ИССЛЕДОВАНИЕ СТРОЕНИЯ И ДИНАМИКИ РАЗВИТИЯ НАУЧНОГО ВЕБ-ПРОСТРАНСТВА НА ПРИМЕРЕ СО РАН

О.А. Клименко
e-mail: klimenko@ict.nsc.ru
И.С. Петров
e-mail: coolwarenik@gmail.com
Институт вычислительных технологий СО РАН, Новосибирск

Аннотация

В настоящее время проблема исследования научного веб-пространства является актуальной в связи со стремительным развитием сети интернет и ресурсов представленных в ней. Эти исследования позволяют определить насколько та или иная научная организация следует мировым тенденциям развития и представляет результаты работ на своём сайте. Обычно, чем более развит сайт организации, тем больший вес имеет она в научных кругах.

Целью нашего исследования является определение структуры и динамики развития части сети интернет, связанной с Сибирским отделением РАН.

Одним из подходов к исследованию веб-пространства является использование методов вебометрики, для получения вебометрических индикаторов, таких как количество страниц на сайте, количество входящих ссылок на ресурс, количество «мощных» файлов (doc, ppt, pdf), индекс цитирования, по которым строится рейтинг сайтов. Эти данные берутся из различных поисковых систем. Вебометрика - наука о количественных аспектах создания и использования информационных ресурсов, структур и технологий в веб-пространстве, выросшая из библиометрики и информатики. Термин вебометрика ввели T. Almind и P. Ingwersen в 1997 году. Методами вебометрики были исследованы сайты институтов СО РАН и построен рейтинг.

Другим подходом для получения данных является написание своей программы-крауера, которая, путём обхода и анализа всех страниц на заданном множестве сайтов, выявляет связи между элементами множества. В результате работы программы получается база данных, в которой содержится вся нужная информация о сайтах. В процессе работы было написано несколько версий краулера. Сначала однопоточная версия, потом многопоточная, где одновременно обрабатываются все сайты, это значительно повысило скорость сбора данных. Также была разработана структура хранения данных на основе файлов, для экономии ресурсов системы. В последней версии программы учитываются только уникальные гиперссылки, что позволяет исключить искажение общей картины случаями, когда одна и та же ссылка находится на каждой странице сайта. Программа имеет большую точность работы, это было проверено с помощью данных, взятых из поисковых систем, которые те позволяют получить. Они совпадали с данными полученными программой. Краулер находится в стадии тестирования, но уже сейчас им удобно пользоваться, и его может использовать любой исследователь, который захочет исследовать взаимосвязи тех или иных сайтов.

На основе этой информации стоится граф взаимосвязей. В этом графе G (V, E) узлы соответствуют организациям. Дуга $(u, v) \in E$; $u, v \in V$, если существует страница на сайте организации u, на которой находится гиперссылка на сайт организации v. На множестве дуг

графа G также определено отображение N_E: $E \rightarrow N^+$. Это отображение показывает, сколько существует ссылок между двумя сайтами в соответствующем направлении. В процессе работы были построены графы взаимосвязей

1) всех организаций СО РАН;
2) научных центров СО РАН;
3) внутри Объединенных ученых советов СО РАН по направлениям наук.

Данные графы позволяют сделать множество выводов о той части научной работы, которая отражена на сайтах научных организаций. В частности, можно заметить, 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. Дополнительную информацию можно извлечь, используя поиск по ключевым словам, которые присутствуют в ссылках и заголовках страниц.

Введение

В настоящее время проблема исследования научного веб-пространства является актуальной в связи со стремительным развитием сети интернет и ресурсов представленных в нём. В частности эти исследования позволяют определить насколько та или иная научная организация следует мировым тенденциям развития и представляет результаты научных работ на своём сайте. С уверенностью можно сказать что, чем более развит сайт организации, тем больший вес имеет она в научных кругах. Это объясняется тем, что на неё больше ссылаются в своих работах различные учёные, так как они имеют свободный доступ к исследованиям, проводимым в этой организации.

Целью нашего исследования является как определение структуры и динамики развития научного веб-пространства, как части сети интернет, так и выявление проблем взаимодействия институтов и организаций, на примере СО РАН.

Подходы к исследованию веб-пространства

Термин «вебометрика» (*webometrics*) был введён в работе T. Almind и P. Ingwersen в 1997 году [2] 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»), являющиеся единственным способом взаимодействия между сайтами. Практическая применимость этих исследований успешно демонстрируется реализацией алгоритмов информационного поиска таких популярных систем, как Google и Яндекс [3,4]. 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 [5,6,7]. Для получения больших объёмов информации о гиперссылках можно применить три подхода.

Первый из них заключается в использовании возможностей поисковых машин, таких как Google, Yahoo, Яндекс (эти три системы наиболее полно индексируют русский сегмент интернета). Используя расширенные возможности этих систем, можно получить данные о количестве страниц на сайтах, количестве внешних ссылок, количестве «мощных» фалов

(pdf, doc, ppt). Проблемы связанные с этим подходам известны достаточно давно [8] и основная из них – это отсутствие открытой информации о работе поисковых роботов, а также стоит отметить 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 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. К ним можно отнести проект РФФИ «Вебометрические исследования научных интернет-ресурсов российского Интернета» [9], проводимый ИПМИ КарНЦ РАН, но у них пока нет в свободном доступе их робота для сбора информации и открытой базы данных. А также ресурсы *Statistical cybermetrics research group* из университета Вулверхемптона [10]. На сайте этой исследовательской группы можно найти базы данных свободного доступа. Правда в них информация только по сайтам университетов Великобритании, Австралии и Новой Зеландии. В *Statistical cybermetrics research group* разработан и поддерживается поисковый робот SocSciBot, его можно свободно использовать в научных целях [11]. Из-за того что робот имеет закрытый код, его нельзя настроить для своих нужд.

Третий подход связан с написанием своей программы-крауера, которая путём обхода и анализа всех страниц на заданном множестве сайтов, выявляет связи между элементами множества. В результате работы программы получается база данных, в которой содержится вся нужная информация о сайтах (ссылки, количество страниц на сайте, количество «тяжёлых» файлов, ключевые слова и т.д.). Плюсы этого подхода в том, что можно вносить изменения в программу, для получения нужных данных.

Официальные сайты научных учреждения СО РАН как целевое множество вебометрических исследований

Многие исследователи отмечают одновременное наличие в веб-пространстве как хаоса, так и порядка, при этом если хаос носит разносторонний и всеобъемлющий характер, то признаки порядка проявляются только 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 веб-пространства, таких как уже упоминавшееся ранее множество сайтов университетов Великобритании и других стран [6,9], с расчетом последующего переноса полученных результатов на более общие случаи. Авторами в качестве такого фрагмента веб-пространства было выбрано множество официальных сайтов научных организаций и учреждений СО РАН. Такой выбор имеет достаточно веские основания.

Известно, что Российская академия наук организована по научно-отраслевому и территориальному принципе и включает в себя 9 отделений по областям науки, 3 региональных отделения, 14 региональных научных центра, 20 научных центров региональных отделений и 470 научных учреждений. Если выбрать из этого только Сибирское отделение РАН, то получим 9 научных центров, 9 отделений по областям науки и 109 научных учреждений СО РАН. Таким образом, мы имеем весьма обширную выборку

научных сайтов организаций и учреждений, относящихся к различным областям науки. Официальная политика в сфере информатизации научных учреждений и организаций РАН [12] позволяет сделать выводы об управляемости процессов построения академических сайтов посредством регламентов или технических заданий, что делает это целевое множество потенциально интересным для статистических измерений и исследований. Например, представляется интересной задача обнаружения зависимостей 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. Кроме того, имеется большое количество материалов в ежегодных отчётах о деятельности СО РАН, позволяющих начинать работу в данном направлении уже сейчас.

Стоит отметить и исследования, проводимые зарубежными коллегами по исследованию академических веб-ресурсов, предоставляющие возможность проведения различных сравнительных исследований. При этом стоит отметить что по к понятию «academic Web» у них относятся сайты университетов, а у нас – институтов.

В Сибирском отделении РАН на данный момент есть 89 научных организаций, имеющих действующие сайты с собственными доменными именами.

Построение графа взаимосвязей между сайтами организаций СО РАН

Веб-сайт - это множество веб-страниц, доступных в интернете через протоколы http и https. Страницы сайта объединены общим доменным именем. Каждая страница характеризуется адресом URL. Каждая страница может содержать ссылки на страницы этого или другого сайта.

Граф взаимосвязей строится на основе собранной информации с сайтов научных организаций. В этом графе G (V, E) узлы соответствуют организациям. Дуга $(u, v) \in E$; $u, v \in V$, если существует страница на сайте организации u, на которой находится гиперссылка на сайт организации v. На множестве дуг графа G также определено отображение N_E: $E \rightarrow N^+$. Это отображение показывает, сколько существует ссылок между двумя сайтами в соответствующем направлении [1].

Пример графа взаимосвязей представлен на рис. 1., где:

- ОУС СО РАН по НИТ – Сайт Объединенного ученого совета СО РАН по нанотехнологиям и информационным технологиям (Новосибирск) http://ousnano.sbras.ru;
- ИВТ СО РАН - Институт вычислительных технологий (Новосибирск) www.ict.nsc.ru;
- ИДСТУ СО РАН - Институт динамики систем и теории управления (Иркутск) www.idstu.irk.ru;
- ИВМ СО РАН - Институт вычислительного моделирования (Красноярск) http://icm.krasn.ru;
- ИФП СО РАН - Институт физики полупроводников им. А.В. Ржанова (Новосибирск) http://www.isp.nsc.ru/.

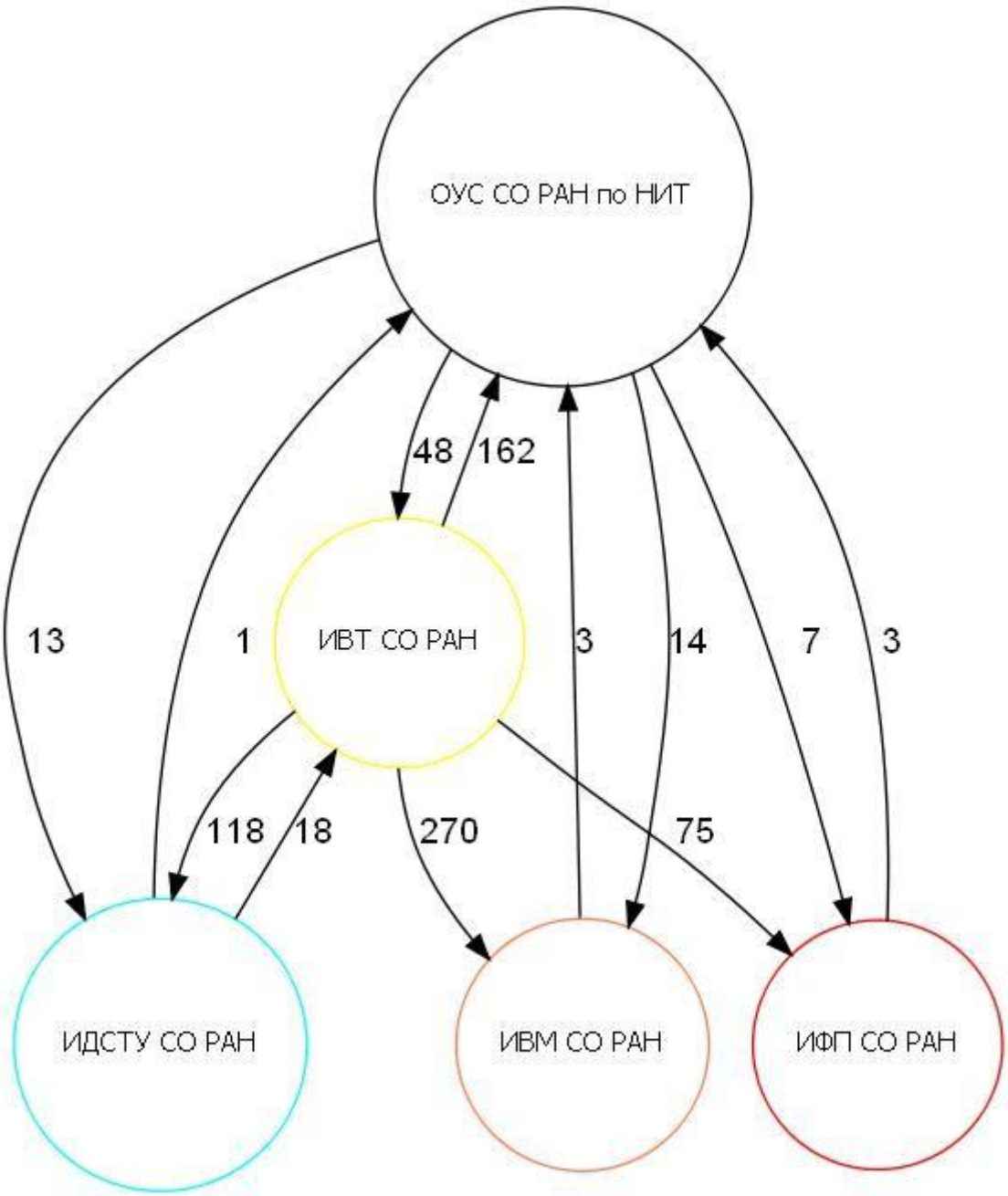

Рис. 1. Граф взаимосвязей сайтов Объединенного ученого совета СО РАН по нанотехнологиям и информационным технологиям.

Алгоритм построения графа

Граф строится с помощью программы-краулера (crawler), которая запрашивает страницы с серверов и анализирует содержимое страниц. 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 $s_i \in S$ определена очередь необработанных страниц Q_i, в которой в начале находится только 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, страницы добавляются в очередь *Qi*.
- 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.

Особенности извлечения данных

Другими исследователями ссылочной структуры Интернет уже создано несколько программ-краулеров, работающих по схожему алгоритму [10], однако для данной задачи было создана оригинальная программа, чтобы учесть ряд особых ситуаций, которые так или иначе 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 в страницы, они содержатся в сценариях Javascript, встроенных в страницу. Такие ситуации нужно обрабатывать индивидуально для 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 в себя:

1. Имя домена
2. Название сайта
3. Исходящие ссылки с сайта
4. Количество «мощных» файлов (doc, pdf, ppt, rtf и т.д.)
5. Ключевые слова, которые содержатся в имени страницы. Метаинформация.
6. Ключевые слова, которые содержатся в анкорах (околоссылочном тексте)
7. Если ключевые слова длинные, то они разбиваются на более мелкие.

Некоторые результаты работы

В процессе работы были построены графы взаимосвязей:

4) Сайтов всех организаций СО РАН;
5) Отдельные графы для каждого из научных центров СО РАН
6) Отдельные графы для каждого из направлений научной деятельности
7) Отдельные графы для интеграционных проектов СО РАН

Данные графы позволяют сделать множество выводов о той части научной работы, которая отражена на сайтах научных организаций. В частности, можно заметить, 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. Дополнительную информацию можно извлечь, используя поиск по ключевым словам, которые присутствуют в ссылках и заголовках страниц.

На графах ясно видно какие сайты хорошо представлены в сети интернет, а какие слабо. Более того, наглядно видно как сайты различных организаций взаимодействуют друг с другом, и на основе этого можно сделать выводы, что в данный момент взаимодействие организаций СО РАН достаточно слабое.

- Было написано несколько версий краулера. Сначала однопоточная версия, потом многопоточная, где одновременно обрабатываются все сайты.
- Была разработана структура хранения данных на основе файлов, для экономии ресурсов системы.
- В последней версии программы учитываются только уникальные гиперссылки.
- Программа имеет большую точность работы.
- Краулер находится в стадии тестирования, но уже сейчас им удобно пользоваться, и его может использовать любой исследователь.

Заключение

К ближайшим задачам можно отнести развитие возможностей программы-краулера и расширения целевого множества за счёт включения сайтов научных организаций не входящих в СО РАН. Так же ведётся работа над более интересным и информативным графическим представлением веб-пространства.

ЛИТЕРАТУРА

[1] Клименко О.А., Петров И.С. Исследование строения и динамики развития научного Веб-пространства на примере СО РАН // Труды XVI Байкальской Всеросийской конференции "Информационные и матетматические технологии в науке и управлении". Часть III. - Иркутск: ИСЭМ СО РАН, 2010. - 92-97с.

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

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Statistical cybermetrics research group. [Электронный ресурс] – 2009. – Режим доступа: http://cybermetrics.wlv.ac.uk.

[11] SocSciBot. [Электронный ресурс] – 2009. – Режим доступа: http://socscibot.wlv.ac.uk.

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