

АНДРЕЙ АНАТОЛЬЕВИЧ ПЕЧНИКОВ

кандидат физико-математических наук, доцент, научный сотрудник Центра ПетрГУ – Метсо систем автоматизации, Петрозаводский государственный университет
pechnikov@krc.karelia.ru

АЛЕКСАНДР ВЛАДИМИРОВИЧ ЧИРКОВ

старший инженер-программист лаборатории телекоммуникационных систем Института прикладных математических исследований, Карельский научный центр РАН
avchirkov@krc.karelia.ru

ЮЛИЯ ВАСИЛЬЕВНА ЧУЙКО

кандидат физико-математических наук, старший научный сотрудник лаборатории информационных компьютерных технологий Института прикладных математических исследований, Карельский научный центр РАН
julia@krc.karelia.ru

ИССЛЕДОВАНИЕ СВЯЗНОСТИ НАУЧНО-ОБРАЗОВАТЕЛЬНОГО ВЕБА

Статья посвящена исследованию связности академического и университетского веб-подмножеств путем определения и анализа структуры связей внутри и между веб-подмножествами, оценки степени связанности подмножеств и взаимного влияния подмножеств на связанность.

Ключевые слова: вебометрика, веб-граф, компонента сильной связности

Рабочая группа по вебометрике Института прикладных математических исследований КарНЦ РАН [1] в течение нескольких лет проводит исследования Веба, основанные на анализе гиперссылок. В 2008–2010 годах исследования были поддержаны Российским фондом фундаментальных исследований (проект № 08-07-00023а).

Объектами исследования в данной работе являются определенные экспертным образом целевые множества, состоящие из официальных сайтов учреждений и организаций РАН и сайтов классических университетов РФ. Множества сайтов и соединяющих их гиперссылок представляются для исследования в виде ориентированного веб-графа, где вершины соответствуют сайтам, а дуги – уникальным гиперссылкам между ними. Уникальность гиперссылок обеспечивается путем объединения каждой группы кратных дуг в одну дугу. Кроме того, в графе нет петель, так как рассматриваются только внешние ссылки, то есть позволяющие переходить именно с одного сайта на другой, а не по страницам одного и того же сайта. Представление множества сайтов с гиперссылками в виде веб-графа сразу дает возможность визуальной оценки степени его связности на основании количества и структуры компонент слабой связности. Каждую компоненту слабой связности в веб-графе составляют сайты, соответствующие тем вершинам в графе, полученном из веб-графа заменой всех дуг на ребра, между любой парой которых существует маршрут. Компоненты сильной связности в таком графе представляют собой множества сайтов, связанных гиперссылками так, что по ним можно перейти с любого сайта на любой сайт данного множества. Вхождение сайта в компоненту силь-

ной связности достаточно большой мощности важно для него, так как повышает вероятность его посещения пользователями. Причем это касается не только так называемых интернет-серферов, которые просматривают веб-ресурсы, переходя между сайтами по попадающимся на веб-страницах гиперссылкам, поисковые роботы также скорее обнаружат и проиндексируют веб-ресурс, на который есть ссылка с уже известных им ресурсов.

В данной работе исследуются такие задачи, как определение структуры связей внутри и между веб-подмножествами, оценка степени связанности подмножеств, взаимное влияние подмножеств на связанность, позволяющие установить, насколько сильно связаны (или разделены) между собой исследуемые веб-сообщества. В свою очередь, это позволяет выдвинуть предположение о том, насколько развито интернет-сотрудничество между владельцами сайтов и отражает ли оно реальную картину сотрудничества между научными и образовательными учреждениями.

В ходе исследования решаются основные подзадачи: сбор исходной информации о веб-множествах, визуализация графов веб-подмножеств, выделение в них компонент сильной связности и определение их мощностей и диаметров, определение изменения структуры компонент связности при объединении подмножеств. Для этого экспертным образом формируются целевые множества сайтов. Путем сканирования веб-ресурсов, входящих в целевые множества, по протоколу http собирается информация о гиперссылках между их узлами. Далее находятся компоненты сильной связности для множеств и их объединений, а также их диаметры. Результаты

графически представляются в наглядном виде для визуальной оценки. Затем делаются выводы о степени связанности подмножеств сайтов между собой и их взаимного влияния.

Для автоматизации проведения исследований рабочей группой по вебометрике был разработан и постоянно развивается ряд инструментов. Для сбора, хранения и обработки информации о гиперссылках создана информационная система для вебометрических исследований (ИС4ВИ), работающая под управлением веб-сервера Apache с интегрированным модулем PHP и СУБД MySQL [5], [6]. Данная система позволяет автоматически сканировать страницы веб-сайта, начиная с указанной страницы первого уровня и переходя дальше на остальные страницы сайта. При этом система собирает со страниц гиперссылки двух видов – внутренние и внешние. Внутренние гиперссылки, позволяющие переходить между страницами одного и того же сайта, используются для составления очереди страниц сайта, подлежащих сканированию. Внешние гиперссылки на другие сайты собираются для дальнейшего исследования связности сообществ сайтов. Необходимо отметить, что сканирование выполняется только для ресурсов, разрешающих это делать. Если в силу технических проблем или административных запретов политики безопасности на стороне сканируемого ресурса системе не удастся последовательно получить содержимое запрашиваемых страниц, то дальнейшие попытки его сканирования прекращаются и ресурс помечается как неотсканированный.

Для автоматизации проведения исследований по анализу связности веб-подмножеств и визуализации результатов было разработано приложение Graph на языке Java 2 SDK, позволяющее в интерактивном режиме выбирать структуру веб-графа из множеств сайтов, находить и отображать компоненты связности и их характеристики. Процедура нахождения компонент связности основана на модификации алгоритма Флойда [2] поиска путей в ориентированном графе. Для визуализации представления веб-графов использована Java-библиотека Jung2 2.0 (<http://jung.sourceforge.net>).

Для исследования были сформированы целевые множества: V, состоящее из 349 официальных сайтов учреждений и организаций РАН, и U, содержащее 56 сайтов классических университетов РФ. Сканирование академического и университетского Веба позволило сформировать базу данных, содержащую информацию о 85 000 уникальных гиперссылок, сделанных с официальных академических сайтов, и о 40 000 уникальных гиперссылок, сделанных с официальных сайтов классических университетов. Информация об исследованиях академического Веба размещена на сайте [1]; некоторые результаты исследований университетского Веба приводятся в [4].

Каждое из исследованных целевых множеств было представлено в виде веб-графа, в котором, как выяснилось, почти все узлы объединены в единственную компоненту слабой связности, за исключением 20 изолированных узлов в веб-графе академического множества и 4 – университетского (рис. 1).

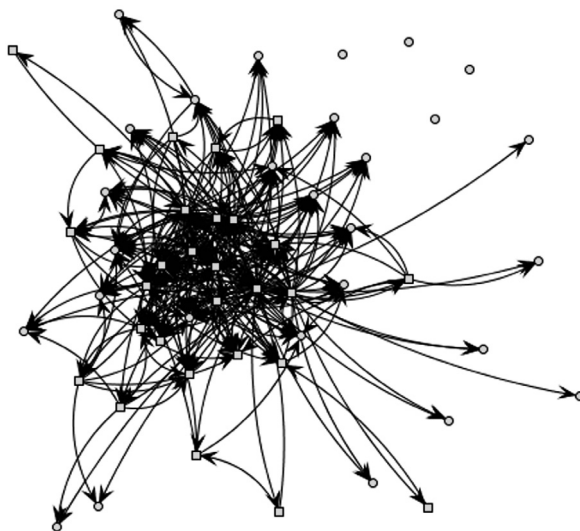


Рис. 1. Веб-граф множества U. Квадратами обозначены узлы, входящие в компоненту сильной связности множества U

Для каждого из целевых множеств были построены компоненты сильной связности. В веб-графе целевого множества V, состоящего из ресурсов академического Веба, была выделена единственная нетривиальная компонента сильной связности, объединяющая 180 узлов и имеющая диаметр 8. Все остальные компоненты тривиальны, то есть каждая состоит из единственного узла. Во множестве U также имеется единственная нетривиальная компонента сильной связности диаметром 5, объединяющая 28 узлов.

В результате можно наблюдать, что структура связанности веб-подмножеств аналогична для академического и университетского Веба:

- в обоих случаях в целевом множестве выделяется единственная нетривиальная компонента достаточно большой мощности и сильной связности;
- примерно совпадают характеристики компонент слабой и сильной связности для целевых множеств относительно мощностей этих множеств.

Однако важно заметить, что для академического Веба характерно вхождение в нетривиальную компоненту сильной связности сайтов наиболее значимых организаций (например, www.gas.ru), в то время как в университетском наблюдается их некоторая обособленность (в частности, www.msu.ru, www.spbu.ru не попадают в компоненту сильной связности).

Далее под термином «компонента сильной связности» будет подразумеваться именно нетривиальная компонента сильной связности.

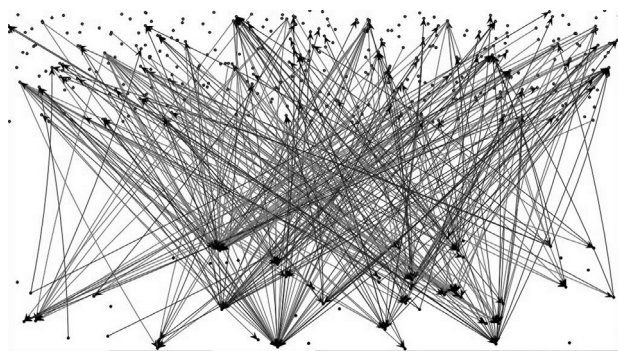


Рис. 2. Взаимные ссылки между подмножествами V и U (вверху – вершины V , внизу – вершины U , стрелки – дуги из множества E)

Следующая задача – объединить два множества с целью исследования структуры взаимной связности научно-образовательного Веба. Рассмотрим ориентированный двудольный веб-граф $G = (V \cup U, E)$, представленный на рис. 2, где V – множество вершин, соответствующих академическим сайтам, U – множество вершин, соответствующих университетским сайтам, E – множество дуг, соответствующих всем уникальным гиперссылкам, сделанным с сайтов из V на сайты из U , и наоборот. Гиперссылки между сайтами внутри множеств U и V здесь не рассматриваются. Веб-граф $g = (V \cup U, E)$, построенный средствами ИС4ВИ и Graph, имеет следующие характеристики: $|V| = 349$, $|U| = 56$, $|E| = 353$, то есть нельзя сказать, что граф сильно насыщен дугами. При этом 262 академических сайта (75 %) не имеют входящих гиперссылок с университетских сайтов, и наоборот, 20 университетских сайтов (36 %) не имеют входящих гиперссылок с академических сайтов. Такие данные являются если не показателем слабого взаимного интереса академических учреждений и классических университетов, то достаточно слабым отражением такого интереса в Вебе.

Аналогичный двудольный граф построен для подмножеств U и V , составляющих их компоненты сильной связности. Он позволяет увидеть, что в данных компонентах присутствуют большие группы узлов, ссылки между которыми связывают между собой обе компоненты сильной связности. Это гарантирует наличие единственной нетривиальной компоненты сильной связ-

ности в объединенном целевом множестве. Причем эта компонента, очевидно, полностью включает обе компоненты сильной связности множеств U и V . Такая компонента сильной связности имеет диаметр, равный 7, и включает 220 узлов – 36 университетских и 184 академических сайтов.

Можно заметить, что при объединении двух целевых множеств связность университетского Веба значительно возрастает (вместо 28 узлов в компоненту сильной связности теперь входят 36), когда в качестве коммуникаторов к нему добавляются академические сайты. Для академических сайтов такое изменение невелико – со 180 на 184.

Компонента сильной связности научно-образовательного Веба содержит все узлы таких компонент академического и университетского подмножеств. Дополнительно она включает в себя множество узлов, не попадавших ранее в компоненты сильной связности подмножеств (например, www.philosophy.nsc.ru, www.ipae.uran.ru, www.ieie.nsc.ru из академических сайтов, www.msu.ru, www.kamgu.ru, www.masu.ru – из университетских). Наиболее значимые организации академического Веба входят в компоненты сильной связности как академического, так и научно-образовательного Веба. Часть значимых организаций университетского Веба, не входивших в его компоненту сильной связности, вошли в компоненту сильной связности научно-образовательного Веба (www.msu.ru), в то время как другая их часть осталась в компоненте слабой связности (www.spbu.ru). Диаметр компоненты сильной связности научно-образовательного Веба меньше диаметра компоненты сильной связности академического Веба, несмотря на меньшую мощность последней.

Проведенные исследования позволяют сделать следующие краткие выводы. Российская академия наук и классические университеты России сильно связаны между собой взаимным сотрудничеством, находящим свое отражение в Вебе. Связность университетского Веба с академическим значительно более важна для университетов, чем для организаций РАН. Высокая степень связности научно-образовательного Веба способствует увеличению посещаемости сайтов обоих целевых подмножеств.

СПИСОК ЛИТЕРАТУРЫ

1. ВЕБОМЕТРИКА. Институт прикладных математических исследований КарНЦ РАН [Электронный ресурс]. Режим доступа: <http://webometrics.krc.karelia.ru>
2. Кристофидес Н. Теория графов: алгоритмический подход: Пер. с англ. М.: Мир, 1978. 429 с.
3. Общий рейтинг выборки классических университетов – Национальный рейтинг российских вузов [Электронный ресурс]. Режим доступа: http://univer-rating.ru/rating_common.asp
4. Печников А. А. О некоторых результатах вебометрических исследований университетского Веба // IV Междунар. науч.-практ. конф. «Современные информационные технологии и ИТ-образование»: Сб. избр. тр. М.: ИНТУИТ.РУ, 2009. С. 530–537.
5. Печников А. А., Луговая Н. Б., Чуйко Ю. В., Косинец И. Э. Разработка инструментов для вебометрических исследований гиперссылок научных сайтов // Вычислительные технологии. 2009. Т. 14. № 5. С. 66–78.
6. Программа для ЭВМ «Информационная система для вебометрических исследований». Свидетельство о государственной регистрации программы для ЭВМ «Информационная система для вебометрических исследований» Федеральной службы по интеллектуальной собственности, патентам и товарным знакам РФ № 2010610941 от 29 января 2010 г.