Являются ли сайты конференций RCDL научными веб-коммуникаторами?♣

© А.А. Печников, Н.Б. Луговая

Институт прикладных математических исследований КарНЦ РАН
{pechnikov, nataly}@krc.karelia.ru

Аннотация

В авторской терминологии модель, называемая схемой научного Веба, конструируется из четырех основных компонент, являющихся непересекающимися подмножествами научных сайтов. Эти компоненты называются административным каркасом, научным подмножеством, ближайшей окрестностью и веб-коммуникатором. При всей кажущейся очевидности того факта, что сайты научных конференций являются средством коммуникации ученых, проведенные исследования показывают, что сайты конференций RCDL веб-коммуникаторами не являются.

1 Схема научного Веба

К актуальным направлениям вебометрики [1], - одного из развивающихся направлений информатики, - относятся исследования гиперссылок (аналогичные термины – «ссылка», «веб-ссылка»), являющиеся единственным способом взаимодействия между сайтами. Практическая применимость этих исследований успешно демонстрируется реализацией алгоритмов информационного поиска таких популярных систем, как Google и Яндекс [2,4]. Теоретические исследования показывают, что изучение гиперссылок имеет достаточный потенциал как в смысле новых источников информации и коммуникации, так и ценности самих веб-страниц [3-6].

В 2008 году в Институте прикладных математических исследований КарНЦ РАН началась работа по проекту «Вебометрические исследования научных интернет-ресурсов российского Интернета». В рамках проекта разработаны поисковый робот для сбора исходящих с сайтов гиперссылок (LPR – аббревиатура от Link, Page и Robot) и база данных, предназначенная для их хранения и обработки (БД ВГ – База Данных Внешних Гиперссылок). Программный комплекс, состоящий из LPR и БД ВГ, создан на языке PHP, работает под управлением веб-сервера Apache с интегрированным модулем PHP и СУБД MySQL и находится в стадии постоянного совершенствования.

Информацию о ходе проекта можно найти на сайте «Вебометрика. ИПМИ КарНЦ РАН» [7]. По данным на март 2009 года (именно на этих данных основан дальнейший материал), проведено сканирование 280 официальных сайтов организаций и учреждений. Российской академии наук (РАН). Обработано более миллиона html-страниц, найдено и сохранено 660000 различных внешних ссылок, из которых 81000 уникальных.

В результате проведенных исследований построена теоретико-графовая модель множества научных сайтов Рунета, получившая название схемы научного Веба. Схема научного Веба представляет собой ориентированный граф, множество вершин которого соответствует исследуемым сайтам, а дуги отражают гиперссылки, существующие между сайтами (считается, что дуга существует тогда и только когда, существует хотя бы одна гиперссылка с одного сайта на другой).

Показано, что в схеме научного Веба можно выделить четыре компоненты.

Первая из них, – административный каркас, – отражает ссылки между сайтами, соответствующие иерархической подчиненности организаций.

Вторая, – множество научных подмножеств, где научное подмножество представляет связи между сайтами родственных организаций.

Третья компонента - это множество ближайших окрестностей официальных сайтов. Ближайшие окрестности содержат вершины, сайты которых имеют имена ddd.nnn.ss∈ nnn.ss, где nnn.ss – доменное имя официального сайта.

И, наконец, четвертая компонента называется множеством научных веб-коммуникаторов и соответствует множеству сайтов, выполняющих коммуникационные функции между официальными научными сайтами, то есть научные сайты имеют много входящих ссылок с сайтов этого множества и много исходящих ссылок на них.

Веб-коммуникаторы, в свою очередь, можно классифицировать по трем типам как «посредник», «индуктор» и «коммутатор». Краткое описание

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

посредника - «много входящих ссылок, много исходящих ссылок», коммутатора – «мало входящих, много исходящих», а индуктора – «много входящих, мало исходящих».

Упрощенный вариант схемы научного Веба изображен на рис. 1.

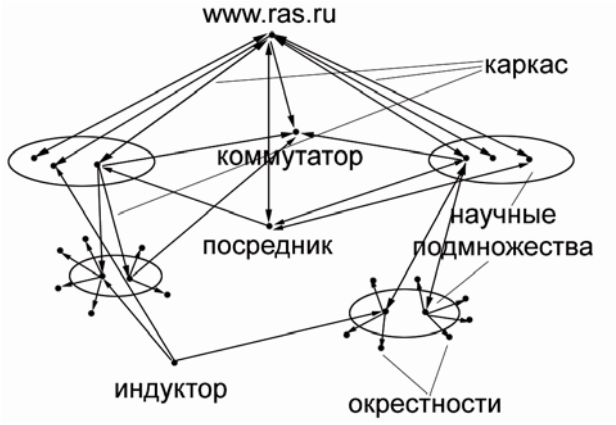

Рис.1 Схема научного Веба.

2 Цель и целевое множество исследования

Основной вопрос данной публикации заключается в следующем: являются ли сайты научных конференций научными веб-коммуникаторами?

В качестве объекта исследования были взяты хорошо знакомые авторам конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (английская аббревиатура – RCDL).

Исследования проводились на множестве сайтов конференций RCDL и научных организаций РАН, являющихся организаторами этих конференций, а также сайтов поддерживающих организаций. Сведения о веб-ресурсах всех 11 конференций приведены на сайте [10], являющемся общим сайтом RCDL и также включенным в число исследуемых сайтов. С помощью LPR было проведено сканирование 10 сайтов конференций из 12. Исключениями являются ресурсы «RCDL 2000: Протвино» и «RCDL 2003: Санкт-Петербург», поскольку они представлены не отдельными сайтами, а директориями на сайтах других организаций (LPR «не умеет» сканировать с произвольной страницы сайта).

Список всех учреждений и организаций РАН, представители которых хотя бы один раз входили в состав организационного или программного комитета, содержит 22 наименования учреждений РАН. По ряду причин (отсутствие сайта, отсутствие на сегодняшний день самой организации и др.) нам пришлось ограничиться 16 сайтами. Список составлен на основе информации, представленной на сайтах конференций и приведен ниже (рис. 2).

Единственной организацией, не входящей в состав РАН, регулярно поддерживающей конференции, является Российский фонд фундаментальных исследований, сайт которого также включен в исследуемое множество.

3 Результаты исследования

Операция БД ВГ, которая называется ПОСТРОЕНИЕ МАТРИЦЫ, позволяет построить матрицу смежности для 16 сайтов учреждений и организаций РАН; матрицы приведена на рис. 2. Элемент матрицы (i,j)=1, если существует хотя бы одна гиперссылка с сайта i на сайт j , и (i,j)=0 в ином случае.

Соответствие номеров вершин матрицы и организаций:

0 Библиотека по естественным наукам РАН
1 Дальневосточное отделение РАН
2 Карельский научный центр
3 Вычислительный центр имени А.А. Дородницына РАН
4 Институт астрономии РАН
5 Институт вычислительных технологий СО РАН
6 Институт космических исследований РАН
7 Институт математических проблем биологии РАН
8 Институт прикладной математики им. М. В. Келдыша РАН
9 Институт прикладных математических исследований КарНЦ РАН
10 Институт проблем информатики РАН
11 Институт систем информатики им. А.П. Ершова СО РАН
12 Институт химической физики им. Н.Н. Семенова РАН
13 Институт цитологии и генетики СО РАН
14 Объединенный институт геологии, геофизики и минералогии
15 Специальная астрофизическая обсерватория РАН

Определяя силу связности научного подмножества сайтов CFC (Community Force of Connectivity) как отношение реального количества дуг к потенциально возможному, для построенной матрицы имеем CFC=0,133. Если же удалить строки и столбцы с номерами 1 и 15 (нулевые), то получаем CFC=0,205. Следует сказать, что для академических сайтов это очень высокий показатель (эксперименты с замерами на научных подмножествах, сформированных по различным признакам принадлежности, дают значения в CFC интервале от 0 до 0,35).

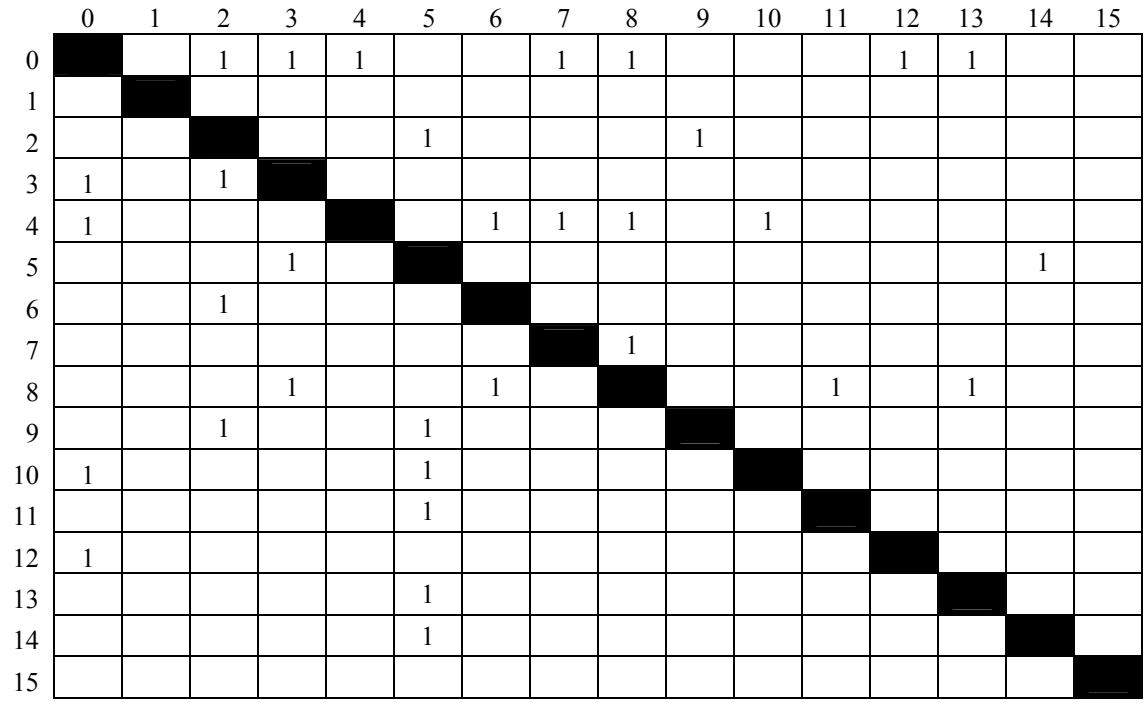

Рис. 2. Матрица смежности веб-графа сайтов организаторов конференций RCDL.

На фоне такого хорошего результата тем более удивительными выглядит картина гиперссылок с сайтов конференций на сайты организаторов выбранного нами множества (и наоборот), приведенная в табл. 1.

Конференция	Ссылки на организаторов	Ссылки с организаторов
RCDL:	1	0
RCDL 1999:	0	0
RCDL 2001:	1	2
RCDL 2002:	1	0
RCDL 2003:	1	1
RCDL 2005:	0	2
RCDL 2006:	2	2
RCDL 2007:	2	3
RCDL 2008:	2	3
RCDL 2009:	3	3
ВСЕГО	13	16

Табл. 1. Количество связей между сайтами конференций и организаторов.

В табл.1 значение, равное трем, на пересечении строки с названием «RCDL 2005: Ярославль» и столбцом «Ссылки на организаторов» означает, что с сайта конференции RCDL 2005 существуют три ссылки на различные сайты организаторов конференций (не обязательно данной конференции). Соответственно значение, равное двум, на пересечении строки с названием «RCDL 2005: Ярославль» и столбцом «Ссылки с организаторов» означает, что существуют два сайта организаторов, имеющие ссылки на сайт этой конференции. То есть за 11 лет проведения конференций с сайтов институтов, сотрудники которых (или хотя бы руководство) не могут не знать об этих конференциях, поставлено всего 16 ссылок. И если можно еще понять отсутствие ссылок на конференцию RCDL 1999: Санкт-Петербург (давно это было!), то отсутствие ссылок на общий сайт не поддается объяснению. Правда, и с основного сайта конференций сделана лишь одна ссылка на организаторов.

После этого вряд ли стоит удивляться, что из остальных академических сайтов (а их 265) ссылки на сайт какой-либо конференции RCDL сделаны лишь с шести. Соответственно, и с сайтов конференций ссылки отсылают лишь к 2 академическим сайтам (из тех же 265).

Интересно заметить, что на сайт РФФИ 13 из 15 сайтов организаторов конференций RCDL содержат гиперссылки (и не по одной, а в среднем по 8), не говоря уже о сайтах самих конференций RCDL.

Не хуже, чем сайты конференций, в качестве веб-коммуникатора для сайтов-организаторов конференции выглядит сайт РОМИП (Российский семинар по Оценке Методов Информационного Поиска) [9], на который сделаны ссылки с 2 сайтов, а он ссылается на 3.

Частичный анализ гиперссылок, сделанных с некоторых сайтов конференций RCDL, позволяет классифицировать исходящие ссылки следующим образом: четверть всех ссылок сделаны на другие сайты конференций RCDL, 17% - на сайт РОМИП, 19% - на сайты, рассказывающие о городах, в которых проводятся конференции, 8% - на Яндекс, 7% - на основного организатора текущей конференции и 5% - на РФФИ.

С помощью средств Google мы проверили, какие же сайты все-таки ссылаются на сайты конференций RCDL, если на них не ссылаются сайты организаций-учредителей из числа институтов РАН. Оказалось, что Google обнаруживает 144 ссылки, из которых 30% приходятся на сайт DELOS an Association for Digital Libraries, 30% - на все сайты RCDL вместе взятые, 10% - сайт Института информационных технологий НАН Азербайджана, по 5% - сайт журнала "Электронные ресурсы в библиотеках" и сайт РОМИП; остальные 15 сайтов с долями меньше 5%.

Заключение

Научные конференции являются коммуникационными площадками для ученых. Проведенные исследования показывают, что сайты конференций RCDL на сегодня с такой ролью справляются не в полной мере. Наверное, на это стоит обратить внимание организаторов конференций и разработчиков соответствующих веб-ресурсов.

Литература

[1] Almind T., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to "webometrics"" // Journal of Documentation. 1997. №53 (4). P. 404-426.

[2] Brin S., Page L. The Anatomy of a large scale hypertextual web search engine // Computer Networks and ISDN Systems. 1998. №30 (1-7). P. 107-117.

[3] Cronin B., Snyder H.W., Rosenbaum H., Martinson A., Callahan E. Invoked on the web // Journal of the American Society for Information Science. 1998. №49 (14). P. 1319-1328.

[4] Flake G. W., Lawrence S., Giles C. L., Coetzee, F. M. Self-organization and identification of web communities // IEEE Computer. 2002. №35. P. 66-71.

[5] Thelwall M. Extracting macroscopic information from web links // Journal of the American Society for Information Science and Technology. 2001. №52 (13). P. 1157-1168.

[6] Thelwall M. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation // Information Research. Vol. 8. №3, April 2003. [Электронный ресурс] - 2003. - Режим доступа: http://informationr.net/ir/8-3/paper151.html.

[7] Вебометрика. Институт прикладных математических исследований КарНЦ РАН. [Электронный ресурс] - 2009. - Режим доступа: http://webometrics.krc.karelia.ru.

[8] Индекс цитирования. [Электронный ресурс] - 2008. - Режим доступа: http://help.yandex.ru/catalogue/?id=873431.

[9] Российский семинар по Оценке Методов Информационного Поиска (РОМИП). [Электронный ресурс] - 2009. - Режим доступа: http://romip.ru.

[10] Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции. [Электронный ресурс] - 2009. - Режим доступа: http://www.rcdl.ru.

Are the RCDL conferenses sites scientific web-communicators?

A.A. Pechnikov, N.B. Lugovaya

In author's terminology the model named the scheme of a scientific Web, is designed from four cores a component which are not crossed subsets of scientific sites. These components are called as an administrative skeleton, the scientific subset, the nearest environs and a web-communicator. At all seeming evidence of that fact that sites of scientific conferences are a communication medium of the scientists, the conducted researches show that sites of conferences RCDL a web-communicators are not.

♣ Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-00023а)