УДК 004.7

Ю. И. Шокин, **О. А. Клименко**, **И. С. Петров**

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

E-mail: klimenko@ict.nsc.ru

АНАЛИЗ СВЯЗЕЙ МЕЖДУ САЙТАМИ ИНСТИТУТОВ СИБИРСКОГО ОТДЕЛЕНИЯ РАН

В работе проведено исследование структуры и динамики развития научного веб-пространства на примере Сибирского отделения Российской академии наук (СО РАН). Для исследования сайтов была создана оригинальная программа, которая учитывает ряд особых ситуаций, которые влияют на полноту и качество собранных данных. Результаты представлены в виде графов, в которых узлы – это сайты, а ребра – гиперссылки. Графы взаимосвязей показали, что некоторые организации взаимно ссылаются друг на друга, другие организации имеют множество исходящих ссылок, третьи изолированы, а четвертые имеют много входящих ссылок, что говорит о ценности информации, размещенной на сайте.

Ключевые слова: веб-пространство, вебометрика, гиперссылка, сайт.

Введение

В настоящее время проблема изучения веб-пространства является актуальной в связи со стремительным развитием сети Интернет и количества ресурсов, представленных в ней. В частности, анализ веб-пространства, связанного с научной деятельностью, позволяет определить, насколько та или иная научная организация следует мировым тенденциям развития и представляет результаты научных работ на своем сайте.

Целью нашего исследования является определение структуры и динамики развития научного веб-пространства и выявление проблем взаимодействия институтов и организаций на примере Сибирского отделения Российской академии наук (СО РАН). Дополнительной целью нашей работы является создание инструментария и методов исследования сайтов.

СО РАН является региональным объединением научно-исследовательских, опытно-конструкторских, производственных организаций РАН, а также подразделений, обеспечивающих функционирование инфраструктуры научных центров, расположенных на территории Сибири в 7 областях, 2 краях и 4 республиках (общая площадь территории около 10 млн кв. км). Каждый из 77 институтов, а также филиалы имеют свои сайты, где представлена научная, научно-организационная, образовательная и другая информация о деятельности учреждения. Была поставлена задача определить, как представлены в веб-пространстве институты СО РАН, есть ли связи между сайтами, какая информация привлекает наибольшее число внешних ссылок.

Подходы к исследованию веб-пространства

Термин «вебометрика» (*webometrics*) был введен в работе T. Almind и P. Ingwersen в 1997 г. [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-пространство). К актуальным направлениям вебометрики относятся исследования гиперссылок (аналогичные термины –

«ссылка», «веб-ссылка»), которые являются способом взаимодействия между сайтами. Практическая применимость этих исследований успешно демонстрируется реализацией алгоритмов информационного поиска таких популярных систем, как Google и Яндекс [2; 3]. Научные исследования в этом направлении показывают, что изучение гиперссылок имеет достаточный потенциал как в смысле новых источников информации и коммуникации, так и ценности самих веб-страниц [4–6]. Для получения больших объемов информации о гиперссылках можно применить три подхода.

Первый из них заключается в использовании возможностей поисковых машин, таких как Google, Yahoo, Яндекс, Rambler (эти системы наиболее полно индексируют русский сегмент Интернета). С помощью расширенных возможностей этих систем можно получить данные о количестве страниц на сайтах, количестве внешних ссылок, количестве так называемых «тяжелых» файлов (pdf, doc, ppt). Проблемы, связанные с этим подходам, известны достаточно давно [7] и основная из них – это отсутствие открытой информации о работе поисковых роботов, а также сложность в извлечении этих данных, так как все поисковые системы вводят ограничения на автоматический сбор информации. К тому же данные, предоставляемые поисковыми системами, не дают всю нужную информацию. Для построения рейтинга сайтов их достаточно, но для более глубокого исследования веб-пространства их не хватает.

Второй подход состоит в использовании информационных источников, созданных другими исследователями и опубликованных в доступном виде, например, результаты исследований, проводимых в ИПМИ КарНЦ РАН [1]. Однако в свободном доступе нет программы для сбора информации и открытой базы данных.

Существуют ресурсы *Statistical cybermetrics research group* из университета Вулверхемптона [2]. На сайте этой исследовательской группы можно найти базы данных свободного доступа. В этих базах содержится информация о сайтах университетов Великобритании, Австралии и Новой Зеландии. В *Statistical cybermetrics research group* разработан и поддерживается поисковый робот SocSciBot, его можно свободно использовать в научных целях [3].

Основные минусы этого подхода.

1. Программы имеют закрытый исходный код, нельзя оценить точность работы программы.

2. Отсутствует либо закрыта для свободного доступа документация по программам, что затрудняет их использование и делает невозможным создания надстроек к ним для расширения стандартного функционала.

Третий подход связан с написанием своей программы-краулера, которая путем обхода и анализа всех страниц на заданном множестве сайтов выявляет связи между элементами множества. В результате работы программы получается база данных, в которой содержится разнообразная информация о сайтах (ссылки, количество страниц на сайте, количество «тяжелых» файлов, ключевые слова и т. д.). Основной плюс этого подхода в том, что программа пишется под определенные задачи, в любой момент можно исправить ошибки или добавить нужный функционал.

Особенности извлечения данных

Существует несколько программ-краулеров, работающих по схожему алгоритму [8], однако для данной задачи было создана оригинальная программа, которая учитывает ряд особых ситуаций, которые влияют на полноту и качество собранных данных.

1. Учет внешних ссылок. При построении графа учитываются только «внешние ссылки» – ссылки с одного сайта на другой. Как известно, внутренние ссылки зачастую не несут смысловой нагрузки, а являются частью навигационной системы сайта.

2. Обработка сценариев. На многих сайтах, которые имеют навигационное меню, URL разделов этого меню не содержатся в ссылках на страницы, они содержатся в сценариях Javascript, встроенных в страницу. Такие ситуации нужно обрабатывать индивидуально для

[1] Вебометрика. Институт прикладных математических исследований КарНЦ РАН. 2009. URL: http://webometrics.krc.karelia.ru

[2] Statistical cybermetrics research group. 2009. URL: http://cybermetrics.wlv.ac.uk

[3] SocSciBot. 2009. URL: http://socscibot.wlv.ac.uk

каждого сайта, поскольку сценарии устроены по-разному. Для рассматриваемых сайтов были выделены конструкции в сценариях, содержащие ссылки.

3. Обработка ошибок. Некоторые ссылки устарели и ведут на несуществующие сервера и страницы. Для того, чтобы отследить такие ситуации, все ошибки сохраняются в базе данных.

4. Учет синонимов доменных имен. Часто у сайта есть не одно доменное имя, а несколько, поэтому необходимо отслеживать количество уникальных страниц и ссылок. Наиболее часто встречающийся пример: использование «www» префикса, например www.math.nsc.ru и math.nsc.ru указывают на один и тот же сервер.

Вся полученная информация сохраняется в базе данных. Структура базы данных включает:

1) имя домена;
2) название сайта;
3) исходящие ссылки с сайта;
4) количество «тяжелых» файлов (doc, pdf, ppt, rtf и т. д.);
5) ключевые слова, которые содержатся в названии страницы;
6) ключевые слова, которые содержатся в анкорах (околоссылочном тексте).

Первые результаты, полученные с помощью написанной программы-краулера [4], позволили впервые наглядно представить веб-пространство, связанное с СО РАН. Сейчас происходит развитие функционала программы-краулера, создание модуля анализа контекста, в котором встречаются ссылки. Это нужно, чтобы точно ответить на вопрос, в какое место на странице ведет ссылка, есть ли взаимосвязь этих страниц по смыслу, а не только по ссылке. Необходимо увеличить стабильность работы программы, так как при большом количестве страниц требуются большие вычислительные ресурсы и ресурсы памяти. Также необходимо разработать новые методы сбора данных, которые бы учитывали особенности защиты сайтов от поисковых роботов.

Алгоритм построения графа

Результаты исследований были представлены в виде графов. В этом графе узлы соответствуют сайтам организаций, направленные дуги соответствуют гиперссылкам. Граф строится на основе данных, полученных программой-краулером (crawler). Для хранения собранных данных используется база данных. Начальные условия задачи – набор адресов сайтов, которые предстоит исследовать, назовем его множеством S.

Работа алгоритма состоит в последовательной обработке сайтов. Для каждого сайта из множества S определена очередь необработанных страниц Q_i, в которой в начале находится главная страница сайта. При просмотре каждой страницы из очереди из нее извлекаются все ссылки, которые затем классифицируются и над ними производятся соответствующие действия.

- Ссылки на другие страницы этого сайта, которые еще не были просмотрены. Страницы, на которые указывают эти ссылки, добавляются в очередь Q_i.
- Ссылки на страницы этого сайта, которые либо просмотрены, либо находятся в очереди Q_i; с этими страницами никаких действий не производится.
- Ссылки на другие сайты из множества S. Страницы, на которые указывают такие ссылки, добавляются к сайту, однако в очередь для обработки не помещаются. Ссылки сохраняются в базе данных.
- Ссылки на страницы сайтов не из множества S. Такие ссылки сохраняются в базе данных, но страницы не просматриваются.

Помимо URL из ссылок также извлекается текст, который заключен между открывающим и закрывающим тегами <a>. Из страниц, в свою очередь, извлекаются их заголовки (тег <title>), а также ключевые слова. Эта информация будет использована для выделения ключевых слов, связанных с теми или иными объектами.

[4] Единая информационная система РАН. 2008. URL: http://www.ras.ru/scientificactivity/eis.aspx

Ведется работа над более интересным и информативным графическим представлением веб-пространства. В этом направлении планируется использовать готовые решения, которые позволяют создавать интерактивные модели, так как эти модели являются наиболее выразительными.

Некоторые результаты работы

В исследовании участвовало 95 сайтов, 85 сайтов институтов СО РАН и 10 внешних сайтов. Были получены графы связей между институтами одной тематической направленности (математические, физические, химические и т. д.), между институтами региональных научных центров (Иркутский, Красноярский, Томский и т. д.). Также были рассмотрены связи между институтами, выполняющими общие интеграционные проекты.

Различные способы визуализации позволили наглядно представить часть веб-пространства, связанную с научной деятельностью СО РАН. С помощью визуализации были найдены подмножества наиболее связанных сайтов, которые образуют скопления в веб-пространстве.

Анализ взаимосвязей всех сайтов СО РАН и разных групп сайтов позволил сделать выводы об устройстве части веб-пространства, связанного с СО РАН:

- наиболее сильно связаны между собой физико-математические институты;
- меньше всего между собой связаны сайты геологических институтов;
- у некоторых институтов есть сильные междисциплинарные связи, например между химическими и биологическими институтами;
- междисциплинарные связи возникают между сайтами институтов, участвующих в интеграционных проектах;

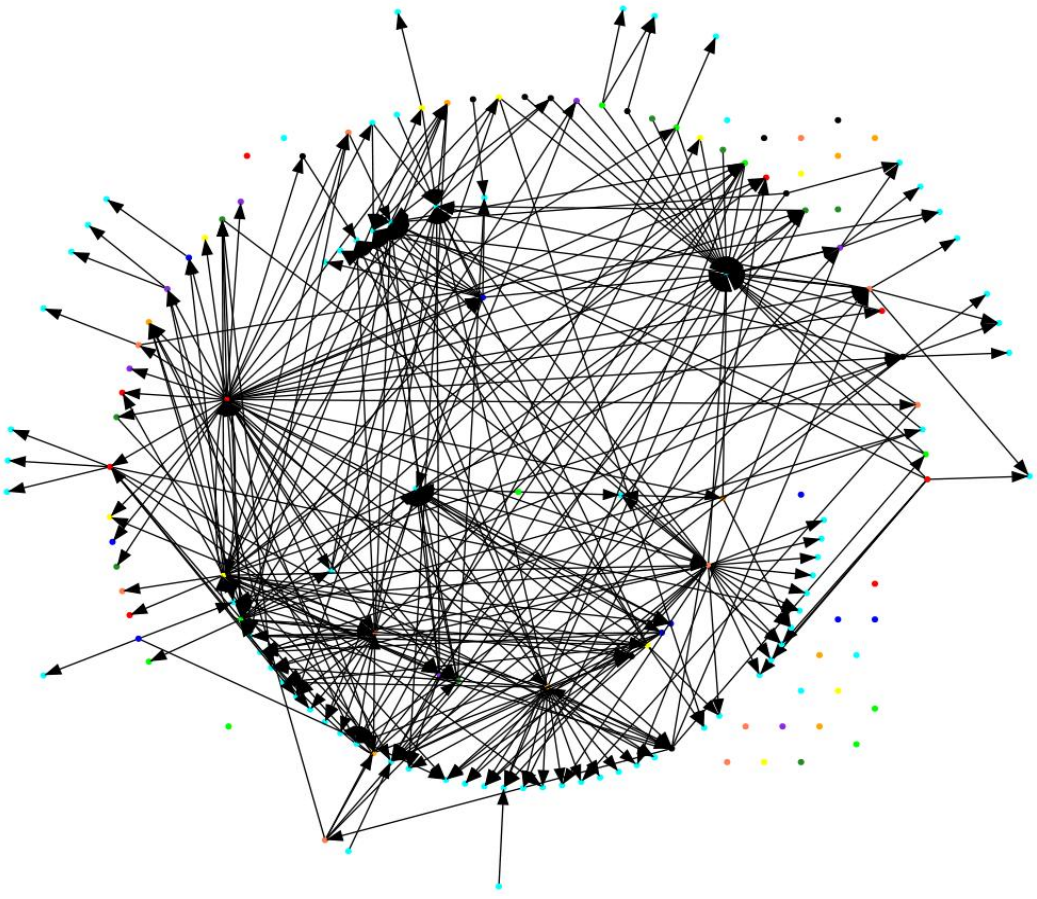

Рис. 1. Представление связей между сайтами институтов СО РАН в виде ориентированного графа

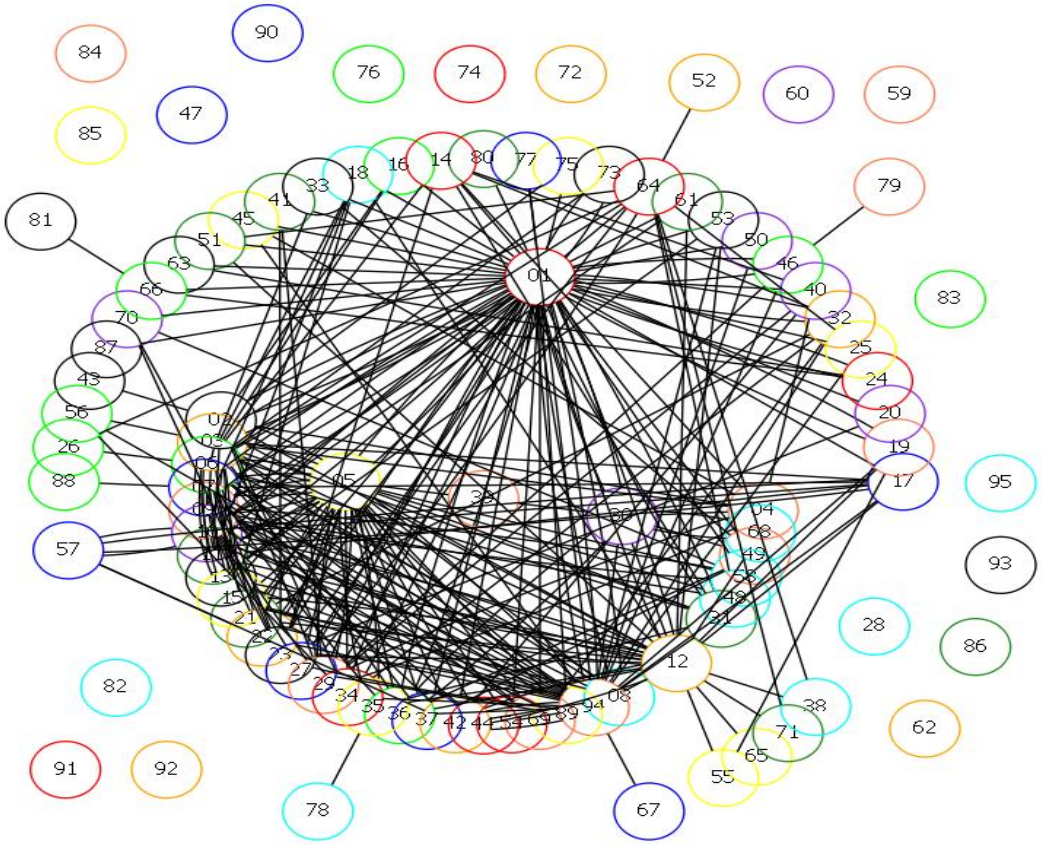

Рис. 2. Представление связей между сайтами институтов СО РАН в виде неориентированного графа

- связи между институтами региональных научных центров слабые, за исключением Иркутского и Красноярского научных центров;
- связи между НГУ и институтами СО РАН сильнее, чем связи между СФУ и институтами СО РАН;
- сайт Института автоматики и процессов управления Дальневосточного отделения РАН хорошо связан с институтами СО РАН физико-математической тематики;
- существует около 20 сайтов с числом входящих и исходящих ссылок более 100;
- существует 10 сайтов с числом входящих и исходящих ссылок более 1 000.

На рис. 1 точками обозначены сайты, стрелками – гиперссылки. Видно, что примерно пятая часть всех сайтов никак не связана с другими. Есть большая группа сайтов, которые слабо связаны с другими.

На рис. 2 указаны номера, которые позволяют определить, какие именно сайты являются центрами скоплений. Это Портал СО РАН (01), сайт Объединенного ученого совета СО РАН по нанотехнологиям и информационным технологиям (02), Новосибирский государственный университет (03), Сибирский федеральный университет (04), Институт вычислительных технологий (05), сайт Президиума СО РАН (06), Институт цитологии и генетики (07), Институт ядерной физики (08), Институт математики им. С. Л. Соболева (09), Государственная научно-техническая библиотека (12).

Заключение

Исследования показали, что чем более развит сайт организации, чем больше на нем представлено полных текстов публикаций, проектов, материалов конференций, баз данных, тем лучше сайт известен в научных кругах. Это объясняется тем, что на него больше ссылаются

сотрудники других организаций, так как они имеют свободный доступ к результатам исследований. Эти результаты совпали с результатами, полученными экспертами при проведении конкурса сайтов Организаций СО РАН в 2010 г. [9].

Список литературы

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. 2009. URL: http://help.yandex.ru/catalogue/?id=873431.

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. *Шокин Ю. И., Клименко О. А., Рычкова Е. В., Шабальников И. В.* Рейтинг сайтов научных организаций СО РАН // Вычисл. технол. 2008. Т. 13, № 3. С. 128–135.

7. *Thelwall M.* What Is This Link Doing Here? Beginning a Fine-Grained Process of Identifying Reasons for Academic Hyperlink Creation // Information Research. April 2003. Vol. 8. № 3. URL: http://informationr.net/ir/8-3/paper151.html.

8. *Клименко О. А., Петров И. С.* Исследование строения и динамики развития научного Веб-пространства на примере СО РАН // Тр. XVI Байкальской Всерос. конф. «Информационные и математические технологии в науке и управлении». Иркутск: ИСЭМ СО РАН, 2010. Ч. 3. С. 92–97.

9. *Бычков И. В., Клименко О. А.* Названы победители конкурса сайтов // Наука в Сибири. 2010. 30 сент. № 38–39 (2773–2774). URL: http://www.sbras.ru/HBC/hbc.phtml?14+562+1

Материал поступил в редколлегию 15.08.2011

Yu. I. Shokin, O. A. Klimenko, I. S. Petrov

ANALYSIS OF LINKS BETWEEN SITES OF INSTITUTES OF THE SIBERIAN BRANCH OF RAS

The work provides analysis of the structure and dynamics of academic web-space development by the example of the Siberian Branch of the Russian Academy of Sciences (SB RAS). An original program which takes into account some special situations having influence on completeness and quality of the collected data has been created. The results are presented as graphs in which nodes are sites, and edges-hyperlinks. The graphs of interlinked data have shown, that some organizations reference each other, other organizations have many outgoing references, the third are isolated, and the fourth have many incoming references, that show a value of information placed on the site.

Keywords: web-space, webometrics, hyperlink, site.