Знания-Онтологии-Теории (ЗОНТ-09)

Об исследованиях российского научного Веба

Печников Андрей Анатольевич[1]

[1]*Институт прикладных математических исследований КарНЦ РАН, ул. Пушкинская, д. 11, г. Петрозаводск , 185910, Россия.*

pechnikov@krc.karelia.ru

Аннотация. *С помощью разработанного специализированного программного обеспечения собран большой объем информации о гиперссылках, исходящих с более чем 270 официальных сайтов учреждений и организаций РАН. Исследования гиперссылок позволяют предложить модель, названную схемой научного Веба, основанную на нескольких структурно составляющих подмножествах сайтов и ссылках между ними. Классификационные исследования гиперссылок показывают возможные пути развития схемы научного Веба. Схема научного Веба может служить основой для постановки и решения таких задач, как типология научных сайтов и математические модели рационального поведения веб-ресурсов, а значит, способствовать более точному пониманию природы Веба.*

Ключевые слова: Веб, вебометрика, гиперссылка, поисковый робот, классификация, организационная модель Веба.

1 Введение

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 [2,3]. Теоретические исследования показывают, что изучение гиперссылок имеет достаточный потенциал как в смысле новых источников информации и коммуникации, так и ценности самих веб-страниц [4-7].

В Институте прикладных математических исследований КарНЦ РАН вебометрические исследования проводятся с начала 2006 года [8]. Для получения, хранения и обработки информации разработан (и постоянно совершенствуется) комплекс программ WebSciRes (от слов Webomerics, Science и Research). В состав WebSciRes входят поисковый робот для сбора исходящих с сайтов гиперссылок (LPR – аббревиатура от Link, Page и Robot) и база данных, предназначенная для их хранения и обработки DB OL (аббревиатура от Data Base of OutLinks – база данных внешних ссылок). WebSciRes создан на языке PHP, работает под управлением веб-сервера Apache с интегрированным модулем PHP и СУБД MySQL.

Запись DB OL о каждой внешней гиперссылке содержит следующую информацию:

- адрес страницы, с которой сделана гиперссылка,
- контекст гиперссылки (в данной версии это выражение, к которому привязана гипертекстовая ссылка, т.е. текст, расположенный между тегами <a> и),
- адрес страницы, на которую сделана гиперссылка.

Одним из проектов, выполняемых при финансовой поддержке РФФИ, является проект по исследованию российского научного Веба. Целевое множество этого исследования представляет собой официальные сайты учреждений и организаций Российской академии наук (РАН). Результаты обследования Рунета позволяют с уверенностью говорить о 340-350 научных организациях РАН, имеющих сайты с собственными доменными именами. На момент

написания статьи DB OL содержит данные, собранные в результате сканирования 275 сайтов. Сюда входят официальный портал РАН, 4 сайта отделений РАН по областям науки, 3 - региональных отделений, 9 – региональных научных центров, 14 – научных центров региональных отделений, 238 – институтов и научных учреждений и 6 сайтов организаций, входящих в состав учреждений РАН.

Проанализировано около 1.5 млн. страниц, найдено 600 тыс. внешних гиперссылок, из которых около 70 тысяч уникальных.

Уникальная ссылка является результатом последовательного выполнения двух операций над множеством ссылок: ликвидации дублирования одинаковых гиперссылок, находящихся на одном уровне сайта и ликвидации дублирования одинаковых гиперссылок на разных уровнях. В результате этих операций вместо нескольких гиперссылок, указывающих на одну и ту же страницу сайта-мишени и имеющих одинаковый контекст, но сделанных с различных страниц сайта-источника, остается только одна гиперссылка со страницы самого высокого уровня.

В данной публикации мы изложим некоторые результаты по структурным и классификационным исследованиям гиперссылок, полученные на конец марта 2009 года.

2 Исследования гиперссылок

2.1 Структурные исследования гиперссылок

Результатом проведенных структурных исследований является теоретико-графовая модель, получившая название схемы научного Веба.

Схема научного Веба представляет собой ориентированный граф, множество вершин которого соответствует исследуемым сайтам целевого множества и всем сайтам, на которые существуют гиперссылки с сайтов целевого множества, 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.

В схеме научного Веба можно выделить четыре компоненты:

1. административный каркас, отражающий связи между сайтами, соответствующие иерархической подчиненности организаций,
2. множество научных подмножеств, где любое научное подмножество отражает связи между сайтами родственных организаций (например, сайты математических институтов или в сайты организаций данного научного центра),
3. множество ближайших окрестностей официальных сайтов; ближайшие окрестности содержат вершины, сайты которых имеют имена ddd.nnn.ss$\in$ nnn.ss, где nnn.ss – доменное имя официального сайта (например, lib.inst.ru – сайт из ближайшей окрестности inst.ru),
4. множество научных веб-коммуникаторов, т.е. сайтов, выполняющих коммуникационные функции между официальными научными сайтами.

Веб-коммуникаторы, в свою очередь, можно классифицировать по трем типам как «посредник», «индуктор» и «коммутатор». Краткое описание посредника - «много входящих ссылок, много исходящих ссылок», коммутатора – «мало входящих, много исходящих», а индуктора – «много входящих, мало исходящих». Хороший пример сайта-посредника – это Общероссийский математический портал MathNet.Ru, а коммутатора – сайт РФФИ (www.rfbr.ru). Упрощенный вариант схемы научного Веба изображен на рис. 1.

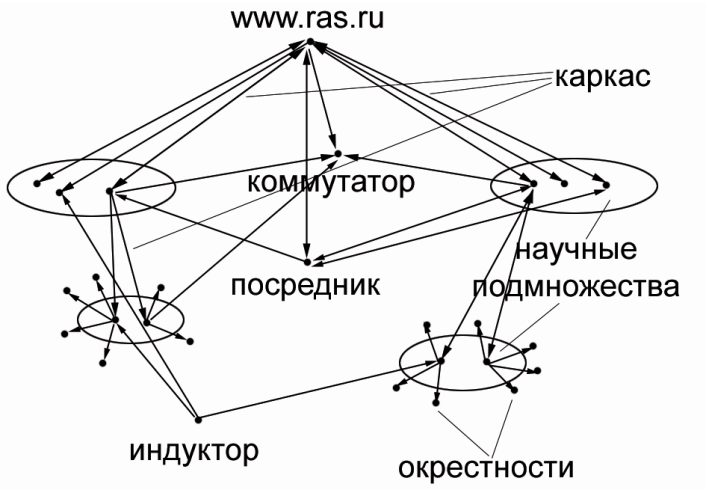

Рис.1. Схема научного Веба.

Результаты проведенных исследований показывают, что множество всех уникальных гиперссылок можно разбить на четыре подмножества в соответствии с тем, какие сайты являются их мишенями и к какой компоненте эти сайты относятся. Из 70 тысяч найденных уникальных гиперссылок к каркасу относится примерно 400 ссылок (0,5% из всего числа), к множеству научных подмножеств 3600 (5%), к множеству окрестностей около 10000 (14%) и к множеству коммуникаторов – около 40%, то есть около 28000 ссылок.

2.2 Классификационные исследования гиперссылок

В работе [9] отмечается, что без рассмотрения вопросов, связанных с классификацией ссылок и мотивацией их создания невозможно ставить и изучать задачи связности академических (academics) сайтов. Основываясь на информации о гиперссылках, содержащейся в базе DB OL, была начата работа по классификации типов гиперссылок. При этом «вручную» были обследованы внешние ссылки с 17 произвольно выбранных сайтов (единственное ограничение заключалось в том, чтобы количество исходящих уникальных ссылок было не менее 100). Обратим особое внимание на то, что нами рассматриваются именно типы гиперссылок, классификация которых основана на трех единицах анализа: исходной странице, контексте и целевой странице. Такой подход существенно отличается от подхода, предлагаемого в [6], когда классифицируются не типы гиперссылок, а типы целевых страниц. Сказанное можно пояснить следующим примером: две внешних ссылки на сайт mathem.krc.karelia.ru, размещенные на одной и той же html-странице www.krc.karelia.ru, в зависимости от контекста могут относиться к различным типам: ссылка на нижестоящую организацию и ссылка на организацию-разработчика сайта. На сегодня можно предложить некоторую предварительную типологию, представленную в таблице 1.

Таблица 1. Типология внешних гиперссылок.

№	Название типа внешней ссылки	Описание внешней ссылки
1	Вышестоящая организация	Ссылка на веб-ресурс организации, структурным подразделением которой является организация-владелец сайта.
2	Нижестоящая организация	Ссылка на веб-ресурс организации, которая является структурным подразделением данной организации.
3	Официальная организация	Ссылка на веб-ресурсы органов государственной власти федерального и республиканского уровня, а также органы местного самоуправления
4	Коммерческая организация	Ссылка на веб-ресурс организации, для которой коммерческая деятельность является основной
5	Фонды	Ссылка на веб-ресурс организации, осуществляющей финансирование проектов.
6	Коллеги	Ссылка на веб-ресурс организации, занимающейся видами деятельности, аналогичными с организацией-владельцем сайта.
7	Партнёры	Ссылка на веб-ресурс организации, с которой осуществляется совместная работа
8	Профессиональное сообщество	Ссылка на веб-ресурс профессионального общественного объединения, ассоциируемого с организацией-владельцем сайта (например, для математических институтов – сайт математического общества).
9	Другое сообщество	Ссылка на веб-ресурс общественного объединения, созданного с определённой целью (например, студенческое общество Петрозаводска или общество пчеловодов)
10	Публикации сотрудников	Ссылка на опубликованные в Вебе статью или тезисы автора(ов), работающего в организации-владельце сайта.
11	Научные труды организации	Ссылка на веб-ресурс, на котором опубликован сборник, монография, диссертация или материалы конференции организации-владельца сайта.
12	Публикации других авторов	Ссылка на публикации авторов, не работающих в организации-владельце сайта
13	Электронное издание	Ссылка на веб-ресурс издания, официально зарегистрированного как электронное, для которого электронная форма публикаций является основной
14	Новостные ленты	Ссылка на новостной веб-ресурс.
15	Научное мероприятие	Ссылка на веб-ресурс с информацией о проведении

		научной конференции, семинара, совещания и др.
16	Конкурс	Ссылка на веб-ресурс с информацией о конкурсе.
17	Справочники и руководства	Ссылка на справочник или руководство в электронном виде.
18	Доступ к базам данных	Ссылка на онлайновые базы данных.
19	Собственные проекты	Ссылка на веб-ресурс проекта, выполняемого данной организацией (возможно, совместно с другими организациями).
20	Другие проекты	Ссылка на веб-ресурс проекта, выполняемого без участия данной организации.
21	Научные журналы	Ссылка на веб-ресурс научного журнала.
22	Научные библиотеки	Ссылка на веб-ресурс научной библиотеки.
23	Официальные документы	Ссылка веб-ресурс, содержащий нормативные акты, техническую документацию, организационно-распорядительные документы и пр.
24	Доступ к программному обеспечению	Ссылка на веб-ресурс, предоставляющий возможности онлайновой загрузки программного обеспечения.
25	Альтернативный сайт	Ссылка на веб-ресурс, представляющий организацию-владельца сайта, но не рассматриваемый как официальный.
26	Личные страницы	Ссылка на персональную страницу сотрудника, расположенную на другом веб-ресурсе.
27	Баннеры	Графические изображения или текстовые блоки рекламного характера, являющиеся гиперссылкой на веб-страницу с расширенным описанием продукта или услуги.
28	Рекламные ссылки	Ссылки на информацию о товарах, услугах, развлекательных мероприятиях.
29	Разработчики сайта	Ссылка на сайт разработчиков сайта данной организации.
30	Счётчики	Ссылка на сайт разработчиков счетчика статистики.
31	Грантодатели и спонсоры	Ссылки на веб-ресурсы организаций, оказавших финансовую поддержку научных исследований и/или мероприятий.
32	Гостевые ссылки (ссылки хостеров)	Ссылки, не имеющие прямого отношения к содержанию сайта и сделанные с веб-ресурсов других организаций, размещенных на сайте организации-владельца (например, веб-страницы профсоюза сотрудников института).
33	Прочее	Все не упомянутые выше ссылки.

3 Заключение

Анализ схемы научного Веба показывает возможность систематизировать в ее рамках 60-70% всех гиперссылок, исходящих с официальных сайтов. Таким образом, сделана попытка минимальным количеством понятий описать достаточно обширный фрагмент Веба. Проводимая работа по классификации гиперссылок показывает, что количество типов ссылок не так уж и велико как могло показаться вначале, и большинство из них имеют научную мотивацию. В перспективе, используя аппарат экспертных оценок, можно построить классификационную схему, где каждый тип ссылки имеет свой вес, определяющий значимость ссылок в научном Вебе.

Отсюда можно сделать вывод о том, что структурные и классификационные исследования научного Веба могут служить информационной основой для постановки и решения таких задач, как типология научных сайтов и математические модели рационального поведения веб-ресурсов, а значит, способствует более точному пониманию природы Веба.

4 Благодарности

Работа выполняется при финансовой поддержке РФФИ (проект № 08-07-00023а).

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] Индекс цитирования. [Электронный ресурс] – 2009. – Режим доступа: http://help.yandex.ru/catalogue/?id=873431.

[4] 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.

[5] 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.

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

[7] 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.

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

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

[10] Payne N., Thelwall M. A Statistical Analysis of UK Academic Web Links // Cybermetrics. Vol. 8, Issue 1, Paper 2. [Электронный ресурс] – 2004. – Режим доступа: http://www.cindoc.csic.es/cybermetrics/articles/v8i1p2.html.