индикаторы, но изменив коэффициенты «значимости». Новая версия доступна на указанном ранее сайте. Эти изменения не слишком влияют на основные мысли, излагаемые в данной статье: нам достаточно тех определений индикаторов, которые приводятся выше, поэтому мы не будем на них останавливаться.

Нам хорошо известны еще несколько проектов по вебометрическому ранжированию, выполняемых исследовательскими группами в России и на Украине:

- «Рейтинг сайтов научных учреждений СО РАН» (http://www.ict.nsc.ru/ranking), Институт вычислительных технологий СО РАН;
- «Вебометрический индекс российских вузов и НИИ» (http://ru-webometrics.info), Институт научной и педагогической информации РАО;
- «Сервис вебометрических исследований научных сайтов» (http://webometrics.fegi.ru), Дальневосточный геологический институт ДВО РАН;
- «Рейтинг сайтов вузов и институтов» (http://webometrics.sfu-kras.ru), Сибирский федеральный университет;
- «Перший український рейтинговий портал» (http://ranking.sumdu.edu.ua), Сумский государственный университет.

Добавим к этому списку и проект, которым руководит автор статьи – «Вебометрический рейтинг научных учреждений России» (http://webometrics-net.ru), Институт прикладных математических исследований Карельского научного центра РАН.

Первые три вебометрических индикатора вроде бы ни у кого не вызывают сомнений, поскольку, как говорится, а что еще можно измерить у сайта? Четвертый индикатор как-то уж слишком упирает на Google Scholar, поэтому исследователи пытаются его различными способами усовершенствовать. Ну и, естественно, все стремятся добавлять что-то свое как в индикаторы, так и в функции ранжирования.

Кстати, как уже можно было увидеть, дизайн веб-сайта, удобство и простота использования (то есть некие эргономические индикаторы), а также количество посещений/посетителей в вебометрическом рейтинге не рассматриваются. Это понятно – вопросов хватает и без них.

Первый вопрос – чем измерять? В ряде авторских работ показано, что все западные поисковые системы существенно занижают значения индикаторов для российских сайтов по сравнению с Яндексом. Такая ситуация вполне объяснима привязкой поисковых машин к национальному фрагменту Веба, и для Яндекса это в первую очередь зона Рунета. Но из этого следует, что при измерении индикаторов сайтов конкретной страны необходимо использовать поисковые машины, наиболее распространенные в этой стране (для России – Яндекс), и только потом – наиболее распространенные в мире (Google). Мы уже знаем, что в списке используемых Cybermetrics Lab поисковых систем Яндекс отсутствует.

Для читателей, не знакомых с этой темой, поясним, что если в Google в поисковой строке набрать текст «site: www.msu.ru», то на экране можно увидеть примерно такой ответ: «Резуль-

татов: примерно 26300...», что и можно принять за значение S, измеренное Google как размер официального сайта МГУ. В Яндексе также есть похожие возможности. Мы в проекте http://webometrics-net.ru, например, используем Яндекс. Вебмастер: если в поисковой строке на странице http://webmaster.yandex.ru/check.xml набрать текст «www.msu.ru» (без кавычек, конечно), то получим «Страницы: 16520». Это страницы сайта, попавшие в поисковый индекс, и понятно, что сюда попадают не все страницы, которые находит Яндекс. Получается, что индикатор, который мы называем «размер сайта», – это количество страниц на сайте, обнаруживаемое соответствующей поисковой системой, да еще и при условии именно такой формы запроса. Попробуйте в Google набрать «site:msu.ru», – увидите много интересного: а) гораздо больше страниц, б) страницы с других сайтов домена "msu.ru".

У нас есть свой краулер BeeCrawler, и мы знаем, что на официальном сайте МГУ страниц гораздо больше, чем 26300. Добавим еще, что мы проводили довольно много измерений сайтов, и можем сказать, что значения S по Яндексу и S по Google не очень-то коррелируют. Так что же мы измерили Яндексом и Google в качестве размера сайта?

Насчет Яндекса ответ кажется очевидным: мы уже сказали, что это количество страниц, попавших в его индекс. По Google ответ не столь очевиден.

Может быть, стоит использовать в качестве S значения, получаемые BeeCrawler? Может быть, и стоит, но только другие исследователи не смогут повторить наши измерения (если, конечно, не воспользуются нашим краулером, но, понятно, это дело достаточно утомительное), тут с Яндексом и Google дело обстоит проще. И потом, где гарантии того, что BeeCrawler дает правильные результаты? Он предназначен для сбора гиперссылок, а не для подсчета страниц, да и вообще у любого краулера столько особенностей, что лучше в рамках этой статьи и не заикаться. Чего стоят одни так называемые «паучьи ловушки» или нормализация гиперссылок!

Но к поисковым системам есть и еще один вопрос. Иногда они дают результаты, которые абсолютно точно являются ошибочными. Встречаются сайты, для которых Google безапелляционно сообщает, что на них 177000 страниц. Ни Яндекс, ни Bee Crawler, ни просмотр «вручную» ничего похожего не показывают. В нашем проекте мы предлагаем подход к «сглаживанию» ошибок поисковых систем, основанный на использовании результатов сканирования сайтов с помощью Bee Crawler.

После сглаживания сайты упорядочиваются по количеству страниц по Яндексу, Google и BeeCrawler, затем выбираются два наиболее близких места у каждого сайта, усредняются, результаты снова упорядочиваются и получается рейтинг сайта по S.

По остальным индикаторам происходит что-то похожее.

Следующий вопрос касается единиц анализа (примеры снова будут об индикаторе S). По методике Cybermetrics Lab в качестве единиц анализа берутся доменные имена официальных сайтов научных организаций. Это не всегда корректно для организаций российского научного веб-пространства. Продемонстрируем сказанное на примере на примере сайта Карельского научного центра РАН (КарНЦ РАН). Детальный анализ перечня страниц сайта КарНЦ РАН, найденных Google по запросу вида «site:krc.karelia.ru», показывает, что сюда же отнесены страницы самостоятельных сайтов, имеющих доменные имена 4-го уровня (rcdl2009.krc.karelia.ru – всероссийская конференция RCDL-2009, ib.krc.karelia.ru – Институт биологии, и многие другие). В то же время измерение значения S, к примеру, для сайта Института биологии КарНЦ РАН дает ненулевое значение, т.е. этот сайт в то же самое время рассматривается Google как самостоятельная единица анализа. Да он таковым и является, поскольку Институт биологии КарНЦ РАН является учреждением РАН и самостоятельным юридическим лицом. Отсюда следует, что реальное значение S для КарНЦ РАН завышено за счет самостоятельных сайтов, имеющих доменные имена 4-го уровня, входящие в домен третьего уровня krc.karelia.ru. Да и сайт конференции rcdl2009.krc.karelia.ru с натяжкой можно считать веб-ресурсом КарНЦ РАН – скорее, он относится к Институту прикладных математических исследований.

В то же время существуют веб-ресурсы институтов, не ассоциируемые по доменному имени с их головным сайтом. Например, в Институте проблем управления РАН (официальный сайт – www.ipu.ru) существует крупный сайт «Теория управления организационными системами» (доменное имя – www.mtas.ru), который не учитывается при измерениях S, V, R

«Единицу анализа», содержащую все веб-сайты учреждения, построить очень трудно»

Ключевые слова:
вебометрические рейтинги, веб-сайт, вебометрический индикатор, поисковая система, краулер

Keywords:
webometric ranking, web-site, webometric indicator, search system, crawler

> ***Что лучше – больше ссылок или больше сославшихся сайтов?***

и Sc поисковыми системами как веб-ресурс института.

Правда, такую вот «единицу анализа», содержащую все веб-сайты учреждения, построить очень трудно. В рамках нашего проекта мы пытаемся это делать (сейчас у нас на 397 официальных сайтов РАН приходится 560 сайтов, не являющихся официальными сайтами, но составляющими множество веб-сайтов этих 397 учреждений). Это, конечно, мало. Но и искать веб-сайты, не ассоциируемые по доменному имени сайта головной организации, непросто. Хотелось бы, чтобы директора, заведующие лабораториями, научные сотрудники, веб-мастера подсказали нам, что у них есть еще. На сайте проекта для этого есть раздел обратной связи.

С другими индикаторами тоже происходит много интересных дел, но мы не будем их описывать в рамках этой статьи. Лучше остановимся на индикаторе, который добавлен нами для множества сайтов РАН. Речь идет о некоей «внутренней ссылочной популярности», или, по-другому, о том, как исследуемые нами академические сайты ссылаются друг на друга. Вроде бы достаточно очевидно, что все сайты РАН должны (или «могли бы») ссылаться на официальный сайт РАН (www.ras.ru). В действительности это происходит не всегда. Если пока не брать в расчет «единицы анализа», содержащие все веб-сайты учреждения (эта часть у нас пока «сырая»), а основываться только на данных по 397 официальным сайтам, то картина получается следующая. На официальный сайт РАН (www.ras.ru) сделана 1081 ссылка с 239 сайтов, на сайт Карельского научного центра РАН (www.krc.karelia.ru) сделано 12499 ссылки с 14 сайтов, а, например, на сайт Института горного дела СО РАН (www.misd.nsc.ru) – 30 ссылок с 9 сайтов. (Здесь «ссылка» на самом деле понимается как «уникальная гиперссылка» – это гиперссылка из множества всех гиперссылок с одинаковым контекстом, которая находится на странице, имеющий максимальный уровень; при этом уровень начальной страницы сайта считается наивысшим. Если, например, на каждой странице сайта стоит баннер с «живой» ссылкой на сайт РАН, то количество таких ссылок у нас равно 1, а не тому, сколько страниц на сайте.) Конечно, 1081 ссылка на сайт Карельского научного центра РАН может вызвать некоторые сомнения. Но здесь есть объяснение: просто со многими сайтами центра работает одна и та же группа веб-мастеров.

И все же, что лучше – больше ссылок или больше сославшихся сайтов? На всякий случай мы пока перемножаем эти два показателя и получаем значение «внутренней ссылочной популярности».

Интегральный показатель в нашем проекте пока строится простым суммированием мест сайта по каждому индикатору и последующим ранжированием по сумме мест (хотя экспериментируем сейчас с методами Борда и Кондорсе – http://ru.wikipedia.org/wiki/ Метод_Борда).

В конце 2013 года первую пятерку у нас составили сайты Физико-технического института им. Иоффе РАН, Института математики им. С.Л. Соболева СО РАН, Института цитологии и генетики СО РАН, Института прикладной математики им. М. В. Келдыша РАН и Института философии РАН.

Но вот что интересно. Мы сравнили полученные результаты по нашему проекту и по проекту Cybermetrics Lab (у нас с ними совпали только 100 учреждений РАН). И получили, что коэффициент ранговой корреляции Кендалла =0.74, что при уровне значимости =0.05 показывает существующую значимую ранговую корреляционную связь. И это наводит на определенные размышления. Как же так, мы используем достаточно примитивный интегральный критерий, а испанцы его всячески обосновывают, мы используем Яндекс, а испанцы нет, и т.д. – а корреляция есть?! Может быть, есть что-то скрытое, глубинное, то, чего мы пока не знаем, а измеряем лишь то, что лежит на поверхности? Вот этот вопрос представляется очень интересным.

И еще. Есть у экономистов такой закон Гудхарта, который заключается в том, что, когда социальный или экономический показатель становится целью для проведения социальной или экономической политики, он перестает быть достойным доверия показателем (http://ru.wikipedia.org/wiki/Закон_Гудхарта). Это очень точно соотносится с вебометрическим ранжированием. В самом начале мы говорили о том, что занимаемся ранжированием сайтов не для того, чтобы назвать лучших и худших (или не только для этого), а для того, чтобы предоставить возможность заинтересованным лицам проанализировать ситуацию. Но ведь и индекс цитирования РИНЦ вначале был «иссле-

довательским», что ли, а сейчас используется при результатах оценки научной деятельности. И сидит ученый, изучает, куда же ему статью отправить: то ли в очень хороший (с его точки зрения) журнал, да РИНЦ маловат, то ли в «так себе» (опять-таки – это его личное мнение, хотя чаще всего совпадающее с мнением коллег), но зато РИНЦ-то какой!

А с другой стороны, официальные сайты, это совершенно очевидно, и должны подвергаться административному воздействию: в конце концов, они создаются на деньги налогоплательщиков, и должны удовлетворять определенным требованиям.

Возвращаясь к сайтам университетов, почитайте 29 статью Закона Российской Федерации «Об образовании в Российской Федерации»: «...Образовательные организации формируют открытые и общедоступные информационные ресурсы, содержащие информацию об их деятельности, и обеспечивают доступ к таким ресурсам посредством размещения их в информационно-телекоммуникационных сетях, в том числе на официальном сайте образовательной организации в сети “Интернет”». И далее – там много интересного, 17 обязательных разделов на сайте должно быть. А в Национальном рейтинге классических университетов (методика 2009 года, http://www.univer-rating.ru/txt.asp?rbr=30&txt=Rbr30Text2268&lng=0) в качестве одного из индикаторов было взято место вуза в рейтинге Webometrics.

Официальные сайты создаются на деньги налогоплательщиков и должны удовлетворять определенным требованиям

И ведь реально за последние 3-4 года официальные сайты университетов изменились в лучшую сторону. Может быть и ладно, ну его, этот закон Гудхарта, зато научные сайты станут лучше?!

Andrey A. Pechnikov
Institute of Applied Mathematical Research Karelian Research Centre of RAS, Leading Researcher, Doctor of Technical Sciences, Associate Professor

Thinking about the Webometric Rating

Webometric Ranking of Spain research group Cybemetrics Lab is well-known in Russia. Several years ago results, which they produce used by the National rating of classical universities (with support of Russian Ministry of Education and Science). The professional interests this article's author is webometric still 2006 year (it is wider than webometric ratings). There are a lot of questions in this sphere. This article is a one way to generalize and discuss them.