Разработка инструментов для вебометрических исследований гиперссылок научных сайтов*

А. А. Печников, Н. Б. Луговая, Ю. В. Чуйко
Учреждение Российской академии наук
Институт прикладных математических исследований КарНЦ РАН,
Петрозаводск, Россия
e-mail: pechnikov@krc.karelia.ru, nataly@krc.karelia.ru, julia@krc.karelia.ru

И. Э. Косинец
Петрозаводский государственный университет, Россия
e-mail: ikos@psu.karelia.ru

Описывается база данных, содержащая информацию о структуре гиперссылок научных веб-сайтов. Обосновывается выбор научных сайтов как целевого множества, излагаются принципы, положенные в основу алгоритмов робота-сборщика ссылок, описываются структура и возможности базы данных и приводятся некоторые результаты анализа ссылок.

Ключевые слова: веб, вебометрика, гиперссылки, базы данных, поисковый робот, краулер, классификация.

Введение

Термин "вебометрика" (webometrics) был введен в работе [1] и обозначает раздел информатики, в рамках которого исследуются количественные аспекты конструирования и использования информационных ресурсов, структур и технологий применительно к World Wide Web (далее — Веб). Одним из актуальных направлений вебометрики стали исследования, посвященные гиперссылкам (далее в качестве аналогов могут использоваться термины "ссылки" или "веб-ссылки"), поскольку они являются единственным способом взаимодействия между сайтами. Хотя может показаться, что разработчики Веба делают ссылки на страницы других сайтов случайным образом, в действительности их поведенческие модели достаточно точны для успешного использования, например, в алгоритмах информационного поиска Google и Яндекса [2, 3]. Научные исследования в данном направлении показывают, что изучение гиперссылок имеет достаточный потенциал как в смысле новых источников информации и коммуникации, так и ценности самих веб-страниц [4–6].

*Работа поддержана Российским фондом фундаментальных исследований в рамках проекта "Вебометрические исследования научных интернет-ресурсов российского Интернета" (грант № 08-07-00023-а).

Для получения больших объемов информации о гиперссылках можно применить три подхода. Первый из них заключается в использовании расширенных возможностей поисковых машин, таких как Google, Яндекс, Yahoo!Search. К примеру, если в поисковой строке Google набрать текст link: `mathem.krc.karelia.ru`, то в результатах поиска мы увидим список, указывающий примерно на 90 страниц различных сайтов, с которых сделаны ссылки на страницы сайта `mathem.krc.karelia.ru`. При дополнительной обработке отсюда можно получить определенную информацию о внешних гиперссылках. Проблемы, связанные с этим подходом, отмечаются достаточно давно [7], и основная из них заключается в отсутствии открытой информации о работе поисковых роботов.

Второй подход состоит в использовании информационных источников, созданных другими исследователями и опубликованных в доступном виде. К ним в первую очередь можно отнести ресурсы Statistical cybermetrics research group из университета Вулверхемптона [8]. На сайте исследовательской группы можно найти базы данных свободного доступа, содержащие сведения о внешних ссылках университетов Великобритании, Австралии, Новой Зеландии. В Statistical cybermetrics research group разработан и поддерживается поисковый робот SocSciBot, решающий задачи сбора внешних ссылок с университетских сайтов, который можно свободно использовать в научных целях [9].

Третий подход связан с разработкой собственного поискового робота и формированием баз данных, содержащих информацию о гиперссылках. Необходимость разработки собственного поискового робота объясняется многими причинами. Одна из них — закрытость кодов и отсутствие технической документации для таких систем, как SocSciBot (публикуется, как правило, только пособие пользователя). Проводимые исследования порождают необходимость в постоянном усовершенствовании и развитии возможностей поискового робота, поскольку накопление и осмысление информации выявляют новые задачи и соответствующие функции. Однако отсутствие документации не позволяет их реализовывать даже в виде надстроек над известными роботами. Неустойчивые каналы связи, существенно влияющие на скорость анализа сайтов, относятся скорее к техническим причинам, но в совокупности с отсутствующей документацией они ведут к вопросу о том, с какой точки происходит сканирование сайта в случае прерывания связи, а значит, к сомнениям в достоверности результатов, получаемых SocSciBot.

Однако наиболее существенной причиной является разница в научных подходах и постановке задач, что находит свое отражение в реализации алгоритмов поискового робота и структурах баз данных о гиперссылках. Один из основных аспектов исследований, определяемых авторами статьи, — неразрывность триады <страница, на которой размещена ссылка> <контекст> <страница, на которую указывает ссылка>, характеризующей любую внешнюю гиперссылку. Этому аспекту будет уделено внимание в следующих разделах статьи, сейчас лишь отметим, что он не совпадает с подходами коллег из Вулверхемптона. Поисковый робот, главной задачей которого является сбор внешних ссылок с сайтов научных организаций и учреждений Российской академии наук (РАН), далее будем обозначать аббревиатурой LPR (от английских ключевых слов Link, Page и Robot, определяющих назначение программы).

В статье дается обоснование выбора академических сайтов в качестве исследуемого целевого множества, излагаются основные принципы, заложенные в основу алгоритмов LPR, описываются структура и наполнение базы данных гиперссылок и приводятся некоторые полученные результаты.

1. Официальные сайты научных учреждений РАН как целевое множество вебометрических исследований

Многие исследователи отмечают одновременное наличие в Вебе как хаоса, так и порядка, при этом если хаос носит разносторонний и всеобъемлющий характер, то признаки порядка проявляются на некоторых его фрагментах. Если к этому добавить перманентную ограниченность исследовательских ресурсов, то следствием из утверждения о хаосе и порядке является концентрация внимания исследователей на достаточно узких фрагментах Веба, таких как уже упоминавшееся множество сайтов университетов Великобритании [6, 8, 10], с расчетом последующего переноса полученных результатов на более общие случаи. Авторами в качестве такого фрагмента Веба выбрано множество официальных сайтов научных организаций и учреждений РАН. Вспоминая 1989 год как официальную дату рождения Веба, можно сказать о том, что предложения сэра Тима Бернерса-Ли касались принципиально нового способа обмена информацией между учеными [11], а это уже могло быть основанием для выбора такого целевого множества и свидетельствовать о его достаточной зрелости. Однако и при более серьезном подходе к делу можно утверждать, что такой выбор имеет веские основания.

Известно, что Российская академия наук организована по научно-отраслевому и территориальному принципу и включает в себя 9 отделений по областям науки, 3 региональных отделения, 14 региональных научных центров, 20 научных центров региональных отделений и 470 научных учреждений (институтов, центров, музеев, станций). Подавляющее большинство организаций и учреждений РАН имеет собственные веб-ресурсы (не обязательно сайты, это могут быть разделы на сайтах вышестоящих организаций). Таким образом, мы имеем весьма обширную выборку научных сайтов организаций и учреждений, относящихся к различным областям науки, находящихся в различных иерархических отношениях и отражающих всю географию Российской Федерации.

Официальная политика в сфере информатизации научных учреждений и организаций РАН [12] позволяет сделать выводы об управляемости процессов сайтостроительства академических сайтов посредством регламентов и/или технических заданий, что делает целевое множество потенциально интересным для статистических измерений и исследований. Например, представляется интересной задача обнаружения зависимости между научной результативностью учреждений и популярностью их веб-ресурсов в научном сообществе, которую можно определить как функцию от количества и типов внешних ссылок на данный ресурс с других научных сайтов. Несмотря на то, что процесс официального оценивания результативности научных организаций находится в самом начале пути [13], есть надежды на то, что в ближайшие год-два он выйдет на необходимый уровень адекватности и прозрачности. Кроме того, имеется большое количество материалов в ежегодных отчетах о деятельности РАН, позволяющих начинать работу в данном направлении уже сейчас.

Немаловажно и проводимое зарубежными коллегами изучение академических веб-ресурсов, предоставляющее возможность сравнительного исследования фрагментов российского и, например, английского Веба. При этом если проводить аналогии между англоязычными и российскими веб-ресурсами, то понятию “academic Web” более соответствуют российские сайты научных институтов, нежели сайты университетов. Исследования (например, [14]) показывают, что в российском сегменте Веба университетские сайты в большей степени ориентированы на решение представительских, учебных

и административных задач, нежели выступают в качестве средства (и источника) для обмена научной информацией, что, по-видимому, более свойственно научным сайтам.

В случае наличия у организации нескольких сайтов в целевое множество включался тот из них, который удовлетворяет одному из следующих условий (приоритет сверху вниз):

— включен в перечень информационных систем научных учреждений РАН [15];

— указан в соответствующем перечне на сайте регионального отделения или регионального научного центра;

— указан в соответствующем перечне на сайте научного центра регионального отделения;

— на самом сайте сказано, что он является официальным сайтом учреждения.

Результаты обследования российского научного фрагмента Веба позволяют с уверенностью говорить о 340–350 научных организациях и учреждениях РАН, имеющих действующие сайты с собственными доменными именами.

2. Поисковый робот LPR

В самой простой форме работу поискового робота можно описать следующим образом: сканирование сайта начинается с начальной страницы, затем используются ссылки на этой странице для перехода на другие. Каждая страница сайта анализируется на наличие требуемой информации, которая копируется в соответствующее хранилище в случае обнаружения. Процесс повторяется с каждой новой страницей, содержащей новые ссылки для перехода до тех пор, пока не будет исследовано требуемое число страниц либо пока не будет достигнута некая цель [16]. Это простое описание вызывает массу вопросов, связанных с точным определением начальной точки входа на сайт, минимальных единиц анализа, внешних и внутренних ссылок, а также эффектами зацикливания (так называемыми “паучьими ловушками” — spider traps), неработающими ссылками, секретными и закрытыми разделами сайта и т. д. На основные из этих вопросов мы попытаемся дать ответы в этом разделе применительно к разработанному поисковому роботу LPR.

Единицей анализа LPR является страница, переданная веб-сервером клиенту по http-запросу и имеющая mime-тип “text/html”. Ссылки на объекты другого типа, например, на документы MS Office, аудио- и видеофайлы, архивы, которые мы будем называть *документами*, LPR не анализируются.

Начальным адресом сканируемого сайта считается только доменное имя сайта, без указания пути. Так как встречаются сайты, которые имеют адреса как с префиксом www, так и без него, мы будем использовать понятие *синонима адреса сайта*. Возможность наличия синонима адреса необходимо учитывать для того, чтобы при сканировании сайта различать внутренние и внешние ссылки, так как некоторые разработчики на одном сайте используют оба обозначения. Таким образом, *начальная страница сканирования сайта* — это страница, передаваемая веб-сервером по запросу URL, содержащему только доменное имя.

В данной версии программы мы рассматриваем лишь гиперссылки, заданные в тегах <a> в значении параметра href либо в тегах <frame> в значении параметра src. Если там обнаруживается абсолютный адрес вида [http[s]://доменное_имя[:протокол]/]путь_к_странице и если доменное имя отлично от адреса исследуемого сайта и его адреса-синонима, то такая ссылка считается *прямой внешней ссылкой*.

В общем случае под *контекстом внешней ссылки* понимаются языковые выражения, окружающие гиперссылку в пределах веб-страницы [17]. В данной версии LPR в качестве *контекста внешней ссылки* рассматривается выражение, к которому привязана гипертекстовая ссылка, т. е. текст, расположенный между тегами <a> и .

Определим понятие *уровня страницы* сканируемого сайта следующим образом: считаем, что начальная страница сайта является страницей нулевого уровня; тогда уровень любой другой страницы — это минимальное количество внутренних ссылок, ведущих от начальной страницы к текущей.

В процессе сканирования сайта могут встретиться внешние и внутренние ссылки, выполнение которых невозможно по ряду причин. Во-первых, это может быть ссылка, указывающая на отсутствующий ресурс (к примеру "битая ссылка", когда данная страница удалена). Во-вторых, это может быть ссылка на ресурс, закрытый для поискового робота (самый простой случай — авторизация пользователя по паролю). И, наконец, ресурс может быть недоступен в данный момент по техническим причинам (неработающий сервер, сбой в каналах связи и т. п.). Для обозначения всего класса ссылок на недоступные для LPR ресурсы мы будем использовать термин *неработающие ссылки*.

Работа LPR начинается с *начальной страницы сканирования сайта*, определяемой через начальный адрес сайта или адрес-синоним, задаваемые пользователем. Обработка любой страницы, включая начальную, начинается с http-запроса к странице и чтения заголовков, по которым выясняется несколько моментов.

1. Существует ли такая страница, и если нет, то ссылка на страницу помечается как неработающая и анализ данной страницы завершается.

2. Не является ли данная страница перенаправлением на другую страницу. Если есть перенаправление, то анализируется ссылка-перенаправление. Если это ссылка на другой сайт, то она записывается как внешняя ссылка, а анализируемая внутренняя ссылка удаляется. Если это ссылка на внутреннюю страницу, то переписывается ссылка и данный этап завершается.

3. Является ли данный объект html-страницей. Если нет, то данная ссылка помечается как "документ" и данный этап завершается.

Далее считывается код всей страницы, который проверяется на наличие перенаправлений на уровне страницы и в случае перенаправления подвергается обработке как в п. 2. Если перенаправлений нет, то из кода извлекаются теги фреймов и гиперссылок, из которых отбирается вся требуемая информация о ссылках (включая контекст), которая записывается в отдельные таблицы внутренних и внешних ссылок (более подробно структура таблиц будет описана в разделе База данных гиперссылок). Ссылки на запрос ресурсов не по протоколу http (например, ссылки на ftp-ресурсы или отправку электронной почты) и java-скрипты отсеиваются. Все ссылки, являющиеся ссылками на документы, записываются в таблицу внутренних ссылок с пометкой "документ". Остальные ссылки с пометкой "непроверенная страница", записываются в эту же таблицу, если их там еще нет.

Следует отметить, что процесс сканирования сайта идет как поиск вширь: начальная страница, потом непроверенные страницы первого уровня, потом второго и т. д. Если по каким-то причинам сканирование было прервано, то оно может быть возобновлено практически с той же точки — сканирование начинается с непроверенных страниц с наименьшим уровнем. Теоретически сканирование сайта должно продолжаться до тех пор, пока не будут проверены все его страницы, хотя LPR может быть в любой момент остановлен пользователем, поскольку он видит текущие результаты благодаря

интерактивному режиму отслеживания работы робота. На практике полное сканирование сайта не всегда приемлемо, поскольку существует по крайней мере два случая, когда поисковый робот может затратить непозволительное количество ресурсов, если не прервать его работу.

Первый случай связан со слишком большим количеством html-страниц сканируемого сайта. Практические исследования показали, что к таким сайтам можно отнести только официальный сайт РАН www.ras.ru, имеющий на первых пяти уровнях более 500 000 страниц; все остальные исследованные сайты имеют значительно меньшие размеры (подробнее об этом несколько ниже). Останов LPR на пятом уровне сайта РАН обосновывается резким уменьшением среднего количества внешних ссылок с увеличением уровня: со страницы первого уровня исходит в среднем 0.38 ссылки, второго — 1.24, а третьего — всего 0.05. Видимо, такую рекомендацию для прерывания работы LPR можно принять в общем случае для больших сайтов.

Второй случай в общей постановке известен достаточно давно и называется "паучья ловушка", когда разработчики сайтов (умышленно или неумышленно) создают условия для зацикливания поискового робота [16]. Поскольку не существует универсального алгоритма для распознавания паучьих ловушек, а процессы возникновения новых типов ловушек идут достаточно быстро, невозможно дать и универсальные рекомендации для прерывания работы робота.

Наши исследования выявили несколько примеров паучьих ловушек на научных сайтах, которые в настоящее время устраняются вручную. К ним относится организация меню сайта в виде дерева, когда в URL передается состояние открытых узлов дерева. Для пользователя содержательная часть страницы может выглядеть одинаково, но из-за разного состояния меню генерируются разные ссылки и разный html-код. Еще один пример — это организация страниц в виде дерева (всевозможные карты и рубрикаторы), когда в URL также передается параметр с состоянием дерева.

Разрабатываемый программный комплекс развивается и совершенствуется в процессе исследований, в него постоянно добавляются новые правила, причем эти правила могут быть связаны не только с паучьими ловушками, но и с некорректным веб-программированием. Понятно, что добавление новых правил, исполняемых в автоматическом режиме, возможно только после многократной ручной проверки их работы. В статье описана версия, реализованная на октябрь 2008 года. Весь программный комплекс, состоящий из LPR и базы данных внешних гиперссылок, разработан на языке PHP и работает под управлением веб-сервера Apache с интегрированным модулем PHP и СУБД MySQL.

3. База данных внешних гиперссылок

База данных внешних гиперссылок научных сайтов РАН (БД ВГ РАН) представляет собой множество взаимосвязанных реляционных таблиц (основных и вспомогательных) и набор операций над ними. К основным таблицам относятся таблица ИНФОРМАЦИЯ О САЙТАХ и множество пар таблиц ВНУТРЕННИЕ ССЫЛКИ и ВНЕШНИЕ ССЫЛКИ, связанных с сайтами, отсканированными LPR.

Запись таблицы ИНФОРМАЦИЯ О САЙТАХ имеет следующую структуру:

<ID_S><DNS_name><DNS_syn><NAME><ADD_inf>,

где ID_S — идентификатор сайта; DNS_name — доменное имя — начальный адрес сайта; DNS_syn — синоним начального адреса; NAME — название сайта; ADD_inf — дополнительная информация, такая как вспомогательные данные, необходимые при выполнении запросов. Кроме того, это могут быть вебометрические сведения о сайте, такие как тематический индекс цитирования [3], если они не оформлены как отдельная таблица.

Значения всех полей записи (кроме вспомогательной информации) задаются пользователем при очередном пуске LPR для сканирования нового сайта.

Запись таблицы ВНУТРЕННИЕ ССЫЛКИ с признаком ID_S для каждого сайта имеет следующую структуру:

<ID_P><PATH_S><ID_P_parent><LEVEL><SIGN><TIME>,

где ID_P — идентификатор страницы; PATH_S — путь к странице; ID_P_parent — идентификатор страницы, содержащей ссылку на данную страницу; LEVEL — уровень страницы; SIGN — статус страницы (0 — необработанная страница, 1 — обработанная страница, 2 — неработающая страница, 3 — страница типа “документ”); TIME — время последнего обновления записи.

Значения полей ID_P и TIME генерируются автоматически, остальные создаются LPR в процессе обработки очередной ссылки.

Запись таблицы ВНЕШНИЕ ССЫЛКИ с признаком ID_S для каждого сайта имеет следующую структуру:

<ID_L><DNST_name><PATH_T><ID_P><LEVEL><CONTEXT><ADD_inf>,

где ID_L — идентификатор внешней ссылки; DNST_name — доменное имя сайта, на который ссылается внешняя ссылка; PATH_T — путь к странице; ID_P — идентификатор страницы сайта, с которой сделана внешняя ссылка; LEVEL — уровень страницы; CONTEXT — контекст внешней ссылки; ADD_inf — дополнительная информация.

Значение поля ID_L генерируется автоматически, остальные значения записи создаются LPR в процессе обработки очередной ссылки.

Основной операцией над множеством таблиц является операция выборки внешних ссылок, исходящих с заданного сайта (операция ВЫБОРКА). В запросе пользователь задает доменное имя анализируемого сайта (или его название). Результатом запроса является таблица, имеющая следующую структуру записей:

<N> <DNST_name><PATH_T><CONTEXT><DNS_name><PATH_S><LEVEL>.

Обратим внимание на то, что пара <DNST_name><PATH_T> полностью идентифицирует страницу, на которую сделана ссылка (так называемые “целевая страница” и соответственно “целевой сайт”), а пара <DNS_name><PATH_S> — страницу, с которой сделана ссылка (“исходная страница”, “исходный сайт”). Операция ВЫБОРКА позволяет получить данные обо всех внешних ссылках, сделанных с сайта DNS_name, с максимально полной информацией о целевой странице, контексте ссылки, исходной странице и ее уровне.

При выборе данных пользователь может уточнить запрос с помощью нескольких фильтров, таких как имя целевого сайта, путь к целевой странице, контекст, путь к исходной странице и ее уровень.

Эксперименты показали, что на одном и том же уровне сайта может встречаться немало одинаковых ссылок, сделанных с разных страниц. Характерный пример — ссылка в виде логотипа фирмы-спонсора конференции, проводимой институтом. Для устранения подобных случаев дублирования на множестве записей, отобранных операцией ВЫБОРКА, определена операция УНИФИКАЦИЯ, которая при наличии записей с одинаковыми значениями полей <DNST_name>, <PATH_T>, <CONTEXT> и <LEVEL> оставляет только одну из них (с минимальным значением ID_L).

Для устранения дублирования одинаковых внешних ссылок с различных уровней сайта на множестве записей, отобранных операцией УНИФИКАЦИЯ, определена операция МИНИМИЗАЦИЯ, которая при наличии записей с одинаковыми значениями <DNST_name>, <PATH_T> и <CONTEXT> (значения остальных полей могут быть различными) оставляет запись с минимальным значением <LEVEL>.

Последовательное применение операций УНИФИКАЦИЯ и МИНИМИЗАЦИЯ приводит к ликвидации повторяющихся внешних ссылок с заданного сайта. Полученное таким образом множество внешних ссылок будем называть уникальным.

4. Некоторые результаты исследований

Исследования российского фрагмента научного Веба, описываемые в статье, начались в феврале 2008 года. Целевое множество, вебометрические индикаторы и внешние гиперссылки постоянно обновляются и пополняются, в статье используются данные исследований на начало декабря 2008 года. Отметим, что рассматриваемая в статье БД ВГ РАН — составная часть более крупного информационного комплекса — базы данных вебометрических исследований РАН (БД ВИ РАН), включающей в себя также компоненты, содержащие данные о вебометрических индикаторах исследуемых сайтов, общую статистику сайтов (число уровней, страниц, документов), данные о неработающих ссылках и многие другие.

Целевое множество исследований содержит 344 записи об официальных сайтах организаций и учреждений РАН (названия организаций и доменные имена) и опубликовано на сайте “Вебометрика” [18] в виде таблицы *.xls. Там же в табличном виде даны результаты измерений основных вебометрических индикаторов сайтов целевого множества.

Проведено сканирование 208 сайтов целевого множества, включая официальный сайт Российской академии наук, 13 сайтов региональных отделений и научных центров, 13 сайтов научных центров региональных отделений и 185 сайтов научных учреждений (институтов, центров, музеев, ботанических садов и т. д.). Общее количество обработанных html-страниц равно 900 000. При этом такой показатель, как среднее количество html-страниц сайта института, поражает своим разнообразием: при среднем значении, равном 1860 страниц, 40 сайтов содержат до 100 страниц, а 7 сайтов — более десяти тысяч.

БД ВГ РАН содержит данные о более чем 520 000 различных внешних ссылках. Операция унификации приводит к тому, что остается 89 000 ссылок (среди которых возможны ссылки на одну и ту же целевую страницу), а последовательная унификация и минимизация оставляет 64 000 уникальных ссылок. Среднее число уникальных внешних гиперссылок, исходящих с одной html-страницы, равно 0.071. Интересно отметить, что если вычислять это значение только для сайтов научных учреждений, то получается значение 0.1535, весьма близкое к результату 0.1589, описанному в [19] для сайтов

университетов Великобритании. Онлайновый доступ к основным разделам БД ВГ РАН реализован в разделе "Исследование внешних ссылок сайтов РАН" сайта [18].

В работе [6] отмечается, что без рассмотрения вопросов, связанных с классификацией ссылок и мотивацией их создания, невозможно ставить и изучать задачи веб-связности академических сайтов. На основе информации о гиперссылках, содержащейся в БД ВГ РАН, была начата работа по классификации типов гиперссылок. Авторами "вручную" обследованы внешние ссылки с 17 произвольно выбранных сайтов (единственное ограничение заключалось в том, чтобы число исходящих 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 Statistical cybermetrics research group [6, 8], 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. Многократные дискуссии авторов и повторные итерации по типологизации гиперссылок позволяют предложить следующую предварительную типологию.

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 органов 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 деятельностью, аналогичной деятельности организации-владельца сайта.

7. **Партнеры.** 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 определенной 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. **Счетчики.** Ссылка на сайт разработчиков счетчика статистики.

31. **Грантодатели и спонсоры.** Ссылки на веб-ресурсы организаций, оказавших финансовую поддержку научным исследованиям и/или мероприятиям.

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.

В работе [19] отмечается, что далеко не все ссылки с сайтов университетов Великобритании ведут на целевые страницы с научным контентом. Классификация внешних ссылок некоторых научных сайтов, проведенная авторами в соответствии с описанными типами, лишь частично подтверждает правоту этого утверждения для российского фрагмента научного Веба. К примеру, баннеры, рекламные ссылки, ссылки на счетчики статистики и разработчиков сайта, а также гостевые ссылки составляют лишь около 20 % всех уникальных ссылок. Из ссылок на страницы с научным контентом 20 % — это

ссылки на публикации, журналы, базы данных, 11 % — на состоявшиеся или анонсируемые конференции и совещания.

Вместе с тем исследования показывают весьма слабую связность сообщества научных сайтов. Если из множества уникальных внешних гиперссылок выделить подмножество "своих ссылок", т. е. ссылок, сделанных на страницы сайтов организаций и учреждений РАН, то подсчеты для всего множества исследованных сайтов показывают, что соотношение своих ссылок ко всем исходящим ссылкам составляет лишь 3.8 %. Это же соотношение для официального сайта РАН составляет 21.2 %, для региональных отделений — 9.7, для научных центров — 8.4, а для институтов — 2.2 %. Закономерность таких соотношений объяснима, поскольку чем выше в организационном плане находится организация, тем большее количество ссылок на подчиненные организации должно появиться на ее сайте. И все-таки средний показатель для институтов, равный шести ссылкам на сайты РАН, кажется слишком низким.

Исследования оставшихся 96.2 % "чужих" ссылок (т. е. ссылок, сделанных с официальных научных сайтов РАН на все другие сайты), позволяют разделить сайты, на которые они сделаны, на два непересекающихся подмножества: сайты веб-окружения и сопутствующие сайты.

Веб-окружением официального сайта называются веб-ресурсы организации-владельца сайта, имеющие доменные имена в зоне доменного имени официального сайта, на которые существуют ссылки с официального сайта. Например, для сайта Института вычислительной математики и математической геофизики Сибирского отделения РАН (www.sscc.ru) в веб-окружение попадает сайт отдела статистического моделирования в физике этого института (osmf.sscc.ru). Из обследованных 344 сайтов организаций и институтов РАН 127 имеют сайты веб-окружения. Общее количество сайтов веб-окружения — около 600, и на них с официальных сайтов сделано около 4000 гиперссылок.

Подмножеством сопутствующих сайтов будем называть все сайты, не являющиеся официальными сайтами научных организаций и учреждений РАН или сайтами их ближайшего окружения, на которые существуют гиперссылки с официальных сайтов. На сегодня найдено около 20000 сопутствующих сайтов, на которые сделано около 30 000 ссылок с официальных сайтов. Отметим большое количество ссылок, сделанных на сайты поисковых систем и сайты фирм, предоставляющих возможности по анализу статистики (около 400 ссылок с более чем 200 официальных научных сайтов). Ссылки на сайты поисковых систем, как правило, имеют сервисный характер для пользователей, а ссылки на сайты статистики являются обязательными в случае использования их возможностей. Видимо, с точки зрения организации научного Веба эти сайты не имеют большого значения и их можно исключить из рассмотрения.

Абсолютным лидером по количеству сайтов, с которых сделаны ссылки на заданный сайт, стал сайт Российского фонда фундаментальных исследований (www.rfbr.ru и www.rffi.ru). На него сделано 500 уникальных ссылок со 120 научных сайтов. На втором месте сайт Научной электронной библиотеки (www.elibrary.ru): 80 ссылок с 60 сайтов. Перечень сайтов сопутствующего множества, на которые сделаны ссылки более чем с 10 научных сайтов, ограничивается первой сотней, зато всего лишь по два сайта ссылаются на 5000 сайтов сопутствующего множества, а по одному сайту — на 15 000 сайтов.

Всего на сайты первой сотни сопутствующего множества сделано около 6000 уникальных ссылок (почти 10 % от всех ссылок), а на первые 4000 сайтов — около 11 000 (17 %). Причем в первые 400 сайтов сопутствующего множества входят те сайты, на ко-

торые ссылаются более пяти научных сайтов. Таким образом, можно выделить достаточно компактное множество для дальнейших исследований, включающее около 350 официальных научных сайтов, 600 сайтов веб-окружения и 400 сопутствующих сайтов.

Заключение

К ближайшим задачам как логическому продолжению изложенных результатов можно отнести развитие возможностей LPR и наполнение БД ВГ РАН данными о ссылках с новых сайтов целевого множества. Кроме того, следует уделить внимание более тщательному изучению сайтов целевого множества и удалению из него ряда сайтов, не поддерживаемых их разработчиками и владельцами.

В более широком плане описанная в статье база данных, содержащая сведения о внешних ссылках с сайтов научных организаций и учреждений РАН, создаваемая с помощью разработанного авторами робота-сборщика ссылок, служит информационной основой для постановки и решения таких задач, как классификация гиперссылок, типология научных сайтов и, в перспективе, оптимизационные математические модели рационального поведения веб-ресурсов, а значит, она способствует более точному пониманию природы Веба.

Список литературы

[1] Almind T., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to "webometrics" // J. of Documentation. 1997. Vol. 53, N 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. Vol. 30, N 1–7. P. 107–117.

[3] Индекс цитирования. [Электронный ресурс]. 2008. 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 // J. of the American Society for Information Science. 1998. Vol. 49, N 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. N 35. P. 66–71.

[6] Thelwall M. Extracting macroscopic information from web links // J. of the American Society for Information Science and Technology. 2001. Vol. 52, N 13. P. 1157–1168.

[7] Bar-Ilan J. How much information the search engines disclose on links to a web page? — A case study of the "Cybermetrics" home page // Proc. 8th Intern. Conf. on Scientometrics and Informetrics. 2001. Vol. 1. P. 63–73.

[8] Statistical cybermetrics research group. [Электронный ресурс]. 2008. Режим доступа: http://cybermetrics.wlv.ac.uk

[9] SocSciBot. [Электронный ресурс]. 2008. Режим доступа: http://socscibot.wlv.ac.uk.

[10] Thelwall M. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation // Information Research. 2003. Vol. 8, N 3. [Электронный ресурс]. 2003. Режим доступа: http://informationr.net/ir/8-3/paper151.html

[11] Berners-Lee T. Information Management: A Proposal. [Электронный ресурс]. 1998. Режим доступа: http://www.w3.org/History/1989/proposal.html

[12] Единая информационная система РАН. [Электронный ресурс]. 2008. Режим доступа: http://www.ras.ru/scientificactivity/eis.aspx

[13] Министерство образования и науки РФ. Типовая методика оценки результативности научных организаций государственного сектора в Российской Федерации (проект). [Электронный ресурс]. 2008. Режим доступа: http://www.mon.gov.ru/work/nti/dok/gsn/tip-metod.doc

[14] Печников А.А. Вебометрические исследования web-сайтов университетов России // Информационные технологии. 2008. Vol. 11. С. 74–78.

[15] Информационные системы научных учреждений РАН. [Электронный ресурс]. 2008. Режим доступа: http://www.ras.ru/sciencestructure/informationsystems.aspx

[16] Pant G., Srinivasan P., Menczer F. Crawling the Web // In Web Dynamics. Springer, 2004. Levene M. and Poulovassilis A., eds. P. 153–178.

[17] Pant G., Srinivasan P. Link Contexts in Classifier-Guided Topical Crawlers // IEEE Transactions on knowledge and data engineering. 2006. Vol. 18, N. 1. P. 107–122.

[18] Вебометрика. Институт прикладных математических исследований КарНЦ РАН. [Электронный ресурс]. 2008. Режим доступа: http://webometrics.krc.karelia.ru

[19] 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

Поступила в редакцию 27 декабря 2008 г.,
в переработанном виде — 20 апреля 2009 г.