Изучение структуры веб-пространства СО РАН методами вебометрики и теории графов*

Ю.И. Шокин
Институт вычислительных технологий СО РАН

А.Ю. Веснин
Институт математики им. С.Л. Соболева СО РАН
e-mail: vesnin@math.nsc.ru

А.А. Добрынин
Институт математики им. С.Л. Соболева СО РАН
e-mail: dobr@math.nsc.ru

О.А. Клименко
Институт вычислительных технологий СО РАН
e-mail: klimenko@ict.nsc.ru

Е.В. Рычкова
Институт вычислительных технологий СО РАН
e-mail: helen@ict.nsc.ru

15 апреля 2013

В работе представлен анализ веб-пространства СО РАН методами вебометрики и теории графов. Методы вебометрики используются для статистического анализа содержания сайтов и связей между ними. Для изучения информационных связей между сайтами организаций также используется их представление в виде веб-графа. Под веб-графом понимается ориентированный граф, вершины которого соответствуют веб-сайтам, а отношение между сайтами определяется существованием ссылок друг на друга. Исследуются структурные и метрические свойства этого веб-графа и его фрагментов. Вычисляются параметры для оценки информационного взаимодействия сайтов.

1. Анализ развития веб-пространства СО РАН методами вебометрики

Изучение научного веб-пространства ведется в Институте вычислительных технологий СО РАН с 2008 года на основе построения рейтинга сайтов научных организаций Сибирского отделения РАН [1]. Работа началась с анализа статистических параметров V, S, R и Ic [2].

Параметр V — видимость сайта. Его значение равно количеству внешних ссылок с других сайтов на данный ресурс; вычисляется на основе данных, полученых из поисковых систем Яндекс [3], Google [4] и Bing [5].

*Работа выполнена при поддержке Президиума СО РАН, в рамках междисциплинарного интеграционного проекта № 21, 2012-2014 гг. и РФФИ (грант № 12-01-00631)

Параметр S — размер сайта. Значение S равно количеству веб-страниц сайта, определяемому поисковыми системами.

Параметр R — насыщенность сайта — определялся как суммарное количество файлов форматов Adobe Acrobat (pdf), Microsoft Word (doc) и Microsoft Powerpoint (ppt), размещенных на сайте. Значение этого параметра определялось с помощью поисковых систем Яндекс и Google.

Параметр Ic — индекс цитирования сайта. Этот параметр является мерой значимости сайта. Для посторения рейтинга использовались данные из системы Google Scholar [6] и индекса цитирования каталога Яндекс [7].

Помимо построения собственно рейтинга сайтов СО РАН существует возможность проследить динамику развития сайтов, сравнив значения параметров V, S, R и индекса цитирования Ic_{Google} за два достаточно удаленных момента времени (2009 и 2013 гг.).

На рис. 1 показано распределение количества сайтов в зависимости от параметра V (число внешних ссылок на сайт) полученые 21 декабря 2009 года (слева) и 28 января 2013 года (справа).

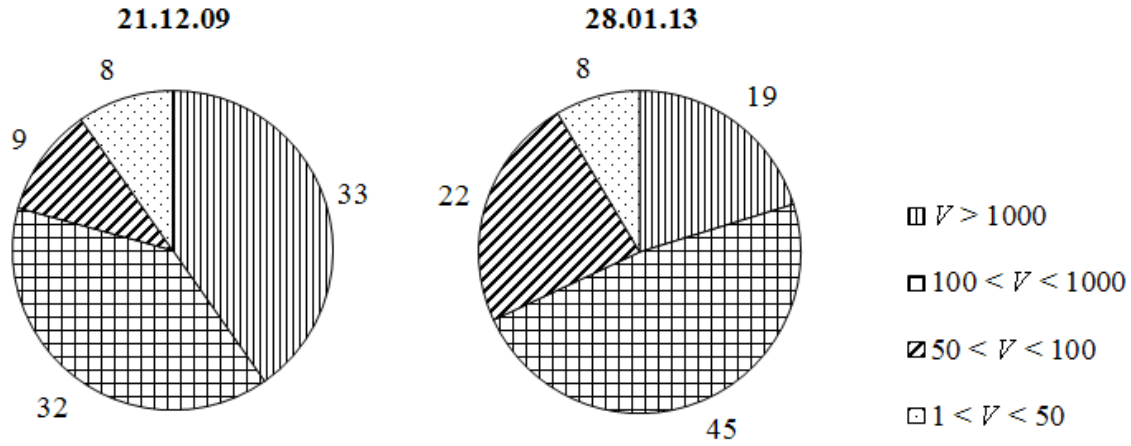

Рис. 1. Количество сайтов в зависимости от числа внешних ссылок (V)

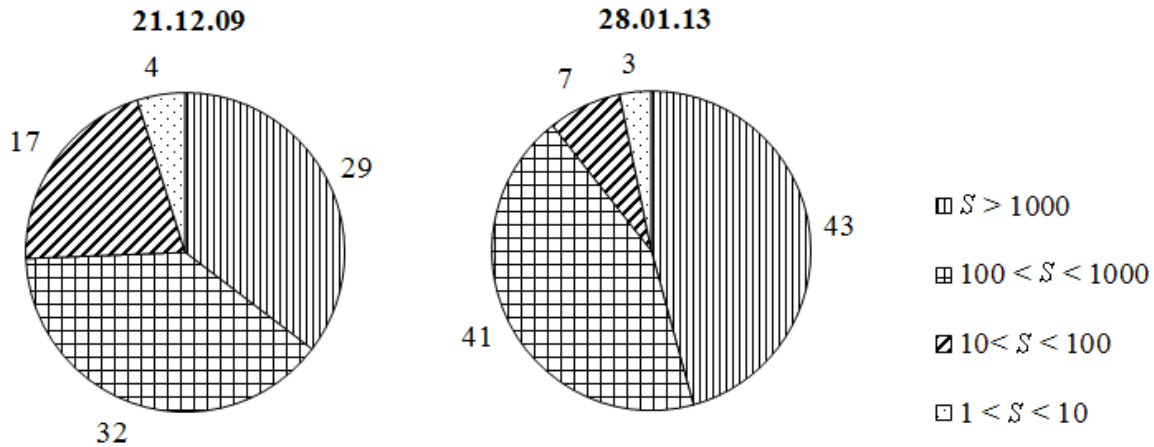

Рис. 2. Количество сайтов в зависимости от числа веб-страниц (S)

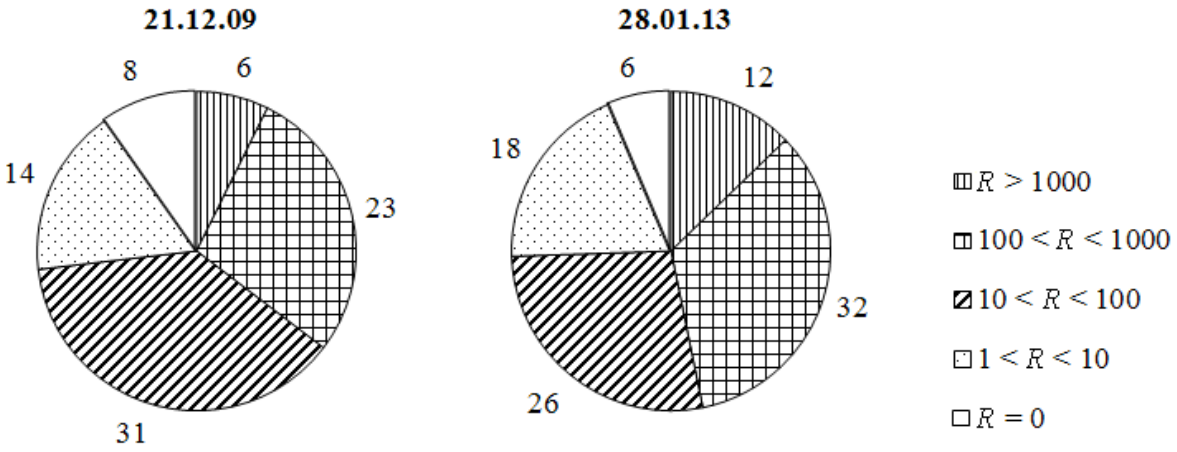

Рис. 3. Количество сайтов в зависимости от числа загруженных файлов (R)

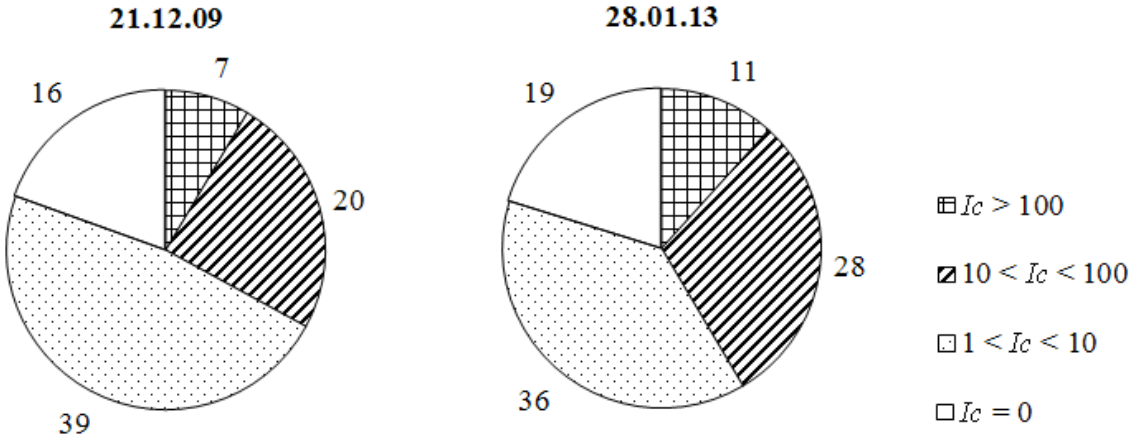

Рис. 4. Количество сайтов в зависимости от величины индекса цитирования (Ic_{Google})

Видно, что в рассматриваемый период уменьшилось количество сайтов, имеющих более 1000 внешних ссылок. Это можно объяснить тем, что Google откорректировал определение данного параметра, убрав слишком похожие ссылки из своих результатов. Кроме того, само веб-пространство — достаточно динамичная система, в которой веб-страницы как появляются, так и исчезают, поэтому и параметр V для каждого конкретного сайта может как увеличиваться, так и уменьшаться с течением времени.

С другой стороны, увеличилось количество сайтов, у которых имеется от 50 до 100 и от 100 до 1000 внешних ссылок.

Размеры сайтов S также возросли: стало больше сайтов объёмом более 1000 веб-страниц (рис. 2). Увеличилось и количество сайтов, на которые загружено более 100 дополнительных файлов (параметр R, рис. 3).

А распределение количества сайтов в зависимости от индекса цитирования Ic_{Google} изменилось мало (рис. 4), причем увеличение количества сайтов с Ic_{Google} от 10 до 100 можно, в том числе, объяснить и тем, что в 2009 году в рейтинге участвовало 82 сайта, а в 2013 году — 94 сайта.

2. Анализ веб-графа научных организаций

В работе [2] был исследован веб-граф $\mathbf{G_0}$, вершинам которого соответствуют сайты научных организаций СО РАН, а отношение между сайтами определяется наличием ссылок друг на друга. Мы полагаем, что 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. Веб-граф $\mathbf{G_0}$ является ориентированным графом, любая пара его вершин может быть соединена одной дугой или двумя противоположно направленными дугами. Вершинам графа $\mathbf{G_0}$ соответствуют научные организации из информационной системы «Организации и сотрудники СО РАН» [8] по состоянию на 10 августа 2012 г. [9]. Число таких организаций равно 88, число ссылок сайтов этих организаций друг на друга, а значит, и число ребер графа $\mathbf{G_0}$ равно 863 [2].

В настоящей работе рассмотрен веб-граф $\mathbf{G}$ информационного пространства СО РАН по состоянию на 2 ноября 2012 года [10]. Он имеет отличия от графа $\mathbf{G_0}$, обусловленные включением в информационное пространство новых сайтов, возникновением новых связей и исчезновением ранее существовавших связей. Граф $\mathbf{G}$ имеет 106 вершин и 1084 дуги и включает как вершины графа $\mathbf{G_0}$, так и новые вершины. В этом параграфе мы сравниваем структурные свойства веб-графов $\mathbf{G_0}$ и $\mathbf{G}$. Диаграммы этих графов представлены в [9] и [10].

Для оценки степени участия вершин и дуг в формировании структуры графа используются следующие численные параметры — инварианты графа.

Индекс вершин. Пусть ориентированный граф G имеет n вершин, и k из них имеют хотя бы одну исходящую или входящую дугу. *Индексом вершин* в графе G будем называть величину $c_v(G) = \dfrac{k}{n}$. Для веб-графа этот параметр характеризует число сайтов, включенных в информационное взаимодействие. Близость величины $c_v(G)$ к нулю говорит о том, что имеется значительное количество изолированных сайтов, то есть таких, которые не связаны с другими сайтами. Равенство $c_v(G) = 1$ означает, что все сайты научных организаций вовлечены во взаимодействие друг с другом.

Индекс дуг. Пусть ориентированный граф G имеет n вершин и t дуг. *Индексом дуг* в графе G будем называть величину $c_a(G) = \dfrac{t}{n(n-1)}$. В [11] эта величина называется *плотностью сети*. Максимальное значение $c_a(G) = 1$ достигается в случае, когда любые две вершины графа G соединены парой противоположно ориентированных дуг. В этом случае все сайты ссылаются друг на друга.

Коэффициент кластеризации. Пусть G – ориентированный граф, а V_2 – множество таких его вершин, для каждой из которых сумма чисел входящих и исходящих дуг не менее 2. Под окрестностью вершины v будем понимать множество вершин графа, соединенных с v дугами без учета их ориентации. Для вершины v графа G обозначим через G_v подграф, порожденный окрестностью вершины v. *Коэффициентом кластеризации вершины* v будем называть величину $c_a(G_v)$, то есть индекс дуг подграфа, порожденного окрестностью вершины [12]. *Коэффициентом кластеризации графа* G будем называть величину $cc(G) = \dfrac{1}{|V_2|} \sum\limits_{v \in V_2} c_a(G_v)$. Таким образом, $cc(G)$ показывает как в среднем заполнена дугами окрестность вершины графа.

Значения указанных параметров для графов $\mathbf{G}$ и $\mathbf{G_0}$ приведены в таблице:

$c_v(\mathbf{G_0}) = 1$	$c_a(\mathbf{G_0}) = 0,11$	$cc(\mathbf{G_0}) = 0,06$
$c_v(\mathbf{G}) = 0,98$	$c_a(\mathbf{G}) = 0,10$	$cc(\mathbf{G}) = 0,07$

Отметим, что граф $\mathbf{G}$ содержит две изолированные вершины.

2.1. Классификация типов вершин

Под расстоянием между парой вершин в ориентированном графе понимается число дуг в кратчайшем ориентированном пути, соединяющим эти вершины. Естественными характеристиками вершины v ориентированного графа являются число исходящих из нее дуг (полустепень исхода) и число входящих в нее дуг (полустепень захода). Увеличение полустепеней вершин графа влечет в общем случае возрастание его компактности. Под этим понимается уменьшение расстояний между вершинами и, как следствие, уменьшение диаметра графа (максимального расстояния между его вершинами). Исходящие и входящие дуги вместе с вершиной образуют легко распознаваемые локальные фрагменты, которые могут быть использованы в качестве классификационных признаков вершин. Минимальная и максимальная степени исхода вершин в графе $\mathbf{G}$ равны 0 и 99, а захода — 0 и 57 (в графе $\mathbf{G_0}$ эти значения равны 0, 87 и 1, 48). Средние полустепени исхода/захода в графе $\mathbf{G}$ равны 10,3 (в графе $\mathbf{G_0}$ — 9,8). Число вершин, из которых нет ни одной исходящей дуги составляет около 24% от всех вершин графа $\mathbf{G}$ (для графа $\mathbf{G_0}$ это значение равно 19%). В $\mathbf{G}$ большое число исходящих дуг имеют пять вершин, соответствующие сайтам: Портал СО РАН (99 дуг), Объединенный ученый совет СО РАН по нанотехнологиям и информационным технологиям (91), ИВТ СО РАН (91), Отделение ГПНТБ СО РАН (89) и Президиум СО РАН (79).

При анализе веб-графа представляет интерес соотношение между полустепенями исхода и захода вершин. На рис. 5 приводятся три варианта возможного распределения входящих и исходящих дуг. Вершины первого типа называют *индукторами* (мало входящих дуг, много исходящих), второго типа — *коллекторами* (много входящих дуг, мало исходящих), а третьего типа — *посредниками* (много входящих и исходящих дуг). Эти типы вершин образуют множество *веб-коммуникаторов* графа.

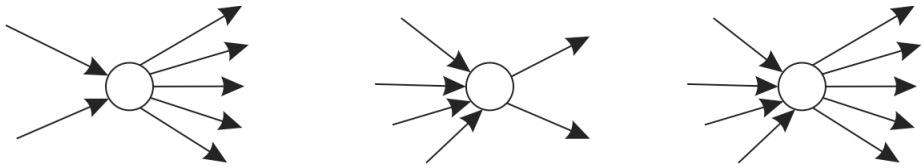

Рис. 5. Веб-коммуникаторы: индуктор, коллектор и посредник

Коллекторы могут соответствовать сайтам организаций, в которых происходит накопление, хранение и обработка данных. Это могут быть библиотеки, банки данных, центры коллективного пользования, справочные ресурсы, журналы. Посредниками могут быть вершины, соответствующие головным сайтам, порталам научных центров, сайтам институтов с высокой степенью научной кооперации. Индукторами могут являться сайты недавно созданных организаций или новые сайты уже существующих институтов. Анализ вершин с большими степенями показывает, что в веб-графе $\mathbf{G}$ индукторами можно назвать сайты Объединенного ученого совета СО РАН по биологическим наукам (45 исходящих ребер, 2 входящих ребра), ОУС СО РАН по НИТ (91, 21) и ИВТ

СО РАН (91, 23), а посредниками – сайты Портал СО РАН (99, 57), Президиум СО РАН (79, 40), ИК СО РАН (30, 15), Отделение ГПНТБ СО РАН (89, 30) и ГПНТБ СО РАН (46, 33). Коллекторов с большой степенью захода в этом графе нет. Посредников в графе довольно много, например, это сайты СОРАН.INFO (18, 25), ИМ СО РАН (28, 22), СФУ (12, 13), ИХБФМ СО РАН (16, 14).

Отнесение вершин графа к веб-коммуникаторам того или иного типа зависит от соотношения между полустепенями захода и исхода, которое можно задавать на основе распределения степеней вершин в графе. В [2] приводятся таблицы с количеством веб-коммуникаторов в графе $\mathbf{G_0}$ при разных отношениях между полустепенями.

2.2. Сильно связная компонента

Для описания структуры веб-графов, особенно больших, используется представление в виде схемы галстука-бабочки [13]. В этой модели в графе выделяется максимальная сильно связная компонента, по отношению к которой классифицируются остальные вершины графа. Подграф называется *сильно связной компонентой* графа, если между любой парой его вершин существует ориентированный путь. Таким образом, проходя по ссылкам соответствующих сайтов, можно обойти всю компоненту. Центральную часть бабочки образует максимальная сильно связная компонента. Левая часть бабочки состоит из вершин, пути из которых ведут в эту компоненту. Правую часть образуют вершины, в которые ведут пути из компоненты (см. рис. 6). В сложных веб-графах есть подмножества вершин, не попадающих в эти части бабочки. На рис. 6 приводятся схемы бабочек для веб-графов $\mathbf{G_0}$ и $\mathbf{G}$, где числа указывают количества вершин в частях бабочки. Максимальное расстояние между вершинами графа $\mathbf{G}$ равно 3, тогда

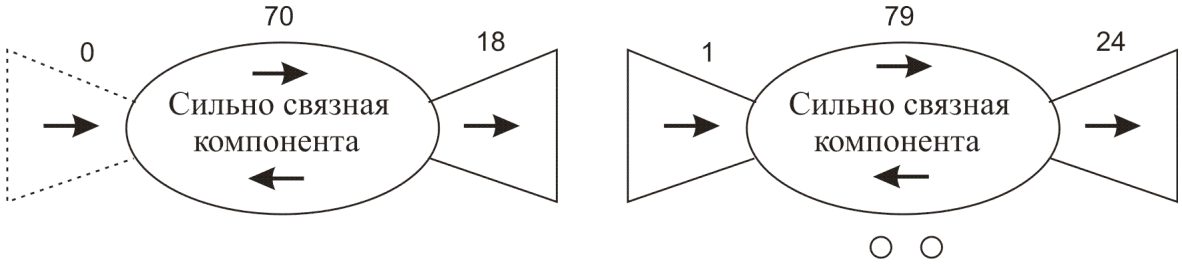

Рис. 6. Представление структуры графов $\mathbf{G_0}$ (слева) и $\mathbf{G}$ (справа).

как в графе $\mathbf{G_0}$ оно равно 4. Малый диаметр графа обеспечивается вершиной Портал СО РАН, имеющей большое число исходящих и входящих дуг.

3. Анализ веб-подграфов

Статическая структура веб-графа научных организаций СО РАН, зафиксированная в какой-то момент времени, отражает текущие информационные связи между институтами. Представляется интересным исследовать веб-подграфы, соответствующие институтам по отдельным наукам за разные годы. Далее будут представлены данные о веб-графах научных организаций по гуманитарным и химико-биологическим наукам за 2010 и 2012 годы. На диаграммах графов две противоположно направленные дуги между парой вершин изображены одной двунаправленной дугой.

3.1. Веб-подграф институтов по гуманитарным наукам

На рис. 7 приводится структура веб-подграфов $\mathbf{H}_{10}$ и $\mathbf{H}_{12}$ сайтов организаций по гуманитарным наукам за 2010 и 2012 годы (включая один экономический институт). Вершины графа $\mathbf{H}_{12}$, отсутствующие в $\mathbf{H}_{10}$, выделены серым цветом.

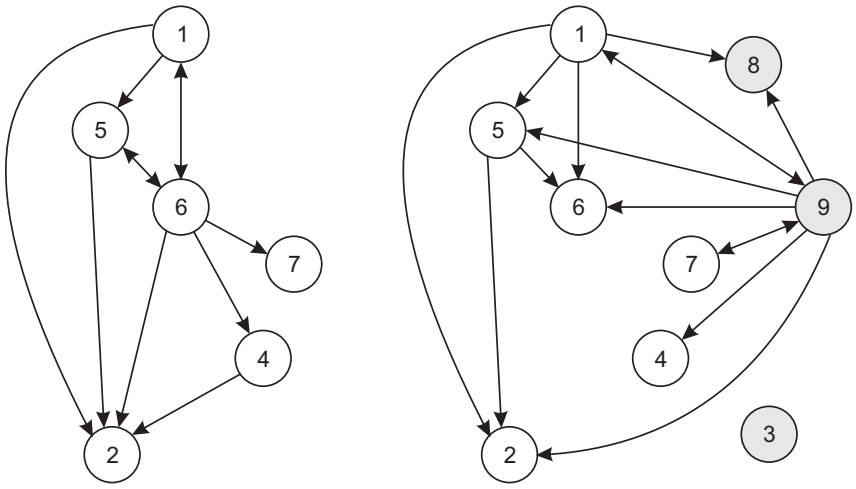

Рис. 7. Веб-графы $\mathbf{H}_{10}$ (слева) и $\mathbf{H}_{12}$ (справа) сайтов институтов СО РАН, относящихся к гуманитарным наукам. 1 — ГПНТБ СО РАН, 2 — ИФПР, 3 — ИПОС, 4 — ИЭОПП, 5 — ИАЭТ, 6 — ИИ, 7 — ИФЛ, 8 — ИМБТ, 9 — Отделение ГПНТБ СО РАН

Сильно связная компонента в обоих графах состоит из трех вершин — в графе $\mathbf{H}_{10}$ это $\{1, 5, 6\}$, а в графе $\mathbf{H}_{12}$ — $\{1, 7, 9\}$. Все остальные вершины графов, кроме одной изолированной, образуют правую часть бабочки. Средняя степень вершин в графах $\mathbf{H}_{10}$ и $\mathbf{H}_{12}$ равна 1,83 и 1,67 соответственно, а диаметр обоих графов равен 2. Параметры графов, характеризующие их структуру: $c_v(\mathbf{H}_{10}) = 1$, $c_a(\mathbf{H}_{10}) = 0,37$, $cc(\mathbf{H}_{10}) = 0,1$ и $c_v(\mathbf{H}_{12}) = 0,89$, $c_a(\mathbf{H}_{12}) = 0,21$, $cc(\mathbf{H}_{12}) = 0,04$. Вершина 6 является индуктором в графе $\mathbf{H}_{10}$, а вершины 1 и 9 — индукторами в графе $\mathbf{H}_{12}$.

3.2. Веб-подграф институтов по химико-биологическим наукам

На рис. 8 приводится структура веб-подграфов $\mathbf{B}_{10}$ и $\mathbf{B}_{12}$ сайтов организаций по химико-биологическим наукам за 2010 и 2012 годы. Вершины графа $\mathbf{B}_{12}$, отсутствующие в $\mathbf{B}_{10}$, выделены серым цветом.

Максимальная сильно связная компонента в графе $\mathbf{B}_{10}$ содержит 6 вершин: $\{1, 2, 4, 7, 10, 13\}$, а в графе $\mathbf{B}_{12}$ — 9 вершин: $\{1, 2, 3, 4, 6, 7, 9, 10, 13\}$. Левая часть бабочки содержит в обоих графах одну вершину, а правая — 2 и 7 вершин соответственно. Средняя степень вершин в графах $\mathbf{B}_{10}$ и $\mathbf{B}_{12}$ равна 2 и 1,56 соответственно, а диаметр обоих графов равен 4. Структура графов характеризуется следующими параметрами: $c_v(\mathbf{B}_{10}) = 0,82$, $c_a(\mathbf{B}_{10}) = 0,2$, $cc(\mathbf{B}_{10}) = 0,09$ и $c_v(\mathbf{B}_{12}) = 0,68$, $c_a(\mathbf{B}_{12}) = 0,07$, $cc(\mathbf{B}_{12}) = 0,09$. В графе $\mathbf{B}_{10}$ большинство вершин являются посредниками, а графе $\mathbf{B}_{12}$ вершина 6 будет индуктором, а вершина 4 — коллектором.

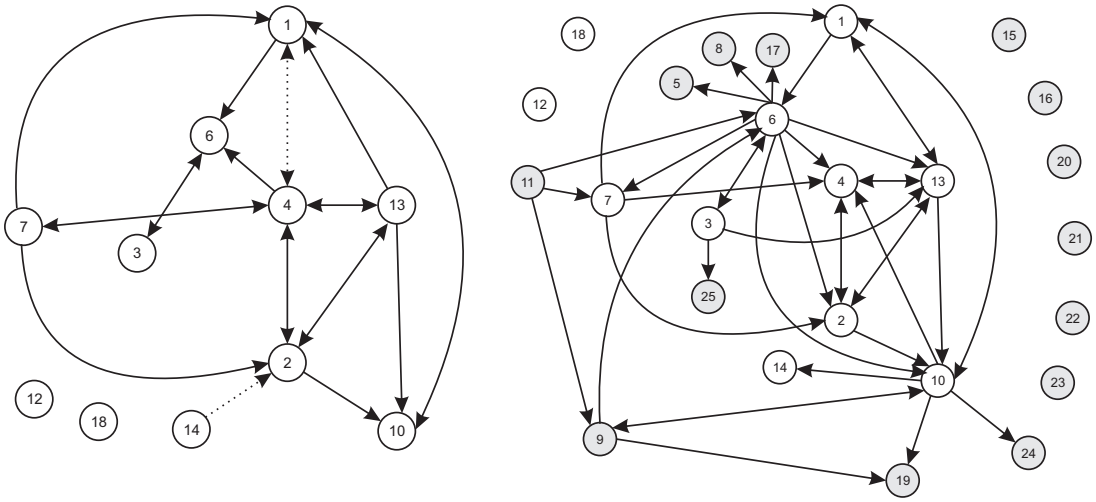

Рис. 8. Веб-графы $\mathbf{B}_{10}$ (слева) и $\mathbf{B}_{12}$ (справа) сайтов химико-билогических институтов СО РАН. 1 — ИЦиГ, 2 — НИОХ, 3 — ИНХ, 4 — ИХКГ, 5 — ИЛ, 6 — ИК СО РАН, 7 — ИХТТМ, 8 — ИХН, 9 — ИБФ, 10 — ИХБФМ, 11 — ИХХТ, 12 — ЦСБС, 13 — МТЦ, 14 — ИСиЭЖ, 15 — ИрИХ, 16 — СИФИБР, 17 — ИПХЭТ, 18 — ИПА, 19 — ИОЭБ, 20 — ИППУ, 21 — ИБПК, 22 — АФ ЦСБС, 23 — ЗСФ ИЛ, 24 — ИМКБ, 25 — ИУХМ

4. Выводы

Методами вебометрики и теории графов проведен анализ сайтов и структуры веб-графа научных организаций Сибирского региона. Представленные данные показывают современное состояние информационной структуры взаимодействия научных организаций на уровне сайтов и позволяют проследить эволюцию рассматриваемого веб-пространства. Изменения происходят непрерывно, без скачков. Замечена тенденция увеличения количества связей между сайтами, которые входят в сильно связную компоненту.

Список литературы

[1] Рейтинг сайтов научных организаций СО РАН. http://www.ict.nsc.ru/ranking/ (дата доступа — 15.04.2013).

[2] Шокин Ю.И., Веснин А.Ю., Добрынин А.А., Клименко О.А., Рычкова Е.В., Петров И.С. Исследование научного веб-пространства Сибирского отделения Российской академии наук // Вычисл. технологии. 2012. Т. 17, № 6. С. 86–98.

[3] Поисковая система Яндекс. http://www.yandex.ru/ (дата доступа — 15.04.2013).

[4] Поисковая система Google. http://www.google.ru/ (дата доступа — 15.04.2013).

[5] Поисковая система Bing. http://www.bing.com/ (дата доступа — 15.04.2013).

[6] Система определения индекса цитирования в веб-пространстве Google Scholar. http://scholar.google.com/ (дата доступа — 15.04.2013).

[7] Индекс цитирования каталога Яндекс. http://help.yandex.ru/catalogue/?id=873431 (дата доступа — 15.04.2013).

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

[9] Веб-граф $\mathbf{G_0}$ организаций СО РАН.
`http://www.ict.nsc.ru/sitepage.php?PageID=975` (дата доступа — 10.08.2012).

[10] Веб-граф $\mathbf{G}$ организаций СО РАН.
`http://www.ict.nsc.ru/sitepage.php?PageID=976` (дата доступа — 02.11.2012).

[11] Hage P., Harary F. Structural models in anthropology. Cambridge University Press: Cambridge, UK, 1983.

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

[13] Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the Web // Computer Networks. 2000. Vol. 33. № 1–6. P. 309–320.