

О СВЯЗНОСТИ МНОЖЕСТВА ОФИЦИАЛЬНЫХ САЙТОВ РОССИЙСКОЙ АКАДЕМИИ НАУК

А.А. Печников, Н.Б. Луговая, Ю.В. Чуйко

*Институт прикладных математических исследований
(Карельский научный центр Российской Академии Наук), г. Петрозаводск*

Ключевые слова и фразы: Веб; вебометрика; гиперссылка; компонента связности; организационная структура; сайт.

Аннотация: Исследованы веб-коммуникации на множестве официальных сайтов научных организаций Российской Академии Наук (РАН), реализуемые через гиперссылки между ними. Показано, что связность такого множества во многом обеспечивается за счёт административного каркаса.

Введение

Термин «вебометрика» (webometrics) введён в работе [5] для обозначения раздела информатики, в рамках которого исследуются аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к *World Wide Web* (далее – Веб). Размеры Веба приводят к концентрации внимания исследователей на его достаточно узких фрагментах, таких как множество сайтов университетов Великобритании [6], с расчётом последующего переноса полученных результатов на общие случаи. Авторами в качестве такого фрагмента, который далее будет называться фрагментом академического Веба, выбраны официальные сайты научных организаций РАН. В случае наличия у организации нескольких сайтов выбирался тот из них, который (приоритет слева направо): включен в перечень информационных систем научных учреждений РАН [1]; указан в перечне на сайте вышестоящего отделения или научного центра; на самом сайте сказано, что он является официальным сайтом организации. Для получения, хранения и обработки вебометрической информации использовался комплекс программ *WebSciRes* (от слов *Webometrics*, *Science* и *Research*), разработанный в Институте прикладных математических исследований КарНЦ РАН. В состав *WebSciRes* входят поисковый робот для сбора исходящих с сайтов гиперссылок *LPR* (*Link, Page, Robot*) и база данных, предназначенная для их хранения и обработки *DB OL* (*Data Base of OutLinks*) [3].

За август 2009 г. проведено сканирование 288 сайтов: официальный сайт РАН, 4 сайта отделений по областям науки, 13 сайтов региональных отделений и научных центров, 15 сайтов научных центров региональных отделений и 255 сайтов научных институтов, центров, станций и т.д. Обработано более миллиона *html*-страниц, найдено и сохранено более 860 000 внешних гиперссылок. Сканирование всех известных авторам официальных сайтов организаций и учреждений РАН оказалось невозможным по таким причинам, как защищённость сайтов от сканирования и неявное формирование ссылок на сайтах с использованием технологий типа *Java Script* и *Adobe Flash*.

Одним из естественных способов описания взаимосвязей сайтов посредством гиперссылок, является аппарат теории графов. Вершины веб-графа соответствуют официальным сайтам, а множество дуг для такого множества вершин отражает наличие гиперссылок между соответствующими сайтами. Такие графовые конструкции, как компоненты связности, позволяют сделать определённые выводы о коммуникациях между сайтами научных учреждений. В статье изложены результаты этих исследований.

1. Компоненты связности фрагмента Веба

Схемой академического фрагмента Веба, построенной на основе веб-графа официальных сайтов научных организаций РАН, будем называть ориентированный граф $G=(V,E)$ без петель и

кратных дуг, где V – множество вершин, соответствующих официальным сайтам РАН, а E – множество дуг, причём дуга $e(i,j) \in E$ существует тогда и только тогда, когда существует хотя бы одна гиперссылка с сайта, соответствующего вершине i , на сайт, соответствующий вершине j . В нашем случае имеем $|V|=288$ и $|E|=1304$.

Для нахождения компонента сильной связности было разработано приложение на языке *Java 2 SDK*, основанное на модификации алгоритма Флойда [2] поиска путей в ориентированном графе.

Оказалось, что $G=(V,E)$ состоит из единственной компоненты сильной связности, содержащей 175 вершин и имеющей диаметр, равный 6, а также 113-ый компонент, содержащих по одной вершине (64 вершины имеют только входящие дуги, 39 вершин имеют только исходящие дуги и 10 изолированных вершин).

Заметим, что в компоненту связности вошёл официальный портал РАН, все сайты научных и региональных отделений, подавляющее большинство сайтов региональных научных центров и научных центров в составе региональных отделений. Среди «изолированных» сайтов довольно много сайтов научных организаций, имеющих в названии слово «центр».

Дальнейшие исследования показали, что веб-коммуникации между сайтами во многом отражают административную подчинённость организаций РАН и мы постараемся обосновать это утверждение в следующем разделе.

2. Административный каркас сайтов РАН

Анализ организационной структуры РАН [4] позволяет построить иерархию, отражающую административную подчинённость между организациями-владельцами сайтов из V .

Административным каркасом схемы академического Веба называется подграф $G_k \subset G$, $G_k=(V,E_k)$, где множество дуг $E_k \in E$ формируется следующим образом: дуга $e(i,j)$ существует тогда и только тогда, когда организация i осуществляет руководство организацией j (каркасная дуга вниз), либо наоборот, организация j находится под руководством организации i (каркасная дуга вверх).

Пусть организация u осуществляет руководство организацией d . Определим каркасную функцию связности сверху вниз SFC (Skeleton Force of Connectivity) между сайтами u и d следующим образом:

$$SFC_{ud}^+ = \begin{cases} 1, & \text{существует каркасная дуга от } u \text{ к } d, \\ 0, & \text{каркасной дуги не существует} \end{cases}$$

Обозначим через D множество организаций, которые находятся под руководством организации u . Тогда функцию связности сверху вниз между сайтом u и всеми сайтами $d \in D$ определим следующим образом:

$$SFC_{uD}^+ = \frac{\sum_{d \in D} SFC_{ud}^+}{|D|}$$

Соответственно можно определить функцию связности снизу вверх для u и d (SFC_{ud}^-) и u и D (SFC_{uD}^-). Функция связности – это $SFC_{ud} = \alpha SFC_{ud}^+ + \beta SFC_{ud}^-$, где α и β являются коэффициентами, отражающими важность иерархии сверху и вниз снизу вверх ($\alpha, \beta > 0$, $\alpha + \beta = 1$). В нашем случае мы приняли α и β равными 0,5.

Результаты исследований приведены на рис. 1. Дуги помечены следующим образом: слева от дуги (или над ней) стоит значение SFC_{ud} для соответствующих u и D ; справа от дуги (или под ней) пара значений в скобках обозначает (SFC_{ud}^+, SFC_{ud}^-) для соответствующих u и D .

Таким образом, административный каркас играет большую роль в формировании академического Веба. Анализ функции связности демонстрирует, что связность каркаса достаточно велика для того, чтобы он выполнял структурно-организующее назначение. Функция связности убывает с изменением уровня иерархии сверху вниз, что, по-видимому, отражает процесс возникновения гиперссылок подчинённости организаций в его развитии. Заметим, что максимальная длина пути между любыми двумя вершинами на административном каркасе равняется шести, то есть совпадает со значением диаметра максимальной компоненты. Отсюда можно сделать вывод о том, что имеющиеся «некаркасные» дуги рассматриваемого фрагмента Веба не повлияли на уменьшение диаметра компоненты сильной связности. В свою очередь, это свидетельствует о незначительном

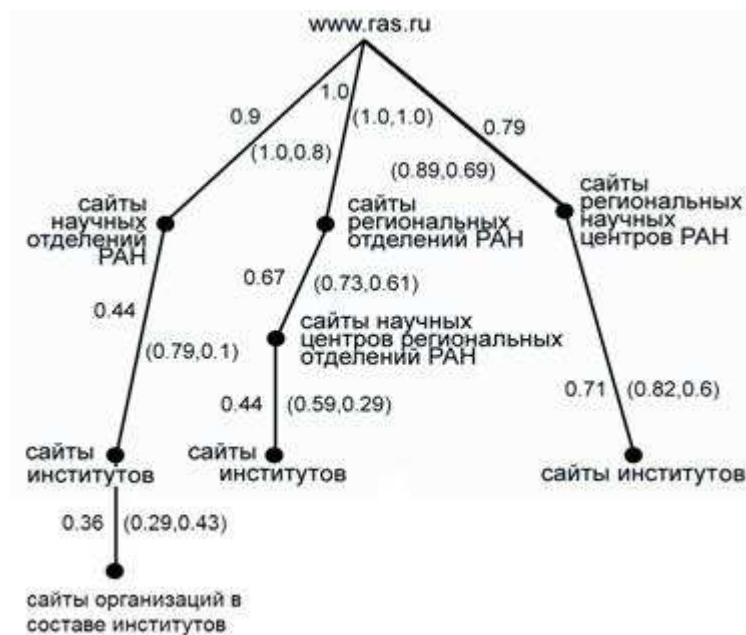


Рис. 1. Значения функции связности на административном каркасе
Заключение

отражении реальных научных связей на множестве официальных сайтов научных организаций РАН.

Не вызывает сомнения, что связность каркаса может быть значительно увеличена за счёт административных воздействий, в виде разработки и внедрения соответствующих регламентов. Относительно «некаркасных дуг» можно высказать предположение, что официальные сайты изначально не рассматриваются как научные коммуникационные площадки, всё зависит от целей, формулируемых при их создании. Возможно, такими площадками должны служить сайты научных библиотек, конференций, форумов.

Список литературы

1. Информационные системы научных учреждений Российской академии наук [Электронный ресурс]. – Режим доступа : <http://www.ras.ru/sciencestructure/informationssystem.aspx>.
2. Кристофидес, Н. Теория графов : алгоритмический подход / Н. Кристофидес. – М. : Мир, 1978. – 429 с.
3. Луговая, Н.Б. Разработка инструментов для вебметрических исследований / Н.Б. Луговая, А.А. Печников // *Телематика'2009* : Труды XVI Всероссийской научно-методической конференции (22–25 июня 2009 года). – С-Петербург, – Т. 2. – С. 341.
4. Российская академия наук. Справочник 2007. Часть II [Электронный ресурс]. – Режим доступа: <http://www.ras.ru/fstorage/download.aspx?id=209736f0-13f2-4509-882f-d90161a9b120>.
5. Almind, T. Informetric analyses on the World Wide Web : Methodological approaches to «webometrics» / T. Almind, P. Ingwersen // *Journal of Documentation*. – 1997. – № 53 (4). – P. 404–426.
6. Thelwall, M. Extracting macroscopic information from web links / M. Thelwall // *Journal of the American Society for Information Science and Technology*. – 2001. – № 52 (13). – P. 1157–1168.

Links between Numerous Websites of RAS

A.A. Pechnikov, N.B. Lugovaya, Yu.V. Chuyko

*Institute of Applied Mathematical Research
 Karelia Research Center of Russian Academy of Science, Petrozavodsk*

Key words and phrases: Web; Web metrics; hyperlink; connected component; organizational structure; website.

Abstract: The paper studies web communication at various official sites of scientific organizations of RAS which is realized via hyperlinks between them. It is shown that administrative frame ensures the cohesion between them.

© А.А. Печников, Н.Б. Луговая, Ю.В. Чуйко, 2009