ПЕЧНИКОВ Андрей Анатольевич *-*
доктор технических наук, доцент,
ведущий научный сотрудник лаборатории
телекоммуникационных систем
Института прикладных математических
исследований Карельского научного центра РАН
(ИПМИ КарНЦ РАН)
Адрес: 185910, Республика Карелия,
г. Петрозаводск, ул. Пушкинская, 11
e-mail: pechnikov@krc.karelia.ru

О некоторых тенденциях изменения связности российского академического Веба*

Введение

Некоторые результаты исследований связности российского академического Веба были опубликованы автором в 2009 году [1,2]. На множестве из 288 официальных сайтов РАН был построен веб-граф, множество вершин которого соответствуют множеству сайтов. Множество дуг веб-графа определяется следующим образом: дуга, соединяющая любую пару вершин, существует тогда и только тогда, когда существует хотя бы одна гиперссылка, сделанная с сайта, соответствующего вершине-источнику дуги, на сайт, соответствующий вершине-приемнику. Построенный таким образом веб-граф является ориентированным графом без петель и кратных дуг. В частности, была обнаружена максимальная компонента сильной связности, содержащая 175 сайтов [2]. Используемые алгоритмы сканирования сайтов и обработки полученных данных описаны в работе [3].

В 2013 году в рамках выполнения проекта «Информационная система вебометрического ранжирования веб-ресурсов научных учреждений России» [4] было проведено сканирование 343 веб-сайтов РАН с целью нахождения исходящих внешних гиперссылок для вычисления индикатора внутренней ссылочной популярности веб-сайтов РАН. Сканирование сайтов было выполнено с помощью программы BeeCrawler [5]. В процессе реализации проекта была разработана база данных внешних гиперссылок. Для исследователей Веба реализован гостевой удаленный доступ по адресу *http://grid.krc.karelia.ru/webometrics/main.php* (имя пользователя: guest, пароль: guest).

Поскольку архивы исследования за 2009 год были сохранены, представилась возможность провести сравнение двух веб-графов РАН и увидеть, какие изменения произошли за почти пятилетний период. Далее для анализа веб-графов используется программный комплекс Gephi - открытая интерактивная платформа визуализации и анализа сетей, сложных систем и графов [6]. Основные результаты этого анализа приводятся в статье.

Множество сайтов РАН для сравнительного анализа

При сравнении множеств веб-сайтов РАН за 2009 и 2013 год обнаруживается их совпадение только по 225 сайтам, которые были взяты в качестве целевого множества для дальнейшего исследования. Казалось бы, множество 2013 года должно почти полностью включать в себя сайты множества 2009 года, однако этого не произошло. Основным объяснением такой ситуации является несовпадение доменных имен сайтов. В проекте 2013 года при фор-

Статья подготовлена при частичной поддержке Российского гуманитарного научного фонда (грант № 12-03-12001).

мировании множества сайтов, идентифицируемых по их доменным именам, обязательным условием являлась их индексация в Яндексе как основных. Например, сайт Сибирского отделения РАН имеет, по крайней мере, два доменных имени www-sbras.nsc.ru и www.sbras.ru, но Яндекс индексирует только первое из них, а второе считает его зеркалом. К сожалению, в 2009 и 2013 годах использовались разные доменные имена сайта Сибирского отделения РАН, и он не вошел в указанные 225 сайтов. Кроме того, значительное количество сайтов РАН за 5 лет поменяло доменные имена.

Тем не менее, исследуемые 225 сайтов являются достаточно представительной выборкой, включающей:

- 1 сайт Российской академии наук;
- 3 сайта научных отделений (математических наук; нанотехнологий и информационных технологий; энергетики, машиностроения, механики и процессов управления);
- 2 сайта региональных отделений (Дальневосточное, Уральское);
- 8 сайтов региональных научных центров (Самарский, Дагестанский, Карельский, Кольский, Санкт-Петербургский, Южный, Пущинский, Троицкий);
- 5 сайтов научных центров Сибирского отделения РАН (Бурятский, Иркутский, Кемеровский, Томский, Тюменский);
- 4 сайта научных центров Уральского отделения РАН (Коми, Пермский, Удмуртский, Челябинский);
- 1 сайт научного центра Дальневосточного отделения РАН (Камчатский);
- 201 сайт институтов, вычислительных центров, музеев.

Результаты сравнительного анализа веб-графов 2009 и 2013 года

Веб-граф, построенный по данным 2009 года, имеет 225 вершин и 898 дуг. Результаты анализа продемонстрированы на **рис. 1** с использованием бабочки Бродера [7], несколько адаптированной для нашей задачи.

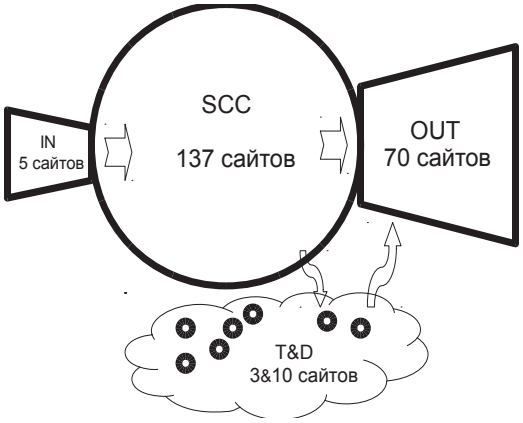

Рис. 1. *Бабочка Бродера веб-графа 2009 года*

Множество вершин веб-графа разбивается на 4 подмножества: *SCC* - вершины, составляющие максимальную компоненту сильной связности, *IN* - вершины, имеющие только исходящие дуги, соединяющие их с вершинами *SCC, OUT* - вершины, имеющие только входящие дуги, исходящие с вершин *SCC*, а *T&D (Tubes and Disconnected)* - изолированные вершины и вершины, которые связаны входящими дугами с вершинами из *IN* и исходящими на вершины *OUT*.

В данном случае обнаружена единственная компонента сильной связности *SCC*, содержащая 137 вершин, 5 вершин принадлежат *IN*, 70 вершин - *OUT*. Кроме того, имеется 10 изолированных вершин и на 3 вершины имеются ссылки с *SCC*, а с них имеются ссылки на *OUT*. Для *SCC* диаметр равен 7, а средняя длина пути между любой парой вершин - 2,409.

Одной из характеристик значимости вершин целевого множества может служить аналог *PR* [8], вычисляемый на заданной структуре веб-графа. В данном случае вершины с большими значениями

Таблица 1

Название учреждения РАН	Доменное имя	PR	Autority	Hub
Российская академия наук	www.ras.ru	0,131	0,089	0,113
Карельский научный центр РАН	www.krc.karelia.ru	0,022	0,015	0,020
Институт вычислительных технологий СО РАН	www.ict.nsc.ru	0,021	0,017	0,022
Уральское отделение РАН	www.uran.ru	0,018	0,019	0,025
Дальневосточное отделение РАН	www.febras.ru	0,017	0,016	0,021
Зоологический институт РАН	www.zin.ru	0,014	0,011	0,014
Институт цитологии и генетики СО РАН	www.bionet.nsc.ru	0,012	0,014	0,018
Пущинский научный центр РАН	www.psn.ru	0,011	0,009	0,012
Институт прикладных математических исследований КарНЦ РАН	mathem.krc.karelia.ru	0,009	0,011	0,015
Институт проблем химической физики РАН	www.icp.ac.ru	0,009	0,008	0,010

PR имеют также и большие оценки авторитетности и посредничества ('authority' и 'hub' по Клейнбергу [9]). Первые десять сайтов, соответствующие вершинам веб-графа, для которых получены наибольшие значения *PR,* приводятся в **таблице 1**. Все вершины, соответствующие сайтам из таблицы 1, входят в подмножество *SCC*.

Веб-граф 2013 года имеет, естественно, те же 225 вершин, но уже 998 связывающих их дуг, что свидетельствует об их приросте на 100 единиц по сравнению с 2009 годом. Бабочка Бродера для данного графа изображена на **рис. 2.**

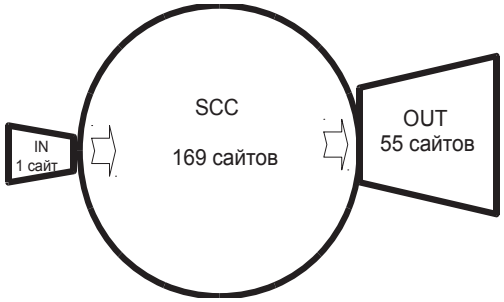

***Рис. 2**. Бабочка Бродера веб-графа 2013 года*

Несложно убедиться в том, что подмножество *SCC* возросло более чем на 23%, исчезло подмножество *T&D*, всего одна вершина осталась в подмножестве *IN* и существенно уменьшилось подмножество *OUT.* Диаметр *SCC* стал равен 6, а средняя длина пути между любой парой вершин - 2,401.

В **таблице 2** приведены первые десять сайтов, у которых вершины веб-графа имеют наибольшие значения *PR.*

Сравнивая таблицы 1 и 2, можно отметить, что в обоих случаях в первой десятке сайтов оказались сайты Российской академии наук, Сибирского и Дальневосточного отделений, Карельского научного центра и Института вычислительных технологий СО РАН. По сравнению с 2009 годом в первую десятку вошли 5 других институтов РАН.

В работе [2] было отмечено, что для формирования связного академического Веба важную роль играют гиперссылки между сайтами, соответствующие административным отношениям учреждений РАН (так называемый «административный каркас» академического Веба).

Наличие официального сайта РАН, двух сайтов региональных научных отделений и регионального научного центра в первой десятке как в 2009, так и в 2013 году подчеркивает этот факт. К примеру, значительное количество гиперссылок с сайта РАН на сайты научных учреждений РАН сделано со страницы «Информационные системы научных учреждений Российской академии наук» *(http://www.ras.ru/sciencestructure/informationsystems.aspx)*.

И, наоборот, многие сайты учреждений РАН имеют ссылку на официальный сайт РАН, как, например, сайт Института археологии РАН - в виде соответствующего значка в правом верхнем углу каждой html-страницы.

Результаты сравнительного анализа веб-графов (без сайтов «административного каркаса»)

Представляет интерес изменение связности веб-графа официальных учреждений РАН без влияния административных воздействий, особенно в свете реформирования РАН и переподчинения институтов РАН Федеральному агентству научных организаций (ФАНО). Для этого рассмотрим веб-граф, в котором оставлена 201 вершина (и связывающие их дуги), соответствующая только сайтам институтов, вычислительных центров и музеев (исключив вершины 24 сайтов собственно РАН, отделений и научных центров). Далее веб-граф, содержащий 225 вершин, будем называть «основным», а граф, полученный из него удалением 24 указанных вершин, - «уменьшенным».

Таблица 2

Название учреждения РАН	Доменное имя	PR	Autority	Hub
Российская академия наук	www.ras.ru	0,169	0,115	0,135
Уральское отделение РАН	www.uran.ru	0,025	0,024	0,028
Дальневосточное отделение РАН	www.febras.ru	0,017	0,014	0,016
Институт вычислительных технологий СО РАН	www.ict.nsc.ru	0,013	0,013	0,015
Институт физики твердого тела РАН	www.issp.ac.ru	0,013	0,011	0,013
Институт археологии РАН	www.archaeolog.ru	0,010	0,006	0,007
Карельский научный центр РАН	www.krc.karelia.ru	0,010	0,011	0,013
Институт философии РАН	iph.ras.ru	0,009	0,009	0,011
Физико-технический институт им. А.Ф. Иоффе РАН	www.ioffe.ru	0,009	0,011	0,013
Институт физики металлов УрО РАН	www.imp.uran.ru	0,009	0,010	0,012

Таблица 3

Название учреждения РАН	Доменное имя	Autority	Hub	Out	In
Институт вычислительных технологий СО РАН	www.ict.nsc.ru	0,025	0,036	5	13
Институт цитологии и генетики СО РАН	www.bionet.nsc.ru	0,023	0,036	3	12
Институт проблем химической физики РАН	www.icp.ac.ru	0,019	0,033	6	10
Зоологический институт РАН	www.zin.ru	0,016	0,024	10	8
Институт ядерной физики им. Г.И. Будкера СО РАН	www.inp.nsk.su	0,016	0,024	6	8
Институт прикладной математики им. М.В. Келдыша РАН	www.keldysh.ru	0,016	0,021	11	8

Уменьшенный веб-граф 2009 года имеет 419 дуг, то есть средняя инцидентность вершины по сравнению с основным веб-графом уменьшилась практически в 2 раза. Можно также отметить, что вместе с 24 вершинами основного графа были удалены и 479 инцидентных им дуг, что еще раз подчеркивает важность «административного каркаса». При этом связность веб-графа очень сильно пострадала: максимальная компонента сильной связности содержит 58 вершин (менее 30% от общего количества), 41 вершина является изолированной, а 67 вершин имеют только входящие ссылки, не имея исходящих.

В уменьшенном веб-графе 2013 года обнаруживается 432 дуги, и, хотя их суммарный прирост невелик, далее будет понятно, что они иначе перераспределены между 201 вершиной. В данном случае на 24 удаленные вершины основного графа пришлось 566 удаленных дуг, то есть «административный каркас» основного веб-графа за 4 года прирос довольно значительно.

В уменьшенном веб-графе 2013 года по сравнению с 2009 годом произошли следующие изменения:

- количество изолированных вершин сократилось с 41 до 31;
- почти на 20% уменьшилось количество вершин, имеющих только входящие ссылки (до 54);
- компонента сильной связности увеличилась до 84 вершин (и достигла почти 42% от общего числа).

В случае уменьшенного веб-графа уже нельзя утверждать, что вершины с большими значениями *PR* также имеют и большие оценки *Autority* и *Hub*. Наибольшие оценки *Autority* и *Hub*, в свою очередь, имеют уже не первые 10, а только первые 6 вершин при упорядочении их по убыванию. В **таблице 3** для шести сайтов, соответствующих таким вершинам уменьшенного веб-графа 2009 года, указаны значения не только *Autority* и *Hub*, но также и количество входящих *(In)* и исходящих *(Out)* дуг, инцидентных этим вершинам.

В **таблице 4** приведены те же данные для уменьшенного веб-графа 2013 года. За четыре года сохранилось лидирующее место сайта Института вычислительных технологий СО РАН. Значительное увеличение количества исходящих ссылок связано с развитием проекта «Рейтинг сайтов научных учреждений СО РАН» *(http://www.ict.nsc.ru/ranking)*. Сохранил четвертое место сайт Зоологического института РАН, но инцидентность соответствующей вершины веб-графа значительно уменьшилась.

В целом сравнение таблиц 3 и 4 позволяет отметить, что, хотя общее количество дуг увеличилось незначительно, они более ровно распределились между вершинами. Об этом говорит более равномерное распределение значений *Autority* и *Hub* в таблице 4.

По изолированным сайтам можно отметить следующий факт: 16 сайтов, не имевших входящих и исходящих гиперссылок в 2009 году, остались в этом же списке и в 2013 году, что позволяет говорить о некотором «устойчивом изолированном ядре» уменьшенного веб-графа. Следует сказать, что и визуальный просмотр сайтов этого ядра оставляет двойственное впечатление. По ряду сайтов можно сказать, что им уделяется недостаточно внимания: обнаруживаются ошибки программирования, пустые разделы, «устаревшие» новости. И,

Таблица 4

Название учреждения РАН	Доменное имя	Autority	Hub	Out	In
Институт вычислительных технологий СО РАН	www.ict.nsc.ru	0,013	0,015	45	13
Физико-технический институт им. А.Ф. Иоффе РАН	www.ioffe.ru	0,011	0,013	3	9
Институт физики твердого тела РАН	www.issp.ac.ru	0,011	0,013	2	10
Зоологический институт РАН	www.zin.ru	0,009	0,011	6	6
Институт прикладных математических исследований КарНЦ РАН	mathem.krc.karelia.ru	0,009	0,011	4	7
Институт философии РАН	iph.ras.ru	0,009	0,011	3	8

хотя по другим сайтам «ядра» таких замечаний нет, гиперссылки с них отсутствуют не только на сайты коллег из исследуемого множества сайтов, но и вообще на сайты РАН.

Заключение

В конце декабря 2013 года Правительство Российской Федерации утвердило перечень организаций, подведомственных ФАНО [10], благодаря которому исследователи российского научного Веба могут узнать, что данным распоряжением в ведение агентства переданы 433 федеральных государственных бюджетных учреждения науки Российской академии наук. Это на самом деле очень важная информация, поскольку теперь можно говорить о том, что выводы по результатам, изложенным в данной статье, касаются большей части официальных сайтов научных учреждений РАН, а, значит, можно смело говорить о наметившихся тенденциях.

Несомненно то, что за период с начала 2009 до конца 2013 года (то есть почти за пять лет) произошли определенные изменения в российском академическом Вебе, и эти изменения можно считать положительными. Обнаруживаются тенденции к существенному увеличению связности сайтов, практически не остается сайтов, на которые не сделано хотя бы одной ссылки, и существенно сокращается количество сайтов, не имеющих ссылок на другие сайты.

Вместе с тем четко прослеживаются тенденции увеличения количества гиперссылок, соответствующих административным отношениям, и практически нулевого роста гиперссылок между институтами. При этом существенно возросла коммуникационная роль официального сайта РАН, сайтов региональных отделений и центров. Резкий рост количества исходящих гиперссылок на сайте одного из институтов не в счет, так как находит свое объяснение в реализации исследовательского проекта по вебометрике.

Если рассматривать Веб как зеркало реальной жизни, то относительно научных взаимодействий можно сказать, что это зеркало достаточно хорошо отражает административные коммуникации РАН и не очень хорошо - содержательные межинститутские [3]. Конечно, при этом следует иметь в виду, что научный фрагмент Веба далеко не исчерпывается официальными сайтами учреждений науки. В проекте [4] кроме 343 официальных сайтов рассматриваются еще 548 сайтов, входящих в веб-пространства научных учреждений РАН (сайты проектов, конференций, лабораторий и т.д.). Будем надеяться, что планируемое изучение такого, по-видимому, более четкого зеркала, даст и более оптимистичные результаты.

Реформа РАН, присоединение к Российской академии наук Российской академии медицинских наук и Российской академии сельскохозяйственных наук, передача в ведение ФАНО около 800 научных учреждений приведут к большим изменениям. Эти изменения, несомненно, затронут и научное веб-пространство России. Хотелось бы, чтобы эта статья помогла создателям и участникам научного веб-пространства России сделать его чуточку лучше.

Литература:

1. Чуйко Ю.В., Печников А.А. Исследование связности российского научного Веба // Когнитивный анализ и управление развитием ситуаций (CASC'2009): Труды Международной конференции, Москва, 17-19 ноября 2009 г. - 2009. - С. 283-286.

2. Печников А.А., Луговая Н.Б., Чуйко Ю.В. О связности множества официальных сайтов РАН // Вопросы современной науки и практики. Университет им. В.И. Вернадского. - 2009. - № 12 (26). - С. 154-158.

3. Печников А.А, Луговая Н.Б., Чуйко Ю.В., Косинец И.Э. Разработка инструментов для вебометрических исследований гиперссылок научных сайтов // Вычислительные технологии. - 2009. - Том 14. - № 5. - С. 66-78.

4. Вебометрический рейтинг научных учреждений России [Электронный ресурс]. - URL: http://webometrics-net.ru (дата обращения 30.12.2013).

5. Печников А.А., Чернобровкин Д.И. Адаптивный краулер для поиска и сбора внешних гиперссылок // Управление большими системами. - 2012. - Выпуск 36. - С. 301-315.

6. Gephi, an open source graph visualization and manipulation software [Электронный ресурс]. - URL:https://gephi.org (дата обращения 31.12.2013).

7. Broder A. Graph structure in the web / A. Broder, R. Kumar, F. Maghoul1, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener // Journal of Computer Networks. - 2000. - № 33(1-6). - P. 309-320.

8. Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine // Computer Networks and ISDN Systems. - 1998. - № 30. - P. 107-117.

9. Kleinberg J. Authoritative sources in a hyperlinked environment // Journal of the ACM. - 1997. - № 46 (5). - P. 604-632.

10. Правительство Российской Федерации. Распоряжение от 30 декабря 2013 года № 2591-р [Электронный ресурс]. - URL: http://government.ru/media/files/41d4b2ee4aa4fdc62ccb.pdf (дата обращения 16.01.2014).