О ВЗАИМОСВЯЗЯХ РОССИЙСКОГО СУПЕРКОМПЬЮТЕРНОГО СООБЩЕСТВА В ВЕБЕ

А.А. Печников

На первой странице веб-ресурса 14-й конференции "Научный сервис в сети Интернет" (http://agora.guru.ru/display.php?conf=abrau2012), говорится от том, что «...на новый уровень выходит взаимодействие суперкомпьютерного сообщества и государственных структур». В данной работе рассматривается взаимодействие российского суперкомпьютерного сообщества с точки зрения взаимосвязей соответствующих веб-сайтов в Вебе.

О некоторых определениях и методах исследования

В работе используются методы, разработанные и примененные ранее для исследований различных фрагментов российского Веба и описанные в общем виде, например, в [1]. Для получения, хранения и обработки вебометрической информации использовалась бета-версия программного комплекса BeeBot [2].

Определим несколько понятий, которые потребуются для дальнейшего изложения.

Определение 1. Уникальной внешней гиперссылкой называется гиперссылка из множества всех гиперссылок с одинаковым адресом и контекстом, которая находится на странице, имеющий максимальный уровень; при этом уровень начальной страницы сайта считается наивысшим. Далее мы будем рассматривать только уникальные внешние гиперссылки, поэтому слова «уникальная» и «внешняя» в большинстве случаев будем опускать.

Определение 2. Целевым множеством называется множество исследуемых сайтов, идентифицируемых уникальными доменными именами, задаваемое прямым перечислением. Здесь следует сделать следующее примечание. Когда мы говорим об элементе целевого множества, то имеем в виду доменное имя, идентифицирующее сайт. Поэтому слова о том, что сайт принадлежит целевому множеству, подразумевают, что ему принадлежит доменное имя, идентифицирующее сайт.

Определение 3. Сопутствующим множеством (по отношению к заданному целевому множеству) называется множество сайтов, не входящих в целевое множество, на которые существуют гиперссылки с сайтов целевого множества.

Несколько слов о процедуре исследования. Вначале определяется целевое множество сайтов, сайты которого сканируются краулером BeeBot’a с целью формирования базы данных внешних гиперссылок. Сформированная база данных позволяет построить сопутствующее множество, в котором можно выделить наиболее интересные для данного исследования сайты (в первую очередь – по количеству ссылок на них с сайтов целевого множества, но также и по другим признакам, например, типу сайта). Сканирование «наиболее интересных» сайтов позволяет дополнить базу данных и построить различные веб-графы фрагмента Веба, у которых вершинами являются сайты расширенного целевого множества, а дугами – гиперссылки между ними. Содержательная интерпретация результатов анализа базы данных гиперссылок и построенных веб-графов позволяет дать рекомендации по улучшению взаимодействия веб-сайтов.

Целевое множество

Для формирования целевого множества использовался веб-сайт Информационно-аналитический центра по параллельным вычислениям Лаборатории параллельных информационных технологий НИВЦ МГУ (www.parallel.ru), который, естественно, первым и был включён в целевое множество. На странице http://www.parallel.ru/russia/organizations.html «...представлены российские организации, использующие или разрабатывающие параллельные (высокопроизводительные) информационные технологии». Здесь указано 44 организации, из которых в целевое множество были включены сайты 38 (например, организаций и/или сайтов не был обнаружен). Сайт Центра высокопроизводительной обработки данных КарНЦ РАН (cluster.krc.karelia.ru), отсутствующий в этом списке, был добавлен автором.

39 сайтов целевого множества с некоторой степенью условности можно разделить на 4 группы (полный список можно найти по ссылке http://webometrics.krc.karelia.ru/doc/TS_supercomp_4_march_12.xls):

(1) 15 сайтов организаций, непосредственно занимающихся параллельными и суперкомпьютерными технологиями (www.parallel.ru, cluster.krc.karelia.ru, lvk.cs.msu.su – Лаборатория вычислительных комплексов ВМиК МГУ, www.csa.ru – Суперкомпьютерный альянс и др.);

(2) 14 сайтов научных учреждений РАН (www.ccas.ru – Вычислительный центр РАН, www.ispras.ru – Институт системного программирования РАН, www.inm.ras.ru – Институт вычислительной математики РАН, и др.);

(3) 5 сайтов фирм и организаций-разработчиков (www.module.ru – Научно-технический Центр "Модуль", www.t-platforms.ru – "Т-Платформы", new.tsniimash.ru – ФГУП Центральный научно-исследовательский институт машиностроения, и др.);

(4) 5 «неклассифицированных» сайтов (www.srcc.msu.ru – Научно-исследовательский вычислительный центр МГУ, uginfo.rsu.ru – Научно-исследовательский институт многопроцессорных вычислительных систем ЮФУ, и др.

Целевое множество является разнородным, что характеризуется, например, количеством html-страниц на них, колеблющимся от десятков до десятков тысяч. Общее количество обнаруженных внешних гиперссылок, сделанных со всех сайтов целевого множества, равняется 24660, при этом сделаны они примерно на 7440 различных сайтов (сюда включены и 39 сайтов целевого множества). Почти половина внешних ссылок (11900 из 24660) сделана с сайта www.parallel.ru на 2440 сайтов. Более 3300 ссылок на 2175 сайтов сделано с сайта Суперкомпьютерного альянса (www.csa.ru) и 1350 ссылок на 1125 сайтов сделано с сайта Вычислительного центра РАН (www.ccas.ru).

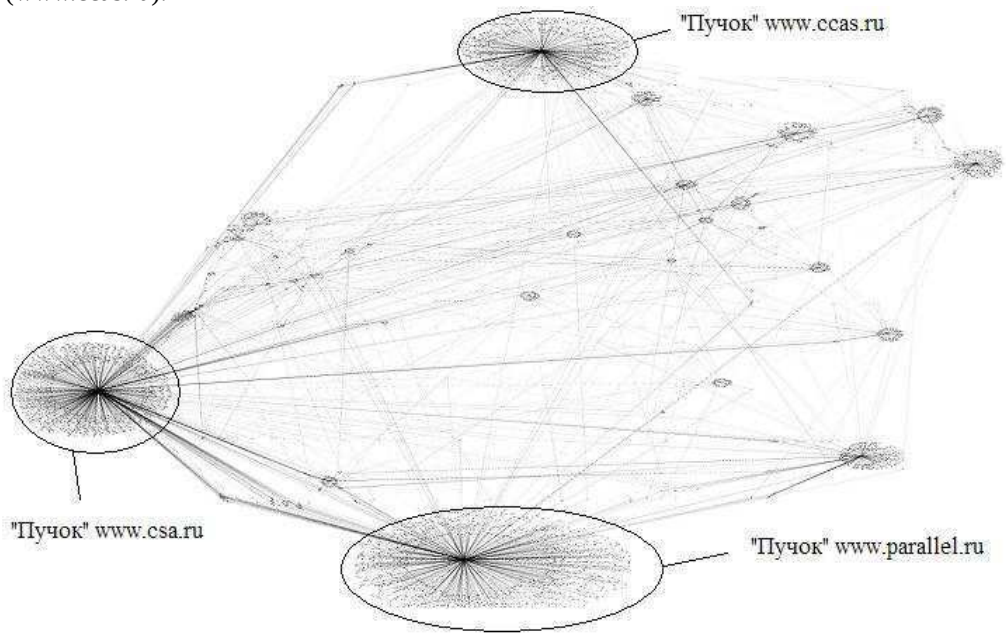

Рис.1. Веб-граф, построенный на гиперссылках, исходящих с целевого множества

На рис. 1 представлен веб-граф, построенный на 7440 вершинах целевого и сопутствующего множеств. Здесь нарисованы только исходящие дуги с вершин, соответствующих сайтам целевого множества (дуга изображена, если существует хотя бы одна гиперссылка, связывающая два соответствующих сайта). Если на рис. 1 оставить только дуги, связывающие сайты целевого множества, то получим веб-граф, приведённый на рис. 2. Четыре сайта целевого множества на рисунке отсутствуют, поскольку их вершины в веб-графе изолированные.

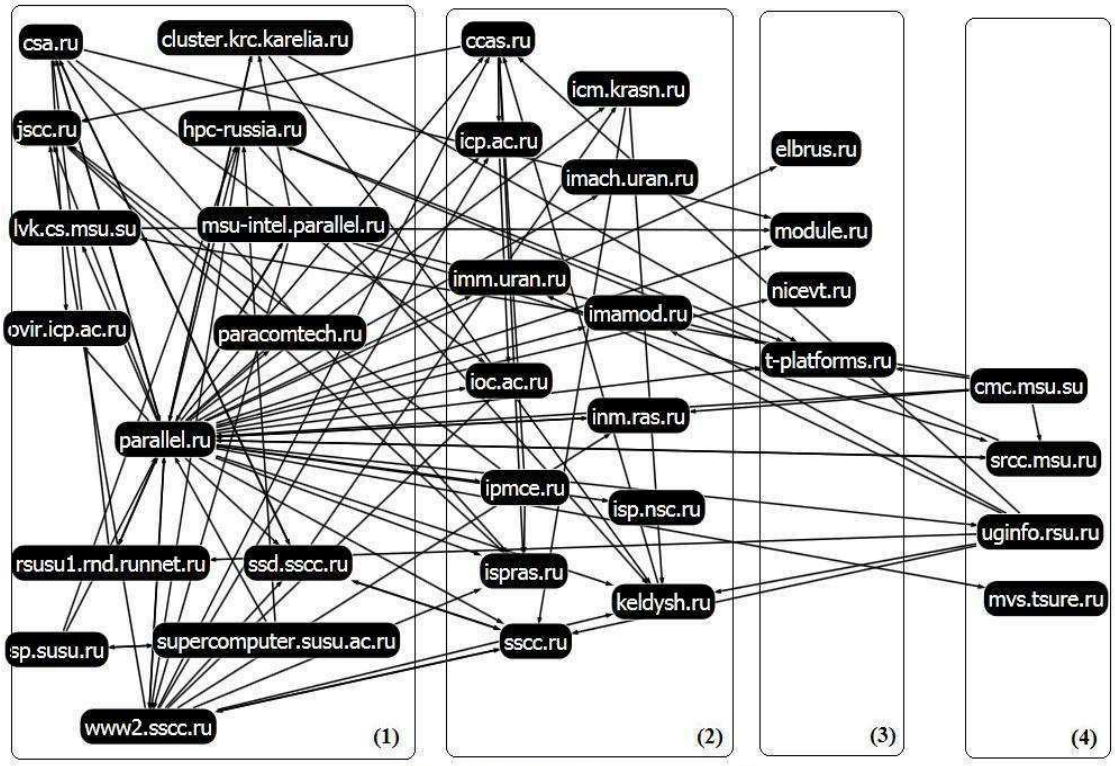

Рис.2. Веб-граф целевого множества.

Первая группа в веб-графе представлена 14 сайтами, вторая – 13-ю, а третья и четвертая – четырьмя сайтами. Веб-граф имеет 97 дуг, но количество гиперссылок между сайтами значительно больше – 329 (например, с сайта parallel.ru на сайт t-platform.ru сделано 59 гиперссылок).

Достаточно ожидаема важная роль сайта parallel.ru, имеющего 47 исходящих и 184 входящих гиперссылок. За ним с существенным отставанием следует сайт sca.ru: 26 исходящих и 23 входящие гиперссылки. На рис. 3 показано количество исходящих и входящих гиперссылок для всех сайтов, изображенных на рис. 2.

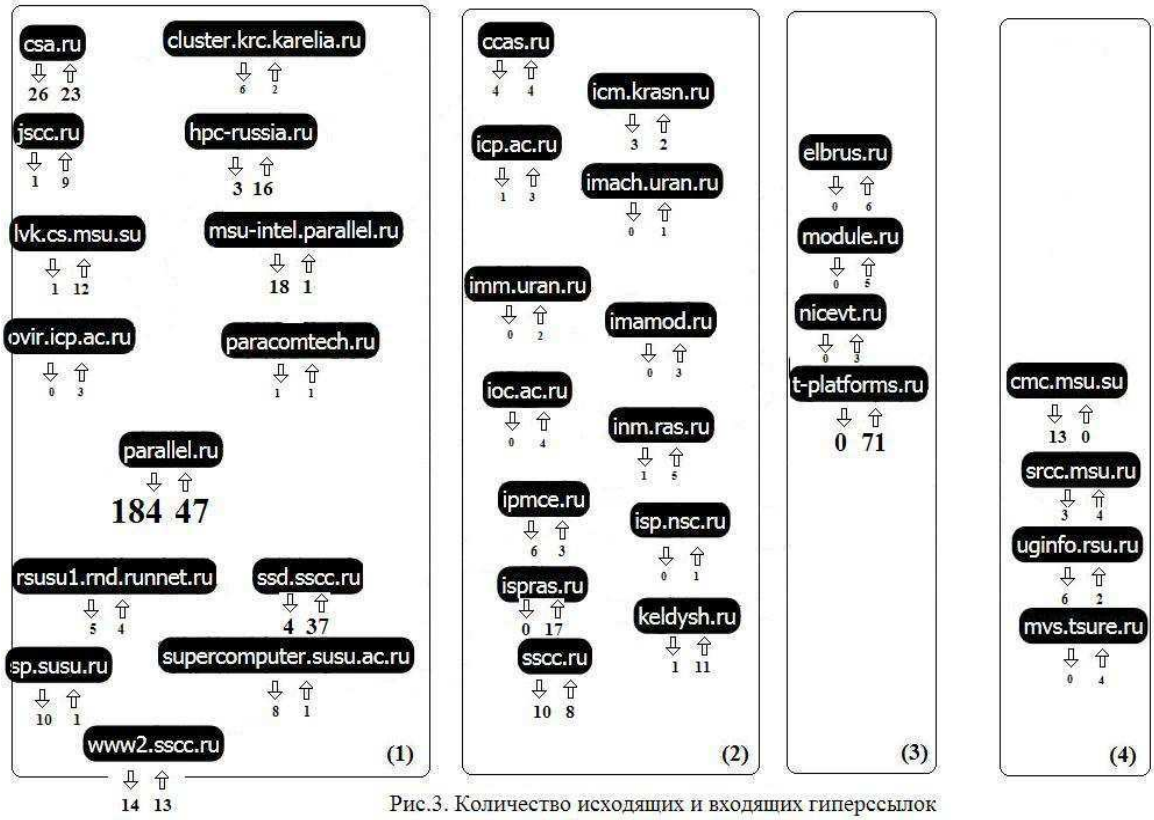

Рис.3. Количество исходящих и входящих гиперссылок

Оценка сайтов целевого множества, основанная на вычислении их «важности» в сообществе по аналогии с PageRank [3], также подтверждает безусловное лидерство сайта parallel.ru: PR сайта parallel.ru почти в два раза больше среднего значения PR по всем остальным сайтам и на треть больше, чем PR следующего по «важности» сайта. Сайт Института системного программирования РАН (www.ispras.ru), занимающий второе место по значению PR, не так просто выявить визуально на рисунке 2, а тем более по количеству исходящих (0) и входящих (7) гиперссылок. Но анализ первого приближения PR подтверждает, что на www.ispras.ru ссылаются 7 сайтов, проявляющих набольшую гиперссылочную активность в сообществе (им инцидентна половина всех гиперссылок целевого множества), существенно повышая его «важность».

Удаление из веб-графа целевого множества всех вершин, имеющих только входящие или только исходящие дуги приводит к «очищенному» веб-графу, содержащему лишь 15 вершин, но зато сайты, соответствующие этим вершинам, в полном смысле можно назвать участниками сообщества: все они имеют прямые и обратные связи с другими участниками.

Сопутствующее множество

Сопутствующее множество содержит около 7400 сайтов, причем только треть имеют имена в доменной зоне .ru, что свидетельствует о нашей международной активности. На первом месте по количеству гиперссылок, сделанных с сайтов целевого множества, находится сайт Top500 (www.top500.org) – 5300 ссылок, сделанные с 6 сайтов целевого множества. В свою очередь, с сайта Top500 сделано одной гиперссылке на srcc.msu.ru и www.jscc.ru (Межведомственный суперкомпьютерный центр РАН).

Из российских веб-сайтов сопутствующего множества, безусловно, выделяются два сайта: agora.guru.ru и sok.susu.ru. На первом из них размещаются разделы, посвященные конференциям «Научный сервис в сети Интернет». На agora.guru.ru сделано 82 гиперссылки с 13 сайтов целевого множества. С сайта agora.guru.ru сделана только одна ответная ссылка. Сайт sok.susu.ru является персональным сайтом проф. Л.Б. Соколинского, здесь мы имеем входящие ссылки с 4 сайтов целевого множества и встречные ссылки на 3 из них.

И, наконец, по поводу взаимодействия с сайтами государственных структур: имеются гиперссылки с сайтов целевого множества на сайты ВАК, Минобрнауки РФ и ряд других. Обратных гиперссылок практически не найдено.

Заключение

Несколько слов о содержательной интерпретации результатов. Группа сайтов организаций, непосредственно занимающихся параллельными и суперкомпьютерными технологиями, проявляет высокую активность в смысле выставления гиперссылок на своих коллег из обозначенного ранее целевого множества. Даже несмотря на то, что 184 таких гиперссылки сделано с сайта parallel.ru, еще 97 приходятся на остальные 14 сайтов этой группы. Не находит отражения в Вебе участие в суперкомпьютерном сообществе фирм и организаций-разработчиков: они вообще не имеют гиперссылок на своих коллег. Большинство сайтов научных учреждений РАН в этом смысле очень похожи на сайты фирм, хотя есть и приятные исключения. Зато три сайта из четвертой группы (Научно-исследовательский вычислительный центр МГУ, Факультет вычислительной математики и кибернетики МГУ и Южно-Российский региональный центр информатизации высшей школы) являются полноправными участниками сообщества.

Отсюда следует и первая рекомендация: если организация позиционирует себя как участника суперкомпьютерного сообщества, то реальное взаимодействие со своими коллегами должно находить отражение в Вебе за счёт выставления соответствующих гиперссылок.

Ещё одна рекомендация связана с разделом конференций «Научный сервис в сети Интернет» на сайте agora.guru.ru. Возможности этого ресурса как средства коммуникации в Вебе между участниками суперкомпьютерного сообщества использованы далеко не полностью. Например, связность веб-сообщества существенно увеличится, если на сайте agora.guru.ru сделать «живые» гиперссылки на организации, представители которых входят в организационный и программный комитеты.

Работа выполнена при частичной поддержке в рамках конкурса Минобразования России на Формирование государственных заданий вузам в части проведения научно-исследовательских работ (Петрозаводский государственный университет, НИР № 632-12) и гранта РГНФ (проект №12-03-12001).

ЛИТЕРАТУРА:

1. Печников А.А. Методы исследования регламентируемых тематических фрагментов Web // Труды Института системного анализа Российской академии наук. Серия: Прикладные проблемы управления макросистемами. Том 59. 2010. – С. 134-145.
2. Печников А.А., Чернобровкин Д.И. Адаптивный краулер для поиска и сбора внешних гиперссылок // Управление большими системами. Выпуск 36. М.: ИПУ РАН, 2012. С.301-315.
3. Brin S., Page L. The anatomy of a large scale hypertextual web search engine // Computer Networks and ISDN Systems. 1998. Т. 30. № 1-7. С. 107-117.