ПЕЧНИКОВ Андрей Анатольевич *– кандидат физико-математических наук, доцент, старший научный сотрудник Института прикладных математических исследований Карельского научного центра РАН*

ИЛЮКЕВИЧ Ольга Геннадьевна *– студентка Петрозаводского государственного университета*

РЕЙТИНГ ОФИЦИАЛЬНЫХ WEB-САЙТОВ УНИВЕРСИТЕТОВ РОССИИ И ФИНЛЯНДИИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Вебометрика (webometrics) как научное направление, занимающимся исследованием количественных аспектов конструирования и использования информационных ресурсов структур и технологий применительно к Web [1], естественно, коснулась Интернет-ресурсов высших учебных заведений, занимающих существенное место во всемирной паутине.

Проект "Webometrics Ranking of World Universities" испанской исследовательской группы Cybermetrics Lab [2] посвящен вебометрическим исследованиям и ранжированию web-сайтов вузов и научно-исследовательских институтов мира. Рейтинг публикуется на сайте проекта [3] с обновлением один раз в полгода. По результатам этих исследований сайты российских вузов занимают достаточно скромные позиции в мировом рейтинге, что, впрочем, не может не вызывать определенных сомнений.

Основываясь на подходах Cybermetrics Lab, авторы провели самостоятельное исследование и сравнительный анализ сайтов ряда классических университетов Северо-Запада России и Финляндии, что позволило выявить ряд интересных моментов, отраженных в данной статье.

1. Подходы и методики Cybermetrics Lab

В соответствии с [3] для построения алгоритма ранжирования сайтов университетов задаются следующие четыре индикатора:

– количество уникальных гипертекстовых ссылок с других ресурсов (V – visibility, цитируемость),
– общее количество страниц сайта (S – size, размер),
– количество полнотекстовых файлов, под которыми понимаются файлы с расширениями pdf, ps, doc, xls, ppt и rtf (R - "rich files"),
– количество статей, размещенных на сайте и их цитирований (Sc – scholar, «научность сайта»).

Индикатор S измеряется с использованием поисковых машин Google, Yahoo, Live Search and Exalead, индикатор V - Yahoo Search, Live Search и Exalead, индикатор R – Google, а индикатор Sc - Google Scholar.

Результирующие значения для S и V определяются как сумма результатов замеров, причем в первом случае минимальное и максимальное значение отбрасываются. Затем по каждому из критериев сайты ранжируются по убыванию соответствующего параметра.

По каждому индикатору производится ранжирование сайтов по убыванию значений соответствующих индикаторов. Для обозначения ранга по заданному индикатору используются обозначения RankV, RankS, RankR и RankSc, соответственно (наивысший ранг равен 1). Интегральный показатель, называемый «вебометрическим рангом» (WR - Webometrics Rank), получается в результате ранжирования суммы рангов, умноженных на коэффициенты, или, как в [3]:

Webometrics Rank (position) = 4*RankV+2*RankS+1*RankR+1*RankSc.

Исходя из значений коэффициентов, можно сделать вывод о том, что наибольшая значимость придается размещенным на сайте полнотекстовым файлам, статьям и их цитированию другими сайтами, что соответствует принципам Открытого доступа, когда интернет рассматривается в первую очередь как средство функционального объединения глобальной базы научных знаний [4]. Следующим по значимости является число страниц на сайте, а затем - количество гипертекстовых ссылок на сайт с других ресурсов, позволяющее, по мысли испанских коллег, оценить актуальность и значимость сайта для профессионального сообщества.

Полагаем, что выбор критериев, индикаторов и коэффициентов остается вопросом, открытым для дискуссий.

Cybermetrics Lab отмечает серьезные проблемы с точным определением того, что понимать под «единицей анализа». Например, многие учреждения поддерживают несколько различных доменных областей. Таким образом, их реальное присутствие в Сети на самом деле должно определяться множеством адресов. Кроме того, различные подразделения одной организации могут иметь (и имеют) собственные Интернет-ресурсы с адресами, не ассоциируемыми поисковыми машинами с адресом основного сайта.

Наконец, есть случаи, когда невозможно разделить академическую и неакадемическую информацию, объединяемую в едином домене верхнего уровня. Примером может служить сайт www.helsinki.fi, который в испанском исследовании используется при замерах индикаторов. На самом деле URL Университета Хельсинки - www.helsinki.fi/yliopisto/, а существенную часть объема основного сайта занимает городская информация.

2. Целевое множество и методы исследования

В качестве целевого множества были выбраны сайты классических университетов Северо-Западного федерального округа РФ (СЗФО) и Финляндии как вузы с примерно аналогичной организационно-управленческой структурой и научным потенциалом преподавательского состава.

В Финляндии это университеты Хельсинки, Йоенсуу, Оулу, Ювяскюля, Куопио, Тампере, Турку, Вааса, Лапландский Университет и Университет Або Академии (так называемый «Шведский университет»). В СЗФО – Петрозаводский, Сыктывкарский, Санкт-Петербургский и Новгородский государственные университеты, Поморский государственный университет (Архангельск) и Российский государственный университет им. Канта (Калининград). Данные по университетам, используемые в работе, взяты из официальных Интернет-источников - Федерального портала «Российское образование» (http://www.edu.ru) и сайта Министерства образования Финляндии (http://www.minedu.fi). В случае необходимости использовалась также информация, размещенная на официальных сайтах университетов.

Выбор в качестве географической зоны исследований Финляндии и СЗФО России вполне объясним, если учесть территориальную близость, давние связи между вузами, а также немаловажный для целей исследования фактор использования английского языка не как родного, а в качестве второго.

Ранее уже отмечалось наличие так называемой «проблемы единицы анализа». Если говорить об Интернет-ресурсах университета в целом, то это сложный информационный комплекс, являющийся, в каком-то смысле, отражением организационно-управленческой структуры вуза. Как правило, в этом комплексе имеется основной (официальный) сайт вуза, а далее следуют сайты факультетов, институтов и кафедр, библиотечный сайт, страницы преподавателей и т.д. Некоторые подразделения вуза (в особенности это свойственно подразделениям, профессионально связанным с информационными технологиями), имеют Интернет-ресурсы, зарегистрированные под именами, не содержащими доменного имени основного сайта вуза. В этом случае только содержательный анализ ресурса может дать ответ на вопрос, является ли этот ресурс частью Интернет-ресурсов университета.

В качестве единиц анализа мы приняли URL главных страниц официальных сайтов университетов, при этом не являющихся директориями некоторого домена (что соответствует и подходам Cybermetrics Lab). Поэтому из заявленного ранее списка классических университетов были вынужденно удалены ресурсы университетов Хельсинки, Йоенсуу и Оулу, как не удовлетворяющие данному требованию (а вот этого у Cybermetrics Lab не сделано). Далее из перечня о 4000 университетах мира («Top 4000 Universities» на сайте [3]) были отобраны сведения об исследуемых сайтах, приведенные в Таблице 1.

Поскольку для дальнейшего рассмотрения нам понадобятся значения WR, из дальнейшего рассмотрения также были исключены сайты Сыктывкарского и Новгородского университетов, как не вошедшие в Top 4000.

Таблица 1.
Выборка от 20 октября 2007 года.

WR	УНИВЕРСИТЕТЫ	URL	RankS	RankV	RankR	RankSc
229	Университет Тампере	www.uta.fi	295	200	410	321
320	Университет Ювяскюля	www.jyu.fi	304	390	199	482
413	Университет Турку	www.utu.fi	387	447	369	823
600	Университет Або Академия	www.abo.fi	635	753	473	742
834	Университет Вааса	www.uwasa.fi	721	983	866	1 406
839	Университет Куопио	www.uku.fi	783	1 170	544	883
900	Санкт-Петербургский государственный университет	www.spbu.ru	999	1 162	656	914
1902	Лапландский Университет	www.ulapland.fi	1 525	2 293	2 367	2 602
2001	Петрозаводский государственный университет	petrsu.karelia.ru	2 493	2 283	1 387	2 618
3231	Поморский государственный университет (Архангельск)	www.pomorsu.ru	2 727	3 979	4 037	3 019
3343	Российский государственный университет им. Канта (Калининград)	www.albertina.ru	3 451	3 380	6 979	1 937

По аналогии с Cybermetrics Lab замеры индикаторов S, V и R проводились с помощью зарубежных поисковых машин, перечисленных в разделе 1. Поскольку изначально авторам было понятно, что российские поисковые системы применительно к данным индикаторам покажут результаты, весьма отличающиеся от зарубежных поисковых машин, индикаторы S , V и R были замерены также с помощью Яндекса как наиболее распространенной российской поисковой системы.

По поводу индикатора Sc следует сказать, что в России единственной поисковой системой, аналогичной Google Scholar, является поисковая машина проекта Scholar.Ru (бета-версия), которая пока не слишком

пригодна для указанных целей. Поэтому в рамках исследования мы ограничились использованием данных, полученных испанскими коллегами.

Следует также четко понимать, что измеренное значение индикатора – это значение, полученное данной поисковой машиной в конкретный момент.

3. Результаты измерений и ранжирования

Результаты измерений индикаторов приведены в таблице 2.

Подробные комментарии по поводу измерений каждого индикатора будут сделаны далее. Сейчас стоит отметить лишь то, что индикатор V с помощью Яндекса измерялся не с помощью поисковой машины, поскольку в ней отсутствует соответствующая опция, а с использованием хостграфа Яндекса. Поясним, что «хостграф - граф ссылок между хостами, известных поисковой системе Яндекс. Все ссылки, исходящие с разных страниц одного хоста на разные страницы другого хоста, заменяются на одну» [5].

Хостграф был передан одному из авторов статьи в рамках проекта «Интернет-математика – 2007» [6] с разрешением использовать его в дальнейших научных исследованиях, за что авторы приносят глубокую благодарность Компании Яндекс. Количество хостов в хостграфе на декабрь 2006 года составляло примерно 2 млн. 700 тыс., поэтому определение индикатора V с помощью такого подхода кажется обоснованным.

Таблица 2.
Измерения 10-17 октября 2007 г.
(V по хостграфу Яндекса - за декабрь 2006 г.)

URL	S					R		V			
	Google	Yahoo	Live Search	Exalead	Yandex	Google	Yandex	Yahoo	Live Search	Exalead	Yandex
www.spbu.ru	1830	2360	847	572	81497	214	**6559**	38500	330	2083	1936
petrsu.karelia.ru	2870	0	7615	3660	38044	273	**1238**	4280	57	5720	940
www.albertina.ru	15	45	567	16	6177	13	**212**	1440	159	845	371
www.pomorsu.ru	3700	18613	1111	543	25867	274	**797**	1690	136	550	713
www.jyu.fi	65000	192000	12927	4462	0	3568	**0**	47100	6150	4346	299
www.uku.fi	50500	89400	14629	4358	0	5156	**0**	5010	5680	4785	189
www.ulapland.fi	27100	40441	5336	1153	0	753	**0**	13200	290	5594	18
www.uta.fi	60000	185310	38528	1128	0	7880	**0**	18700	11800	1878	380
www.utu.fi	6740	18547	7821	5605	0	877	**0**	9550	4000	4306	189
www.uwasa.fi	10100	27886	11515	4433	0	1286	**0**	6520	7900	3533	267
www.abo.fi	18800	27900	19883	10342	0	3678	**0**	23000	5950	11301	313

В Таблице 3 приведены значения рейтингов, вычисленные для нашего целевого множества по данным таблицы 2. Поясним обозначения колонок:

– WR (в соответствии с Табл.1) – сохранен порядок Top 4000, приведенный в таблице 1;

– WR (по данным Табл.2) – рейтинги вычислены в точном соответствии с методиками Cybermetrics Lab на основании замеров, приведенных в таблице 2;

– WR (с добавлением измерений по Yandex) - рейтинги вычислены в соответствии с методиками Cybermetrics Lab на основании замеров, приведенных в таблице 2 с добавлением замеров с помощью Яндекса.

–

Таблица 3.

УНИВЕРСИТЕТЫ	WR (в соответствии с Табл.1)	WR (по данным Табл.2)	WR (с добавлением измерений по

			Yandex)
Университет Тампере	1	2	2
Университет Ювяскюля	2	1	1
Университет Турку	3	8	8
Университет Або Академия	4	3	4
Университет Вааса	5	7	6
Университет Куопио	6	5	5
Санкт-Петербургский государственный университет	7	4	3
Лапландский Университет	8	6	6
Петрозаводский государственный университет	9	9	9
Поморский государственный университет (Архангельск)	10	10	10
Российский государственный университет им. Канта (Калининград)	11	11	11

Отметим высокую степень корреляции рейтингов, полученных по каждому из вариантов. При этом даже беглый взгляд на таблицу 2 вызывает целый ряд вопросов, на которые мы попробуем ответить ниже.

4. О целевом множестве и применимости «измерительных приборов»

Ранее уже отмечалось, что некоторые подразделения вуза могут иметь свои собственные Web-ресурсы, зарегистрированные под именами, не содержащими доменного имени основного сайта вуза. Этот вопрос является весьма существенным по причине того, что, ограничиваясь в рамках исследования только официальными сайтами университетов, мы на самом деле уходим от формулировки основной задачи – вместо «Вебометрического ранга университетов мира» имеем «Вебометрический ранг официальных сайтов университетов мира», что практически меняет тему исследования.

Официальные сайты университетов не предназначены для решения задачи функционального объединения глобальной базы научных знаний; скорее она решается в рамках сайтов научных подразделений и электронных библиотек вуза. Поэтому сомнительно включать в оценки рейтинга официального сайта вуза такой индикатор, как Sc - «научность сайта». Конечно, он имел бы право на существование в случае, если под единицей анализа понималась бы вся совокупность Интернет-ресурсов университета, но из методологии Cybermetrics Lab этого не следует.

Остановимся подробнее на выборе в качестве средств измерения поисковых машин.

Во-первых, определение значений индикаторов с помощью поисковых машин сразу же вызывает необходимость обоснования того, почему именно данные машины выбраны. Как можно видеть из таблицы 2, все западные поисковые машины существенно занижают значения индикаторов для российских сайтов по сравнению с Яндексом. В свою очередь, Яндекс «не видит» страниц финских сайтов, хотя и индексирует ссылки на них. Такая ситуация вполне объяснима привязкой поисковых машин к индексируемым сайтам 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).

Во-вторых, измерение количества страниц на сайте с помощью поисковой машины не является корректным хотя бы потому, что механизмы индексации страниц являются «черным ящиком» и секретом разработчиков. Если этот индикатор должен показывать реальное количество страниц сайта, то надежней измерять его с помощью программ-краулеров (например, [7]), дающих точные оценки. В противном случае следует говорить о размере сайта с точки зрения поисковой машины.

Поэтому лучшее, что можно сделать в случае использования замеров поисковыми машинами, – это взять максимальное значение из количества страниц, найденных разными поисковыми машинами. Это наверняка ближе к истине, чем применение алгоритма отбрасывания максимального и минимального и суммирование остальных значений (см. раздел 1). То же самое можно отнести и вычислению итогового значения количества полнотекстовых файлов.

Что касается индикатора цитируемости V, оценивающего актуальность и значимость сайта для профессионального сообщества, то он действительно может измеряться поисковыми машинами, поскольку характеризует ссылочную популярность с точки зрения данной поисковой машины. В случае Яндекса может быть также использован тематический индекс цитирования (ТИЦ), характеризующий популярность сайта или, как в нашем случае, данные по хостграфу Яндекса.

Однако в этом случае интегральный показатель должен являться некоторой взвесью замеров. Например, в случае равнозначности поисковых машин, сайты могут ранжироваться по замерам каждой машины (наибольший результат получает ранг 1), затем ранги по сайту суммируются и суммы вновь ранжируются, давая в итоге RankV.

Небезынтересно отметить, что только Yahoo Search предоставляет возможность исключить из множества ссылок на заданный сайт ссылки, сделанные с этого же сайта. Эта особенность имеет немаловажное значение, поскольку беглый анализ первых 100 ссылок на официальный сайт Петрозаводского государственного университета показал, что 58% ссылок на него сделано с его же страниц, а 24% - со страниц других Интернет-ресурсов Петрозаводского государственного университета. Причем это характерно не только для Петрозаводска, например, у Лапландского Университета 8% ссылок на основной сайт сделано с сайта Арктического центра Лапландского Университета (http://www.arcticcentre.org) и 72% - с собственных страниц.

В заключение приведем таблицу ранжирования официальных сайтов университетов, полученную в результате уточнений к модели, сделанных в предыдущем разделе. Интуитивно такое ранжирование официальных сайтов университетов кажется более близким к реальности.

Таблица 4.

ранг	УНИВЕРСИТЕТЫ
1	Университет Тампере
2	Университет Ювяскюля
3	Санкт-Петербургский государственный университет
4	Университет Або Академия
5	Университет Куопио
6	Петрозаводский государственный университет
7	Университет Вааса
8	Лапландский Университет
9	Университет Турку
10	Поморский государственный университет (Архангельск)
11	Российский государственный университет им. Канта (Калининград)

Сайты, составляющие Интернет-ресурсы университета, создаются для достижения разных целей с различным акцентированием на тех или иных функциях. Достаточно очевидно, что официальный сайт университета имеет иную цель, нежели сайт электронной библиотеки. Возможно, основной трудностью при разработке моделей ранжирования сайтов как раз и является отсутствие четко сформулированных целей их создания. Вместе с тем, мы должны уметь формулировать эти цели, если хотим уметь оценивать эффективность наших затрат.

Возможно, вебометрические исследования рейтингов сайтов смогут послужить основой для формулирования этих целей посредством количественного сравнительного анализа некоторых характеристик уже созданных ресурсов. А, значит, позволят определить и направления их дальнейшего развития.

Литература

1. Mike Thelwall, Liwen Vaughan, Lennart Björneborn (2005). 'Webometrics'. Annual Review of Information Science and Technology, 39, pp. 81-135.

2. Portal de estudios cuantitativos en Internet. [Электронный ресурс] – 2007. – Режим доступа: http://internetlab.cindoc.csic.es.

3. Webometrics Ranking of World Universities. [Электронный ресурс] – 2007. – Режим доступа: http://www.webometrics.info.

4. "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities", Conference on Open Access to Knowledge in the Sciences and Humanities, October 20-22, 2003, Berlin, [Электронный ресурс] – 2007. – Режим доступа: http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.

5. Наборы данных. Набор данных "Хостграф". [Электронный ресурс] – 2007. – Режим доступа: http://company.yandex.ru/grant/datasets_description.xml.

6. А.А.Печников, Ю.В.Чуйко Математические модели согласованного поведения малых Интернет-сообществ. Интернет-математика 2007, Екатеринбург: Изд-во Уральского университета. - С. 164-170.

7. SocSciBot 3. Link crawler for the social sciences. [Электронный ресурс] – 2007. – Режим доступа: http://socscibot.wlv.ac.uk.