УДК З 97
ББК 681

П. А. Бузунов
г. Чита, Россия

Ранжирование официальных сайтов университетов мира как метод построения рейтинга посредством присутствия в сети Интернет

На основе настоящих данных проведён общий обзор состояния вебометрических исследований в России и зарубежом. Проанализирована методология определения рейтинга доменов учебных заведений в сети Интернет на примере исследования технологии обработки входных-выходных данных с сайтов представителями испанской исследовательской группы Webometrics.

Ключевые слова: вебометрика, индикаторы, поисковые машины, ранжирование сайтов.

P.A. Buzunov
Chita, Russia

Ranking of the Official Websites of World Universities as a Method of Rating by the Presence in the Internet

On the basis of current data the overview of webometric researches in Russia and foreign countries is carried. The methodology of ranking of official websites of educational institutions in the Internet are analyzed on the example of research of technology processing of input-output data from the sites by the representatives of Spanish research group Webometrics.

Keywords: webometrics, indicators, search engines, ranking of sites.

Введение

Существует достаточно много методик для сравнения эффективности представления университетского веб-пространства в сети Интернет. Один из наиболее качественных подходов — использование инструментов вебометрики (веб-индикаторов), доверием к измерению которых не пренебрегает ни один из академических кругов.

Термин «вебометрика» («webometrics») был введен сравнительно недавно – в 1997 году. П. Игверсен и Т. Алминд в своей работе «Informetric analyses on the World Wide Web: Methodological approaches to "webometrics"», опубликованной в журнале Journal of Documentation, определили данное понятие после расширения понятия «библиометрика» на веб-пространство. «Вебометрика» здесь понимается как информетрический метод для World Wide Web (далее, Веб). Другими словами, вебометрика занимается количественными аспектами конструирования (создания) и использования информационных ресурсов, структур и технологий применительно к Вебу [4].

В 1998 году П. Ингверсон ввел понятие Web Impact Factor *(WIF)* и определил его как отношение числа входящих ссылок V на сайт к общему количеству страниц S на сайте:

$$WIF = \frac{V}{S}$$

Это, по сути, первая достойная внимания вебометрическая характеристика, предложенная в то время, которая использовала веб-индикаторы для создания общей картины ранжирования. Конечно, она уже позволяла составить рейтинг сайтов высших учебный заведений и научных организаций, но зависимость этого отношения только от двух веб-индикаторов вызывало сомнение в том, что с помощью этой характеристики можно провести всесторонний анализ сайта. К тому же, значение данных индикаторов разнилось для одного и того же сайта в зависимости от использования для их определения разных поисковых машин.

На сегодняшний день существует достаточно рейтинговой информации по университетам мира, следовательно, разработано множество методик по ранжированию сайтов университетов.

Существуют несколько зарубежных источников по ранжированию университетов мира [1]:
1) Webometrics: World Universities' Ranking on the Web by Cybermetrics Lab;
2) The Times Higher Education - QS World University Rankings;
3) Academic Ranking of World Universities by SJTU;
4) Top 100 Global Universities by Newsweek;
5) G-Factor International University Rankings by Google Search;
6) Professional Ranking of World Universities by MINES Paris Tech;
7) Performance Ranking of Scientific Papers for World Universities by HEEACT;
8) Global University Ranking by Wuhan University, China и др.

В России же можно отметить несколько исследований, направленных на изучение национальных веб-ресурсов университетов и научных отделений РАН: исследование научных сайтов Сибирского отделения РАН, рейтинг научных учреждений СО РАН, рейтинги университетов северо-запада России и Финляндии [5].

Исследование подхода к построению рейтингов доменов университетов мира на примере проекта Webometrics

Из мировых рейтинговых работ можно отметить испанскую исследовательскую группу Cybermetrics Lab, принадлежащей крупнейшему исследовательскому Центру информации и документации Испании Consejo Superior de Investigaciones Cienthficas (CSIC) с их успешным проектом Webometrics Ranking of World Universities (кратко, проект Webometrics). Изначально сайт проекта – http://www.webometrics.info — разрабатывался для привлечения научных публикаций в глобальную сеть, но, со временем, практика показала, что полученные данные с сайтов можно использовать с целью ранжирования [6]. Первая публикация рейтингов появилась в 2004 году, начиная с 2006 года, в год появляется по 2 рейтинга (в январе и июле) [2].

Для своих исследований Webometrics использовали 4 веб-индикатора:

1) Размер сайта (Size) – количество страниц на сайте, найденных четырьмя поисковыми машинами: Google, Yahoo, Live Search и Exalead [3];

2) Видимость сайта (Visibility) – общее количество уникальных внешних ссылок на сайт. С большей степенью достоверности может быть получено только в Yahoo Search [3];

3) Число файлов на сайте (Rich Files) – количество размещенных на сайте научных работ, публикаций исследований в форматах Adobe Acrobat (.pdf), Adobe PostScript (.ps), Microsoft Word (.doc) и Microsoft Powerpoint (.ppt). для поиска файлов используют ресурс Google [3];

4) Научность сайта (Scholar) – научная популярность, или индекс цитирования, сайта с точки зрения поисковой системы Google Scholar. Здесь определяется количество работ и цитат на сайте, используемых в мире в различных областях знаний [3].

Формула, предложенная Webometrics, представляет собой многочлен с весовыми коэффициентами:

$$WR = \alpha V + \beta S + \gamma R + \delta Sc, \tag{1}$$

где:

WR (Webometrics Rank) – результирующее значение в статистике Webometrics (ранг);

$\alpha, \beta, \gamma, \delta$ – весовые коэффициенты;

V, S, R, Sc – нормированные вебометрические параметры: видимость сайта, его размер, научные файлы сайта и индекс цитирования соответственно.

Остановимся отдельно на каждом параметре и на весовых коэффициентах.

V – параметр, представляющий нормированный на единицу по всем исследуемым учебным заведениям результат, полученный с помощью Yahoo Search.

В общем виде нормирование параметров происходит по формуле [2]:

$$N_a = \frac{ln(n_a + 1)}{ln(max(n_i) + 1)}, \tag{2}$$

где:

N – поисковая машина (Google, Yahoo, Live Search или Exalead);

a – исследуемый домен (сайт);

n_a – значение параметра поисковой машины N для домена a;

i – количество исследуемых сайтов учебных и научных заведений.

Тогда нормированный параметр видимости домена (сайта) a примет вид:

$$V_a = Y_a,$$

где:

$$Y_a = \frac{ln(y_a + 1)}{ln(max(y_i) + 1)},$$

здесь

V_a – значение видимости (наличия внешних ссылок) домена (сайта) a;

Y_a – нормированное по методу (2) значение числа внешних ссылок на домен (сайт) a, полученное с помощью Yahoo Search.

S – индикатор индексированной части сайта, представляющий собой нормированные на единицу по всем исследуемым учебным заведениям значения, полученные поисковыми машинами Google, Yahoo, Live Search, Exalead значения объема сайта, усреднённые с помощью метода медиан [2].

Используя (2), запишем:

$$S_a = \frac{1}{2}\big((G_a + Y_a + L_a + E_a) - max(G_a, Y_a, L_a, E_a) - min(G_a, Y_a, L_a, E_a)\big),$$

где:

S_a – усреднённый параметр нормированных значений числа страниц домена (сайта) a;

G_a, Y_a, L_a, E_a – нормированные по методу (2) значения числа страниц домена, а по поисковым машинам Google, Yahoo, Live Search, Exalead соответственно;

R – параметр, представляющий собой сумму нормированных на единицу по всем исследуемым учебным заведениям значений числа файлов по отдельности четырёх форматов (*.pdf, *.ps, *.doc, *.ppt) [2]:

$$R_a = PDF_a + Ps_a + DOC_a + PPT_a,$$

где:

R_a – результирующее значение показателя числа файлов с научными данными домена (сайта) a, нормированного отдельно по каждому формату;

$PDF_a, Ps_a, DOC_a, PPT_a$ – нормированные значения числа файлов домена (сайта) a, найденные Google. Для нормирования используется формула (2).

Sc – популярность сайта, представляющая собой нормированное на единицу по формуле (2) количество страниц и ссылок на сайт учебного заведения, найденных с помощью поисковой системы Google Scholar [6].

$$Sc_a = G_a,$$

где:

Sc_a – нормированный индекс цитирования домена (сайта) a.

Весовые коэффициенты за историю Webometrics менялись 2 раза. Главным условием было сохранение для коэффициентов отношения:

$$\frac{k(V)}{k(S) + k(Sc) + k(R)} = 1.$$

С начала исследования весовые коэффициенты распределялись как:

$$k(V) : k(S) : k(Sc) : k(R) = 0,5 : 0,25 : 0,125 : 0,125.$$

Затем параметрам индекса цитируемости (Sc) и количества файлов домена (R) были увеличены весовые коэффициенты.

$$k(V) : k(S) : k(Sc) : k(R) = 0,5 : 0,20 : 0,15 : 0,15. \quad (3)$$

Так весовые коэффициенты выглядят и в настоящее время.

Исходя из (3), формула (1) принимает вид:

$$WR = 0,5V + 0,2S + 0,15R + 0,15Sc,$$

или:

$$WR = (10V) + (4S + 3R + 3Sc).$$

Что касается позиции России в рейтинге Webometrics в январе 2011 года:

1) В Топ-200 не попал никто;

2) В Топ-500 один вуз;

3) В Топ-1000 5 вузов.

По сравнению с июлем 2010 года изменилась ситуация только с Топ-1000, куда попали ещё 2 вуза.

Интересно привести мнение руководителя проекта Webometrics, Isidro F. Aguillo из Центра научной информации и документации при Высшем совете по научным исследованиям Испании относительно мест в этом рейтинге российских вузов [7]:

«. . . По моему мнению, у России в Топ-100 должны быть, по крайней мере, два университета – Московский и Санкт-Петербургский. Сегодня они далеки от этого.

Возможно, первая причина – недостаток контента, что удивительно, тем более, если учесть, что это очень большие учреждения с множеством профессоров и студентов. Следует обязать все факультеты и исследовательские группы увеличить свое присутствие в сети, и, конечно, создать архивы публикаций.

Вторая причина – недостаток английских текстов. В ее решении надо руководствоваться не формулой "английский вместо русского", а "английский плюс русский"».

Многие исследователи рейтингов в России используют в своих работах статистику поисковых машин Yandex, Rambler, которые лучше индексируют русскоязычные сайты, и, основываясь на данных которых, сравнивают рейтинги сайтов каждый год. Но прогресс в рейтинге среди поисковых машин России не говорит о росте в рейтинге на арене более популярных поисковых систем. Поэтому целесообразнее использовать более популярные в мире Google, Yahoo, и др.

Список литературы

1. Engr Muhammad Ismail. Ranking of Universities // National University of Science & Technology. URL: http://www.nust.edu.pk/general/Ranking%20of%20Universities%20-%20Engr%20Muhammad%20Ismail.doc. (дата обращения 01.06.2011).

2. Isidro F. Aguillo. Webometrics Ranking of World Universities// 3rd Meeting of the International Rankings Expert Group (IREG-3). Shanghai Jiao Tong University, 2007. URL: http://ed.sjtu.edu.cn/IREG-3/PPT/4%20No%20PPT%20Isidro%20Aguillo%20(table)/webometrics_aguillo.ppt (дата обращения 05.05.2011).

3. Methodology. Ranking Web of World Universities. URL: http://www.webometrics.info/methodology.html (дата образщения 31.05.2011).

4. Вебометрика. Институт прикладных математических исследований КарНЦ РАН. URL: http://webometrics.krc.karelia.ru (дата обращения 31.05.2011).

5. Мазалов В. В., Печников А. А. О рейтинге официальных сайтов научных учреждений северо-запада России // Управление большими системами. Вып. 24. М.: ИПУ РАН, 2009. С. 130–146.

6. Рейтинг Webometrics: мировые вузы в виртуальном пространстве / Материалы независимого агентства РейТОР. URL: http://www.reitor.ru/ru/analitic/experience/index.php?id19=304 (дата обращения 04.06.2011).

7. Стерлигов И. Самый дешевый способ помочь российской науке // Ученый совет. 2008. №8. С. 21–23.

Рукопись поступила в редакцию 14 июня 2011 г.