Исследование научного веб-пространства Сибирского отделения Российской академии наук*

Ю. И. Шокин[1], А. Ю. Веснин[2], А. А. Добрынин[2],
О. А. Клименко[1], Е. В. Рычкова[1], И. С. Петров[1]
[1]*Институт вычислительных технологий СО РАН,*
[2]*Институт математики им. С. Л. Соболева СО РАН, Новосибирск, Россия*
e-mail: vesnin@math.nsc.ru, helen@ict.nsc.ru

Представлен анализ веб-пространства Сибирского отделения Российской академии наук методами вебометрики и теории графов. Рассматривается более 90 сайтов научных организаций СО РАН. Содержание сайтов и связи между ними анализируются с помощью сервисов поисковых систем и специальных программ. Описаны критерии, использованные для составления рейтинга сайтов ведущих институтов СО РАН. Выделены сайты, на которые особенно много ссылаются российские и международные научные организации. Исследуются структурные и метрические свойства веб-графа сайтов Сибирского отделения и его фрагментов.

Ключевые слова: вебометрика, теория графов.

Введение

В современных подходах к изучению информационных процессов в World Wide Web (веб-пространстве) активно используются методы вебометрики. Термин вебометрика (webometrics) обозначает раздел информатики, в рамках которого исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к веб-пространству. Развитие этого направления началось в 1997 г. после работы Т. Алминда и П. Ингверсена [1]. Методы вебометрики носят статистический характер и не претендуют на описание всего разнообразия информационных процессов, происходящих в веб-пространстве. Поэтому, используя только данные методы, невозможно построить математическую модель веб-пространства и математически обосновать критерии оценки информационных ресурсов в интернете. В настоящей работе для анализа структуры веб-пространства привлечены методы теории графов.

Анализ свойств веб-пространства как математического объекта впервые был начат в работах Р. Алберта и А.-Л. Барабаши [2]. Возникающая проблематика включает поиск адекватных представлений веб-пространства в виде сложной сетевой структуры, исследование её свойств, нахождение математических параметров, характеризующих такую сеть, определение и предсказание изменений этих параметров при эволюции сети. Для изучения содержательных и логических связей между объектами веб-пространства удобно использовать их представление в виде веб-графа. В настоящей

*Работа выполнена при финансовой поддержке Президиума СО РАН (Междисциплинарный интеграционный проект № 21, 2012–2014 гг.) и РФФИ (грант № 12-01-00631).

работе под веб-графом понимается ориентированный граф, вершины которого соответствуют веб-сайтам. Отношение между сайтами определяется наличием ссылок с одного сайта на другой.

1. Анализ веб-пространства СО РАН методами вебометрики

Регулярные исследования университетского и академического веб-пространства ведутся в лаборатории Cybermetrics Lab исследовательского центра CSIC (Consejio Superior Investigaciones Cienthficas) в Испании. В рамках этих исследований реализуется проект "Ranking Web of World Research Centers" [3], в котором определяется рейтинг сайтов университетов и научных организаций для отдельных стран и всего мира. Для некоторых стран количество организаций, представленных в рейтинге, существенно меньше, чем реальное количество организаций, имеющих сайты. В частности, в выборке для России [4] в этом рейтинге фигурируют только 20 из более чем 90 сайтов организаций СО РАН. В табл. 1 приведены позиции сайтов организаций Сибирского отделения РАН в мировом рейтинге сайтов научных организаций по данным на июль 2012 г. (названия организаций и адреса сайтов взяты из [4]). Всего в рейтинг включены 182 сайта научных организаций России.

Начиная с 2008 г. в Институте вычислительных технологий СО РАН строятся рейтинги сайтов научных организаций Сибирского отделения РАН [5, 6]. При формиро-

Т а б л и ц а 1. Сайты организаций СО РАН в мировом рейтинге сайтов

Научная организация	Адрес сайта	Место в мировом рейтинге
Russian Academy of Sciences Siberian Branch	www.nsc.ru	42
Boreskov Institute of Catalysis RAS	www.catalysis.ru	574
Institute of Cytology and Genetics RAS	www.bionet.nsc.ru	763
Institute of Computational Technologies RAS	www.ict.nsc.ru	840
Sobolev Institute of Mathematics RAS	www.math.nsc.ru	912
Institute of Computational Mathematics and Mathematical Geophysics RAS	www.sscc.ru	1024
Budker Institute of Nuclear Physics RAS	www.inp.nsk.su	1324
Ershov Institute of Informatics Systems RAS	www.iis.nsk.su	1680
Institute of Solar-Terrestrial Physics RAS	www.iszf.irk.ru	1823
Kirensky Institute of Physics RAS	www.kirensky.ru	1829
Institute of High Current Electronics RAS	www.hcei.tsc.ru	2037
Institute of Computational Modelling RAS	icm.krasn.ru	2679
Institute of Automation and Electrometry RAS	www.iae.nsk.su	2756
Lavrentyev Institute of Hydrodynamics RAS	hydro.nsc.ru	3059
Institute of Strength Physics and Materials Science RAS	www.ispms.ru	3202
Institute of Chemical Kinetics and Combustion RAS	www.kinetics.nsc.ru	3209
Energy Systems Institute	www.sei.irk.ru	3433
Institute of Semiconductor Physics RAS	www.isp.nsc.ru	3868
Institute of System Dynamics and Control Theory RAS	www.idstu.irk.ru	5141
International Tomography Center RAS	www.tomo.nsc.ru	6339

вании рейтингов используется методика из [3]. В данной работе для оценки сайтов использовались следующие параметры.

Параметр V — видимость сайта. Его значение равно количеству внешних ссылок с других сайтов на данный ресурс. Этот параметр вычислялся посредством усреднения количества внешних ссылок, найденных с помощью поисковых систем Яндекс [7], Google [8] и Bing [9]:

$$V = (V_{\text{Яндекс}} + V_{\text{Google}} + V_{\text{Bing}})/3.$$

Параметр S — размер сайта. Значение S равно количеству веб-страниц сайта, определяемому поисковыми системами. Важно отметить, что поисковые системы не всегда корректно определяют количество веб-страниц, поэтому значение данного параметра может отличаться от реального размера сайта. Параметр S вычислялся посредством усреднения значений размера сайта, полученных с помощью указанных выше поисковых систем:

$$S = (S_{\text{Яндекс}} + S_{\text{Google}} + S_{\text{Bing}})/3.$$

Параметр R — насыщенность сайта — определялся как суммарное количество файлов форматов Adobe Acrobat (pdf), Microsoft Word (doc) и Microsoft Powerpoint (ppt), размещенных на сайте. Предполагается, что популярность сайта выше, если на нём размещены в свободном доступе документы, статьи, презентации и т. п., представленные в удобном для читателя виде. Информацию о наличии на сайте файлов указанных выше форматов позволяют получать поисковые системы Яндекс и Google. Значение параметра насыщенности вычислялось путём усреднения данных, полученных с помощью этих систем:

$$R = (R_{\text{Яндекс}} + R_{\text{Google}})/2.$$

Параметр Ic — индекс цитирования сайта. Этот параметр является мерой значимости сайта. Участники проекта [3] использовали сведения из системы Google Scholar [10]. В данном исследовании применялся также индекс цитирования Яндекса [11], который определяет “авторитетность” интернет-ресурсов с учётом не просто количества ссылок на них с других сайтов, но и качественных характеристик этих ссылок.

Определение рейтинга сайтов научных организаций СО РАН включало следующие этапы.

1. Вычисление значений параметров видимости V, размера S и насыщенности R для каждого исследуемого сайта.

2. Ранжирование значений параметров V, S, R. Массив значений параметра V для всех сайтов упорядочивался по убыванию. Сайту, имеющему максимальное значение V, был присвоен ранг $V_r = 1$. Сайтам с одинаковыми значениями V присваивались одинаковые ранги. Таким образом, сайт с минимальным значением V будет иметь ранг не более 93 (количество организаций, участвующих в исследовании).

Аналогичным образом вычислялись ранги S_r и R_r параметров S и R.

3. Вычисление ранга Ic_r индекса цитирования Ic. Сначала были независимо вычислены ранги для $Ic_{\text{Яндекс}}$ и Ic_{Google}. Затем для каждого сайта полученные ранги суммировались и величина Ic_r строилась ранжированием этих сумм. Сайт с наименьшей суммой получил ранг $Ic_r = 1$.

4. Суммирование определённых выше рангов для каждого исследуемого сайта

$$W = V_r + S_r + R_r + Ic_r.$$

5. Формирование рейтинга сайтов упорядочением значений W по возрастанию. Таким образом, итоговый ранг (позиция в текущем рейтинге) будет тем выше, чем меньше значение W. Сайтам с одинаковыми значениями W присваивались одинаковые рейтинги.

В табл. 2 представлены значения параметров V, S, R и индекса цитирования Ic_{Google} для сайтов, занимающих первые 20 мест в рейтинге (данные на 10 августа 2012 г.).

Таблица 2. Рейтинг сайтов научных организаций СО РАН

Научная организация, адрес сайта	V	S	R	Ic	Место в рейтинге
Портал СО РАН, www.sbras.ru	54863.3	73363.3	10438.0	620	1
Институт вычислительных технологий СО РАН, www.ict.nsc.ru	68066.7	107935.0	794.5	154	2
Институт цитологии и генетики СО РАН, www.bionet.nsc.ru	6045.7	9196.7	1653.0	258	2
Институт ядерной физики им. Г. И. Будкера СО РАН, www.inp.nsk.su	23608.3	5850.0	2354.5	149	4
Институт математики им. С. Л. Соболева СО РАН, www.math.nsc.ru	4226.3	7233.3	1336.5	182	5
Институт вычислительного моделирования СО РАН, icm.krasn.ru	4914.7	5742.7	5750.5	474	5
Государственная публичная научно-техническая библиотека СО РАН, www.spsl.nsc.ru	5110.0	7653.3	417.5	136	7
Институт систем информатики им. А.П. Ершова СО РАН, www.iis.nsk.su	2352.0	13562.3	591.5	105	8
Отделение ГПНТБ СО РАН, www.prometeus.nsc.ru	4896.7	12370.0	241.0	94	9
Институт автоматики и электрометрии СО РАН, www.iae.nsk.su	2815.0	3982.7	3392.5	24	10
Институт проблем освоения Севера СО РАН, www.ipdn.ru	3637.3	9320.0	1540.5	57	11
Институт неорганической химии им. А. В. Николаева СО РАН, www.nioch.nsc.ru	1788.7	4733.3	2384.0	16	12
Институт катализа им. Г. К. Борескова СО РАН, www.catalysis.ru	13441.3	178713.0	153.0	12	13
Президиум СО РАН, www.sbras.nsc.ru	5346.7	11826.7	1489.0	0	14
Институт физики им. Л. В. Киренского СО РАН, www.kirensky.ru	1424.3	3263.7	835.0	31	15
Институт теоретической и прикладной механики им. С. А. Христиановича СО РАН, www.itam.nsc.ru	1535.0	5241.0	350.0	42	16
Институт философии и права СО РАН, www.philosophy.nsc.ru	4806.0	2043.3	350.5	96	17
Институт химической кинетики и горения СО РАН, www.kinetics.nsc.ru	710.7	2002.0	1056.5	24	18
Институт вычислительной математики и математической геофизики СО РАН, www.sscc.ru	386.7	5843.7	209.0	33	19
Институт криосферы Земли СО РАН, www.ikz.ru	2001.0	3150.0	369.0	60	20

Полный рейтинг сайтов организаций СО РАН представлен в [6]. Приведённые в [6] данные позволяют проанализировать принимаемые значения параметров V, S, R. Для 44 организаций значение параметра V (количество внешних ссылок на сайт) превышает 100, для 23 организаций $V > 1000$ (рис. 1, *а*). Таким образом, 72 % сайтов имеют достаточно много внешних ссылок. Для сравнения, в 2008 г. порог 1000 ссылок на сайт преодолели только 13 организаций [5].

Размер сайтов S варьируется от нескольких десятков до ста тысяч страниц, при этом у 39 организаций сайты имеют более 1000 страниц. В 2008 г. таких сайтов было только 19. 41 организация (40 %) имеет сайты с достаточно большим числом страниц — от 100 до 1000, остальные — менее 100 страниц.

Насыщенность сайтов R (количество загруженных файлов) больше у тех организаций, которые предоставляют свободный доступ к полным текстам документов (статей, докладов, отчётов, инструкций и т. п.). 31 организация имеет сайты с количеством загруженных файлов более 100 (рис. 2, *а*), у восьми организаций на сайтах размещены более 1000 файлов форматов Adobe Acrobat (pdf), Microsoft Word (doc) и Microsoft Powerpoint (ppt). В 2008 г. таких сайтов было 5. Значение индекса цитирования $Ic_{\text{Google}} > 100$ зафиксировано у 11 сайтов (рис. 2, *б*).

Анализируя положения сайтов в рейтинге за достаточно большой период, можно увидеть, что одни сайты стабильно занимают высокие позиции, а другие постепенно

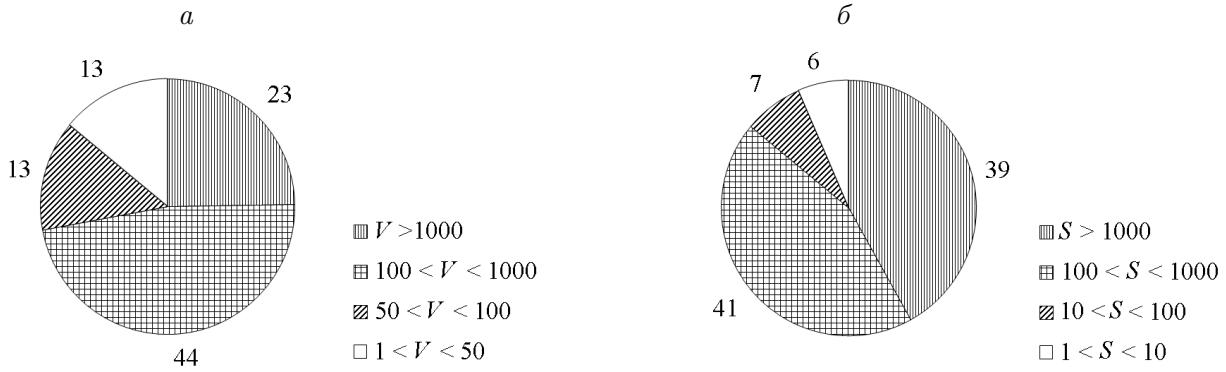

Рис. 1. Количество сайтов в зависимости от числа внешних ссылок (*а*) и веб-страниц (*б*)

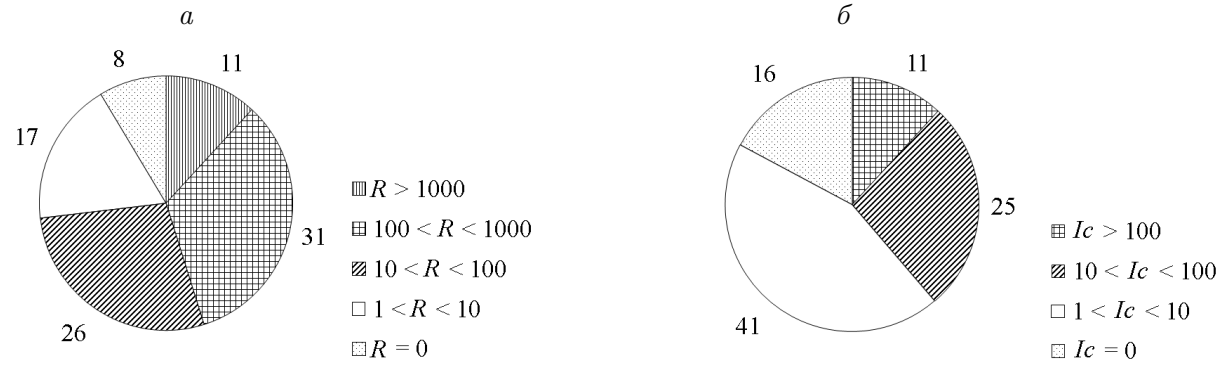

Рис. 2. Количество сайтов в зависимости от количества загруженных файлов (*а*) и величины индекса цитирования (*б*)

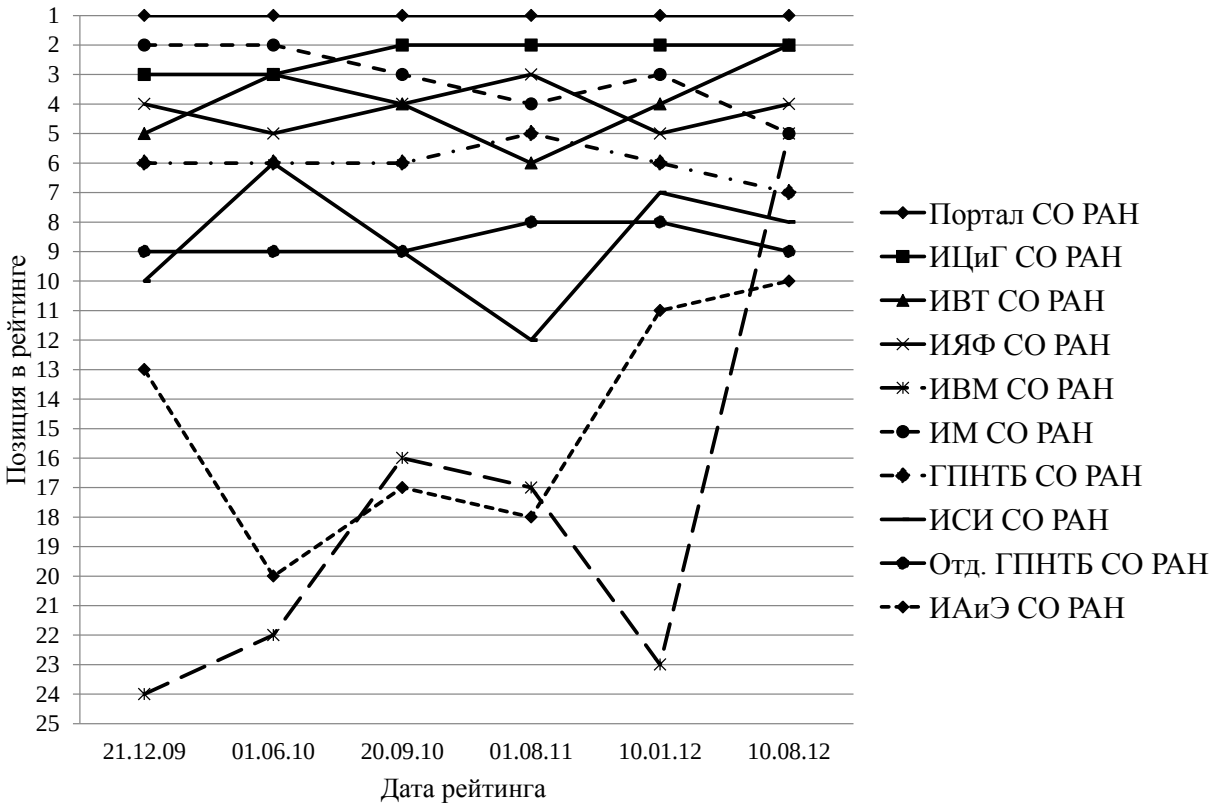

Рис. 3. Топ 10 рейтинга сайтов СО РАН с 2009 по 2012 гг.

улучшают свои характеристики. На рис. 3 приведен тренд положения в рейтинге (начиная с 2009 г.) сайтов, занимающих первые 10 позиций по состоянию на 10 августа 2012 г.

2. Анализ веб-графа научных организаций СО РАН

Рассматривается веб-граф $\mathbf{G}$, вершинам которого соответствуют сайты научных организаций СО РАН, а отношение между сайтами определяется наличием ссылок с одного сайта на другой. Дуга графа выходит из вершины v и заходит в вершину u, если сайт, соответствующий вершине v, содержит хотя бы одну ссылку на сайт, соответствующий вершине u. Таким образом, полученный веб-граф является ориентированным графом, любая пара вершин которого может быть соединена одной дугой или двумя противоположно направленными дугами. Веб-граф $\mathbf{G}$ организаций СО РАН содержит 88 вершин и 863 дуги. В этот граф были включены научные организации из Информационной системы “Организации и сотрудники СО РАН” [12], имеющие сайты на 10 января 2012 г. (к моменту построения рейтинга 10 августа 2012 г. сайтов стало 93). Диаграмма графа представлена в [13].

Так как в граф включены сайты институтов из разных областей науки, то его структура далека от графа, в котором любые две вершины соединены двумя дугами. Для оценки степени участия вершин и дуг в формировании структуры графа будем использовать численные параметры.

Первый параметр оценивает число вершин, еще не включённых в информационное взаимодействие. Индекс вершин в графе $c_v(G)$ определяется как отношение числа вершин k, имеющих хотя бы одну исходящую или входящую дугу, к числу n всех вершин графа, $c_v(G) = k/n$. Близость $c_v(G)$ к нулю указывает на большую долю изолированных вершин в G, не связанных с другими вершинами графа. При максимальном значении $c_v(G) = 1$ все сайты институтов, хотя бы попарно, вовлечены во взаимодействие друг с другом. Для рассматриваемого графа $\mathbf{G}$ организаций СО РАН $c_v(\mathbf{G}) = 1$.

Второй параметр характеризует глобальную интенсивность взаимодействия сайтов друг с другом. Индекс дуг графа G с n вершинами и t дугами задается отношением $c_a(G) = t/(n(n-1))$ (плотность сети [14]). Максимальное значение $c_a(G) = 1$ достигается на полном графе, любые две вершины которого соединены парой противоположно ориентированных дуг. В этом случае все сайты ссылаются друг на друга. Для графа сайтов организаций СО РАН выполняется $c_a(\mathbf{G}) = 0.11$.

Третий параметр характеризует локальную интенсивность взаимодействия сайтов. Под окрестностью вершины v будем понимать множество вершин графа, соединённых с v дугами без учёта их ориентации. Коэффициент кластеризации вершины v определяется как $c_a(G_v)$, где G_v — подграф, порождённый окрестностью вершины v [15]. Для графа G коэффициент кластеризации $cc(G)$ есть среднее значение по множеству вершин U, для каждой из которых общее число входящих в неё и исходящих из неё дуг не менее 2, $cc(G) = \sum\limits_{v \in U} c_a(G_v)/|U|$. Таким образом, этот параметр показывает, как в среднем заполнена дугами окрестность вершины. Для графа сайтов организаций СО РАН коэффициент кластеризации $cc(\mathbf{G}) = 0.06$.

2.1. Характеристики связей вершин графа

Под расстоянием между парой вершин в графе понимается число дуг в кратчайшем ориентированном пути, соединяющем эти вершины. Естественными характеристиками вершины v ориентированного графа являются число исходящих из неё дуг $deg_+(v)$ (полустепень исхода) и число входящих в неё дуг $deg_-(v)$ (полустепень захода). Увеличение полустепеней вершин графа вызывает в общем случае возрастание его компактности, под которой понимается уменьшение расстояний между вершинами и, как следствие, уменьшение диаметра графа (максимального расстояния между его вершинами). Если в графе не учитывается ориентация дуг (неориентированный граф), то последнее справедливо в ещё большей степени. Исходящие и входящие дуги вместе с вершиной образуют легко распознаваемые локальные фрагменты, которые могут быть использованы в качестве классификационных признаков вершин. В неориентированном графе степень $deg(v)$ вершины v равна сумме её полустепеней исхода и захода:

Т а б л и ц а 3. Распределение вершин графа $\mathbf{G}$ по полустепеням исхода

deg_+	0	1	2	3	4	5	6, 7	8, 9	10	11	12	13, 15, 16, 18, 25	26	43, 77, 82	83	87
N	17	13	11	4	7	2	4	2	3	4	2	1	2	1	2	1

Т а б л и ц а 4. Распределение вершин графа $\mathbf{G}$ по полустепеням захода

deg_-	1	4	5	6	7	8	9	10	11	12, 13	14	15, 17, 18, 19, 27, 29, 38, 48
N	3	4	9	7	15	6	7	10	7	5	2	1

Т а б л и ц а 5. Распределение вершин графа **G** по сумме полустепеней

deg	1	3, 4	5	6	7	8, 9, 10	11	12	14	15	16	17	18	19
N	2	1	7	5	10	4	5	4	3	4	3	2	3	1
deg	20	22, 24	25	26, 27, 36, 41, 42, 72, 101, 102, 109, 115, 135										
N	6	3	2	1										

$deg(v) = deg_+(v) + deg_-(v)$. В табл. 3–5 приводятся данные о степенях вершин веб-графа **G** сайтов организаций СО РАН. В верхней строке таблиц указаны значения степеней, нижняя строка содержит количество вершин N с соответствующими степенями.

Минимальная и максимальная степени исхода и захода вершин равны 0, 87 и 1, 48 соответственно. Средние полустепени исхода/захода вершин равны 9.8 (сумма полустепеней исхода всегда равна сумме полустепеней захода). Если граф рассматривается как неориентированный, то минимальная степень вершин равна 1, максимальная 135, а средняя 7.6. Число вершин, из которых нет ни одной исходящей дуги, составляет около 19 % от всех вершин графа. Входящие дуги имеются у всех вершин графа. В графе есть единственная вершина, соответствующая Порталу СО РАН, из которой дуги ведут во все остальные вершины графа. В эту вершину входят дуги из 48 других вершин графа. Также большое число исходящих дуг имеют четыре вершины, соответствующие сайтам ОУС СО РАН по НИТ (83), ИВТ СО РАН (83), Отделения ГПНТБ СО РАН (82) и Президиума СО РАН (77).

2.2. Классификация типов вершин

При анализе веб-графа представляет интерес соотношение между полустепенями исхода и захода вершин. На рис. 4 приводятся три варианта возможного распределения входящих и исходящих дуг. Вершины первого типа называют индукторами (мало входящих дуг, много исходящих), второго — коллекторами (много входящих дуг, мало исходящих), третьего — посредниками (много и входящих, и исходящих дуг). Эти типы вершин образуют множество веб-коммуникаторов графа.

Коллекторы могут соответствовать организациям, в которых происходит накопление, хранение и обработка данных. Это — библиотеки, банки данных, центры коллективного пользования, справочные ресурсы. Посредниками могут быть вершины, соответствующие головным сайтам, порталам научных центров, сайтам институтов с высокой степенью научной кооперации, индукторами — сайты недавно созданных организаций или новые сайты для существующих институтов. Визуальный анализ вершин с большими степенями показывает, что в веб-графе организаций СО РАН индукторами можно назвать сайты ОУС СО РАН по НИТ (83, 19) и ИВТ СО РАН (83, 18), а по-

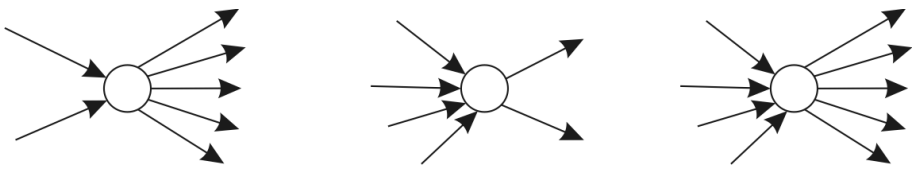

Рис. 4. Веб-коммуникаторы: индуктор, коллектор и посредник

Т а б л и ц а 6. Число индукторов и коллекторов в графе **G**

rel	2													3					
md	2	3	4	5	6	7	8	10	11	19	20	29	39	2	3	4	19	20	28
Индуктор	7	7	7	7	7	7	6	5	4	3	2	1	0	3	3	3	2	1	0
Коллектор	21	11	7	3	2	0	0	0	0	0	0	0	0	9	2	0	0	0	0

Т а б л и ц а 7. Число посредников в графе **G**

rel	1.2			1.4			1.6					1.9						
md	10	11	12	10	11	12	10	11	12	17	29	10	11	12	15	17	29	48
Посредник	2	3	2	3	4	2	4	4	2	1	1	4	4	2	1	1	1	1

средниками — Портал СО РАН (87, 48), сайты Президиума СО РАН (77, 38), ГПНТБ СО РАН (43, 29) и Отделения ГПНТБ СО РАН (82, 27) (в скобках указаны полустепени исхода и захода вершин). Отнесение вершин графа к веб-коммуникаторам того или иного типа зависит от соотношения между полустепенями. Будем характеризовать индукторы (коллекторы) двумя параметрами (md, rel), где md означает полустепень захода (исхода), а rel — отношение полустепени исхода (захода) к md. Например, если задано $(md, rel) = (5, 3)$, то индукторами будут вершины v, в которые входят $deg_-(v) \geq 5$ дуг и выходят $deg_+(v) \geq deg_-(v) \cdot rel$ дуг, а коллекторы будут определяться значениями $deg_+(v) \geq 5$ и $deg_-(v) \geq deg_+(v) \cdot rel$. В табл. 6 показано, как изменяется количество вершин указанных типов при $rel = 2$ и 3 в веб-графе **G** (приведены значения md, на которых происходит изменение числа индукторов или коллекторов).

Для поиска значимых веб-коммуникаторов при выборе значения md можно учитывать средние полустепени вершин.

Для посредников значение md задает наименьшую полустепень, а rel — отношение между полустепенями. Например, параметры $(md, rel) = (15, 1.1)$ определяют вершины-посредники, в которых меньшая полустепень составляет не менее 15, а бóльшая полустепень превышает её не более, чем на 10 %. В табл. 7 приводятся данные по числу посредников в рассматриваемом графе, указаны значения степени md, на которых происходит изменение числа посредников.

С течением времени структура веб-графа может меняться. Вершины с малой степенью могут соответствовать, например, сайтам недавно созданных институтов. Степень таких вершин будет возрастать при установлении новых связей с сайтами других институтов.

2.3. Сильно связная компонента

Для описания больших веб-графов используется представление их структуры в виде схемы галстука-бабочки [16]. В этой модели в графе выделяется максимальная сильно связная компонента, по отношению к которой классифицируются остальные вершины графа. В подграфе, называемом сильно связной компонентой графа, существует ориентированный путь между любой парой вершин. Поэтому, проходя по ссылкам соответствующих сайтов, можно обойти все вершины компоненты. Центральную часть бабочки образует максимальная сильно связная компонента. Левая часть бабочки состоит из вершин, пути из которых ведут в эту компоненту. Правую часть образуют

Рис. 5. Сильно связная компонента графа и её окружение

вершины, в которые ведут пути из компоненты (рис. 5). В сложных веб-графах имеются подмножества вершин, не попадающих в эти части бабочки. Для веб-графа **G** сайтов организаций СО РАН единственная максимальная сильно связная компонента имеет большой размер и содержит 70 вершин (всего в графе 88 вершин), левая часть бабочки не содержит вершин, а оставшиеся 18 вершин входят в правую часть.

Максимальное расстояние между вершинами графа **G** равно 4 (диаметр графа). Малый диаметр обеспечивается вершиной, соответствующей Порталу СО РАН, которая имеет максимально возможное для данного графа число исходящих (87) и большое число входящих дуг (48). Все диаметральные цепи графа начинаются в вершинах, соответствующих сайтам ИЛФ и ИрИХ СО РАН. Вторые концевые вершины этих цепей лежат как в сильно связной компоненте, так и вне её. Через вершину, соответствующую сайту ИНЦ СО РАН, проходят все диаметральные цепи.

3. Анализ веб-подграфов

При анализе веб-графа институтов авторы исходят из предположения о том, что его статическая структура, зафиксированная в какой-то момент времени, отражает текущие информационные связи между институтами. Поэтому представляется интересным исследовать веб-подграфы, соответствующие институтам по отдельным наукам, парам наук и т. д. Принадлежность института к конкретной науке определялась его вхождением в соответствующий Объединённый ученый совет СО РАН [12].

3.1. Веб-подграф химических институтов

Веб-граф **G**(X) сайтов химических институтов СО РАН содержит 11 вершин и 20 дуг. Структура графа изображена на рис. 6, в подрисуночной подписи приведены сокращённые названия институтов. В скобках после названия института указан размер его сайта (количество страниц). Вершины большего диаметра соответствуют сайтам большего размера. Если между двумя институтами есть контур длины 2, то такая пара противоположно направленных дуг будет для удобства изображаться одной двунаправленной дугой (например, дуга между вершинами 3 и 8).

Согласно классификации веб-коммуникаторов сайт ИК СО РАН соответствует коллектору (вершина 4, входящих 7 дуг и исходящих 3 дуги), сайт НИОХ СО РАН — посреднику (вершина 1, входящих 3 и исходящих 4 дуги), а сайты ИХКГ СО РАН и МТЦ СО РАН можно отнести как к индукторам, так и к посредникам (вершины 3 и 8, входящих 2 и исходящих 4 дуги). Вершины сайтов ИрИХ СО РАН и ИППУ СО РАН являются в этом веб-подграфе изолированными, т. е. не имеют никаких связей с другими вершинами. Единственная сильно связная компонента графа **G**(X) содержит все вершины, за исключением вершин 6, 7, 9, 10 и 11. Вершина 7 образует левую часть бабочки, а вершины 6 и 10 — её правую часть. Диаметр графа **G**(X) равен 2, что обес-

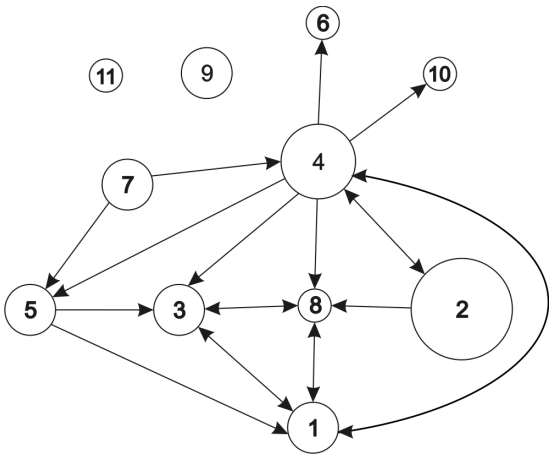

Рис. 6. Веб-подграф $\mathbf{G}(\mathrm{X})$ сайтов химических институтов СО РАН. 1 — НИОХ (3195), 2 — ИНХ (34821), 3 — ИХКГ (2478), 4 — ИК (18331), 5 — ИХТТМ (1517), 6 — ИХН (273), 7 — ИХХТ (1818), 8 — МТЦ (347), 9 — ИрИХ (1426), 10 — ИПХЭТ (276), 11 — ИППУ (106)

печивается наличием вершин с большими степенями. Минимальная и максимальная полустепени исхода и захода вершин графа равны 0, 7 и 0, 4 соответственно. Средние полустепени захода и исхода вершин равны 1.82. Для неориентированного графа $\mathbf{G}(\mathrm{X})$ минимальная степень вершин равна 0, максимальная — 8, средняя — 1.36. Индексы вершин и дуг графа принимают значения $c_v(\mathbf{G}(\mathrm{X})) = 0.82$ и $c_a(\mathbf{G}(\mathrm{X})) = 0.18$, коэффициент кластеризации $cc(\mathbf{G}(\mathrm{X})) = 0.17$. К самому заметному нарушению коммуникаций в веб-графе приводит прекращение работы сайта ИК СО РАН (вершина 4). Удаление этой вершины приводит к декомпозиции графа на большее число не связанных друг с другом подграфов, чем удаление любой другой вершины.

3.2. Веб-подграф научных центров

Веб-подграф $\mathbf{G}(\mathrm{Ц})$ головных сайтов научных центров СО РАН содержит 10 вершин и 25 дуг. Структура графа и наименования центров приводятся на рис. 7. Величина вершины отражает размер соответствующего сайта (число страниц указано в скобках).

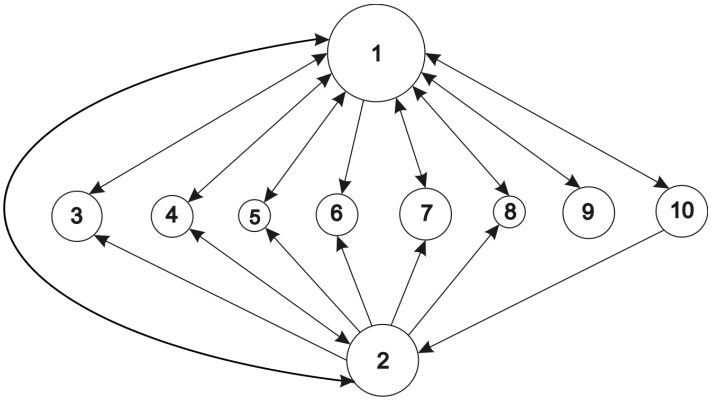

Рис. 7. Веб-подграф $\mathbf{G}(\mathrm{Ц})$ сайтов научных центров СО РАН. 1 — Портал СО РАН (143729), 2 — Президиум СО РАН (26304), 3 — ТНЦ (562), 4 — КНЦ (103), 5 — КемНЦ (47), 6 — ИНЦ (181), 7 — ТюмНЦ (308), 8 — ОНЦ (23), 9 — БНЦ (312), 10 — ЯНЦ (397)

Из диаграммы графа видно, что только Портал СО РАН и сайт Президиума СО РАН (вершины 1 и 2) объединяют сайты научных центров в связную структуру, так как между другими вершинами непосредственные связи отсутствуют. На Портал СО РАН есть ссылки почти из всех центров (7 из 8), в то время как на сайт Президиума СО РАН есть ссылки только из двух центров. По классификации веб-коммуникаторов Портал СО РАН является посредником (8 входящих и 9 исходящих дуг), а сайт Президиума СО РАН можно отнести скорее к индукторам (3 входящих и 7 исходящих дуг).

Сайт Президиума СО РАН не имеет ссылок на сайты БНЦ и ЯНЦ СО РАН. В графе есть единственная сильно связная компонента, которая содержит все вершины графа, за исключением вершины 6, т. е. с сайта ИНЦ СО РАН нельзя попасть ни на один сайт научных центров. Вершина 6 образует правую часть бабочки. Диаметр графа $\mathbf{G}(\text{Ц})$ равен 2 из-за двух вершин с большими полустепенями. Минимальная и максимальная полустепени исхода и захода вершин равны 0, 9 и 1, 8 соответственно. Средние полустепени вершин равны 2.5. Если граф рассматривается как неориентированный, то эти степени равны 1, 9 и 1.6. Индексы вершин и дуг в графе равны $c_v(\mathbf{G}(\text{Ц})) = 1$ и $c_a(\mathbf{G}(\text{Ц})) = 0.28$, значение коэффициента корреляции $cc(\mathbf{G}(\text{Ц})) = 0.07$. К полному нарушению коммуникаций в веб-графе научных центров приведёт прекращение работы сайтов Портала и Президиума СО РАН (вершины 1 и 2).

3.3. Анализ веб-графов институтов из разных областей науки

Если полагать, что идеальной структурой взаимодействия институтов в одной области науки является сильно связная компонента, между любой парой вершин которой есть контур длины 2, то для институтов из нескольких областей такая структура взаимодействия представляется нереальной. Как правило, некоторая часть институтов одного профиля связана с какими-то институтами другого профиля. Далее нас не будут интересовать связи между институтами внутри одной области науки. Поэтому в общем случае будет рассматриваться многодольный подграф, в котором все вершины согласно числу рассматриваемых областей науки разделены на несколько непересекающихся подмножеств (долей). Дуги могут соединять вершины только из разных долей. Пусть подграф G_1 имеет n_1 вершин, а подграф G_2 — n_2 вершин. Тогда индекс вершин для двудольного подграфа $G = G_1 \cup G_2$ определим как $c_v(G) = k/(n_1 + n_2)$, где k равно числу вершин, в которые входит или из которых выходит хотя бы одна дуга. Для индекса дуг графа положим $c_a(G) = t/2n_1n_2$, где знаменатель равен максимально возможному числу дуг между долями размеров n_1 и n_2. Для многодольного графа G параметры определяются аналогично. Например, для графа с тремя долями размеров n_1, n_2 и n_3 (три группы институтов) $c_v(G) = k/(n_1 + n_2 + n_3)$ и $c_a(G) = t/2(n_1n_2 + n_1n_3 + n_2n_3)$.

3.4. Веб-подграф химических и физических институтов

Веб-подграф $\mathbf{G}(\text{Х}, \Ф)$ институтов СО РАН, проводящих исследования в областях химии и физики, содержит 21 вершину и 18 дуг. Структура графа, наименования институтов и их принадлежность к долям (Х или Ф) приводятся на рис. 8. Величина вершины отражает размер соответствующего сайта (число страниц указано в скобках). Вершины физических институтов для наглядности располагаются только в центральном ряду диаграммы графа.

Вершины 7 и 9 сайтов химических институтов ИК и ИХТТМ СО РАН являются в этом графе индукторами (1 входящая дуга и 4 исходящих дуги). Изолированными

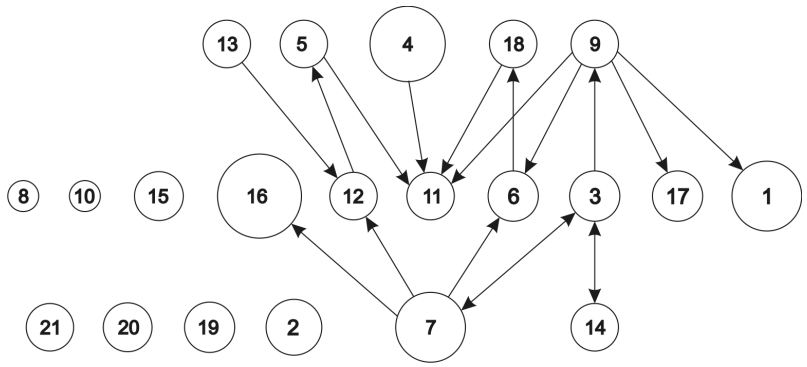

Рис. 8. Веб-граф $\mathbf{G}(\mathrm{X}, \Phi)$ институтов СО РАН в областях химии (X) и физики (Ф). 1 — ИЯФ (Ф, 14448), 2 — НИОХ (X, 3195), 3 — ИФ (Ф, 1954), 4 — ИНХ (X, 34821), 5 — ИХКГ(X, 2478), 6 — ИАиЭ (Ф, 3448), 7 — ИК (X, 18331), 8 — ИСЗФ (Ф, 0), 9 — ИХТТМ (X, 1517), 10 — ИКФИА (Ф, 0), 11 — ИФП (Ф, 1114), 12 — ИСЭ (Ф, 941), 13 — ИХН (X, 273), 14 — ИХХТ(X, 1818), 15 — КТИ НП (Ф, 380), 16 — ИОА (Ф, 297171), 17 — ИЛФ (Ф, 61), 18 — МТЦ (X, 347), 19 — ИрИХ (X, 1426), 20 — ИПХЭТ (X, 276), 21 — ИППУ (X, 106)

являются 7 вершин. Максимальная сильно связная компонента содержит всего три вершины: 3, 7 и 14. Остальные не изолированные вершины, кроме вершин 4 и 13, образуют правую часть бабочки. Из вершин 4 и 13, образующих "отростки" в модели бабочка, пути ведут в правую часть.

Диаметр графа $\mathbf{G}(\mathrm{X}, \Phi)$ равен 5. Минимальная и максимальная полустепени исхода и захода совпадают и равны 0 и 4. Средние полустепени вершин равны 0.86. Если граф рассматривается как неориентированный, то эти степени равны 0, 5 и 0.76. Индексы вершин и дуг в графе $c_v(\mathbf{G}(\mathrm{X}, \Phi)) = 0.67$ и $c_a(\mathbf{G}(\mathrm{X}, \Phi)) = 0.16$. По построению графа $\mathbf{G}(\mathrm{X}, \Phi)$ коэффициент корреляции будет всегда равен нулю, так как окрестность любой вершины целиком лежит в одной из долей и не содержит дуг. К сильному нарушению коммуникаций в веб-графе $\mathbf{G}(\mathrm{X}, \Phi)$ приведёт удаление вершин 7, 9 и 11, т. е. сайтов ИК, ИХТТМ и ИФП СО РАН.

Таким образом, среди всех рассмотренных сайтов научных организаций СО РАН наиболее развитыми в плане информационного взаимодействия являются сайты Портала СО РАН, ИВТ, ИК, ИМ, ИЦиГ и ИХБФМ СО РАН. Сайты, на которые ссылаются большое число российских и международных научных организаций, следующие — Портал СО РАН, ИВТ, ИЯФ, ГПНТБ и ИК СО РАН. Сайтами с высоким индексом цитирования являются Портал СО РАН, ИВМ, ИКФИА, ИЛ, ИЦиГ, ИМ, ИВТ, ИЯФ и ГПНТБ СО РАН.

Проведённый анализ показывает современное состояние информационной структуры взаимодействия институтов СО РАН на уровне сайтов и может способствовать дальнейшему развитию веб-пространства СО РАН.

Список литературы

[1] Almind T., Ingwersen P. Infometric analyses on the World Wide Web: Methodological approaches to 'webometrics' // J. of Document. 1997. Vol. 53, No 4. P. 404–426.

[2] Albert R., Barabási A.-L. Statistical mechanics of complex networks // Rev. of Modern Phys. 2002. Vol. 74, No 1. P. 47–97.

[3] Проект Ranking Web of World Research Centers. http://research.webometrics.info/ (дата доступа — 10.08.2012).

[4] Проект Ranking Web of World Research Centers, выборка данных по стране Россия. http://research.webometrics.info/en/Europe/Russian%20Federation (дата доступа — 10.11.2012).

[5] Шокин Ю. И., Клименко О. А., Рычкова Е. В., Шабальников И. В. Рейтинг сайтов научных организаций СО РАН // Вычисл. технологии. 2008. Т. 13, № 3. С. 128–135.

[6] Рейтинг сайтов научных организаций СО РАН. http://www.ict.nsc.ru/ranking/ (дата доступа — 10.08.2012).

[7] Поисковая система Яндекс. http://www.yandex.ru/ (дата доступа — 10.08.2012).

[8] Поисковая система Google. http://www.google.ru/ (дата доступа — 10.08.2012).

[9] Поисковая система Bing. http://www.bing.com/ (дата доступа — 10.08.2012).

[10] Система определения индекса цитирования в веб-пространстве Google Scholar. http://scholar.google.com/ (дата доступа — 10.08.2012).

[11] Индекс цитирования каталога Яндекс. http://help.yandex.ru/catalogue/?id=873431 (дата доступа — 10.08.2012).

[12] Информационная система “Организации и сотрудники СО РАН”. http://www.sbras.ru/sbras/db/ (дата доступа — 10.08.2012).

[13] Веб-граф организаций СО РАН. http://www.ict.nsc.ru/ranking/graph_sbras_2012.jpg (дата доступа — 10.01.2012).

[14] Hage P., Harary F. Structural Models in Anthropology. Cambridge Univ. Press, 1983.

[15] Watts D., Strogatz S. Collective dynamics of ’small world’ networks // Nature. 1998. Vol. 393. P. 440–442.

[16] Broder A., Kumar R., Maghoul F. et al. Graph structure in the Web // Comput. Networks. 2000. Vol. 33, No 1-6. P. 309–320.

Поступила в редакцию 5 октября 2012 г.,