О СВЯЗНОСТИ И ДИНАМИКЕ ВЕБ-ГРАФА НАУЧНЫХ УЧРЕЖДЕНИЙ

А.А. Печников

Институт прикладных математических исследований Карельского НЦ РАН

Введение

Задачи, связанные с исследованиями связности веб-графов тематических фрагментов Веба, сегодня уже являются достаточно традиционными. На первый план выходят задачи, связанные с динамикой Веба, в данном случае — с изменением связности во времени. Сложность таких исследований связана, в частности, с недостатком исходных данных, которые требуется собрать за достаточно продолжительный отрезок времени. В работе предлагается модель динамики веб-графа научных учреждений, основанная на фиксации веб-графа, построенного на заданный (конечный) момент времени и «возврате назад» посредством удаления кратных гиперссылок. В теоретическом плане данная модель подтверждает такие известные идеи, используемые при моделировании веб-графов, как предпочтительное присоединение и начальная притягательность вершин, в содержательном — подтверждает ряд известных фактов по созданию веб-пространств ряда научных организаций РАН.

О некоторых терминах, определениях и средствах исследования

В статье рассматриваются веб-ресурсы организаций, которые до начала реформы РАН позиционировались как научные учреждения и структурные подразделения Российской академии наук. В соответствии с Распоряжением Правительства Российской Федерации большинство из них переданы в ведение Федерального агентства научных организаций [1], а их уставы находятся в процессе разработки и утверждения, поэтому далее мы будем называть их просто научными учреждениями, подразумевая под этим названием собственно РАН, отделения по областям науки, региональные отделения и научные центры, научные институты — всего 397 единиц исследования.

Уточним несколько понятий, которые потребуются для дальнейшего изложения.

Определение 1. Веб-сайт (сайт) — это совокупность *html*-страниц и веб-документов, связанных внутренними гиперссылками и обладающих единством содержания, идентифицируемый в Вебе по его доменному имени.

Для большинства научных учреждений характерно наличие официальных сайтов. В данной работе рассматриваются 397 официальных сайтов учреждений. В качестве основы использовались данные проекта по вебометрическому ранжированию научных учреждений [2]; список исследуемых сайтов можно увидеть на сайте проекта в разделе «Целевое множество».

Определение 2. Уникальной внешней гиперссылкой называется гиперссылка из множества всех гиперссылок с одинаковым адресом и контекстом, которая находится на странице, имеющей максимальный уровень; при этом уровень начальной страницы сайта считается наивысшим. Далее мы будем рассматривать только уникальные внешние гиперссылки, поэтому слова «уникальная» и «внешняя» в большинстве случаев будем опускать.

В общем случае веб-графом называется ориентированный граф, множество вершин которого соответствует множеству исследуемых страниц в Вебе, а множество дуг — множеству гиперссылок, связывающих эти страницы. Рассматривая все страницы одного веб-сайта как единое целое и соответствующим образом агрегируя гиперссылки, можно получить веб-граф, построенный на некотором множестве сайтов. Используя только уникальные гиперссылки (в смысле определения 2) мы получаем веб-граф на множестве сайтов с уникальными гиперссылками. В этом графе отсутствуют петли, но допустимы кратные дуги, поскольку гиперссылок между сайтами может быть несколько.

Множество вершин веб-графа, соответствующих официальным сайтам научных учреждений, обозначим S. Множество дуг $V(n)$ в нашем случае зависит от параметра n: в веб-графе существуют только те дуги, для которых количество гиперссылок, связывающих соответствующие сайты, не менее n. Таким образом, рассматриваемые далее графы не имеют кратных дуг (и петель).

Например, веб-граф $G(S, V(1))$ – это граф, построенный на множестве официальных сайтов, у которого дуга, соединяющая пару вершин, существует тогда, когда существует хотя бы одна гиперссылка, соединяющая соответствующие сайты, а $G(SB, V(10))$ – это граф, построенный на множестве пучков, у которого дуга, соединяющая пару вершин, существует тогда, когда существует 10 и более гиперссылок, соединяющих соответствующие сайты.

Для нахождения и сбора гиперссылок использовалась специализированная программа BeeCrawler [3], а для анализа веб-графов – открытая программная платформа Gephi [4]. Данные хранятся и обрабатываются в базе данных внешних гиперссылок [5].

Свойства веб-графа на множестве официальных сайтов

По данным, полученным в результате сканирования веб-сайтов научных учреждений, была построена последовательность веб-графов $G(S, V(1))$, $G(S, V(2))$, …, $G(S, V(10))$.

Рассмотрены три следующие характеристики связности веб-графа: количество вершин, имеющих хотя бы одну инцидентную ей гиперссылку (далее – «неизолированные вершины»), общее количество дуг и количество вершин в максимальной компоненте сильной связности (КСС). Понятно, что уменьшение параметра n, означающее смягчение требований к количеству гиперссылок, связывающих сайты, ведёт к возрастанию значений этих характеристик. Для последовательности веб-графов, построенных на множестве официальных сайтов S, изменения характеристик в зависимости от n приведены в виде графиков на рисунке 1. Значения n по оси абсцисс взяты в обратном порядке, поскольку такой порядок будет использоваться в дальнейшем.

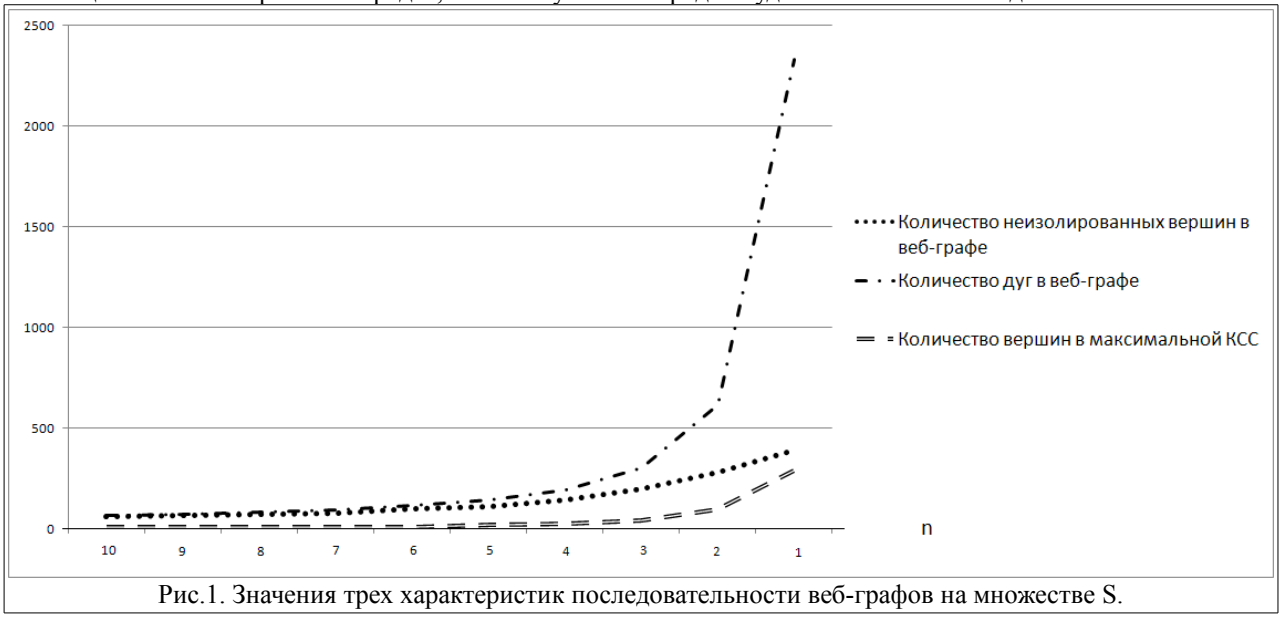

Рис.1. Значения трех характеристик последовательности веб-графов на множестве S.

Из графиков видно, что чем более мягким является требование к n, тем большие значения принимают все характеристики, однако порядок роста различен; экспоненциальный рост демонстрирует количество дуг в веб-графе, количество неизолированных вершин растет как полином второй степени, а количество вершин в максимальной КСС как полином третьей степени. При этом понятно, что количество неизолированных вершин и вершин в максимальной КСС ограничены сверху $|S|$, т. е. при n=1 близки к пределу.

Максимально возможное количество дуг в веб-графе равно $|S|*(|S|-1)$, для нашего случая это более 157000 дуг, то есть до насыщения графа дугами ещё очень далеко.

Представляется интересным рассмотреть последовательность из 10 веб-графов, построенных на множестве официальных сайтов, как модель динамики веб-графа научных учреждений.

Пусть $t(1)$, $t(2)$,...,$t(10)$ – дискретные моменты времени; $t(i)<t(i+1)$.

Будем считать, что граф $G(S, V(10))$ отражает состояние веб-пространства научных учреждений в начальный момент $t(1)$, граф $G(S, V(9))$ – в момент $t(2)$, и т. д. Из рисунка 1 понятно, что количество дуг в графе $G(S, V(9))$ больше, чем в графе $G(S, V(10))$, и этот прирост мы можем объяснить тем, что за промежуток времени, прошедший от $t(1)$ до $t(2)$ появились новые гиперссылки между сайтами. Такая интерпретация позволяет рассматривать веб-граф научных учреждений в его развитии, связанном с ростом количества дуг. На рисунке 2 показано, как ведут себя его компоненты сильной связности, неизолированные и изолированные вершины с изменением времени.

Остановимся подробнее на том, что на этом рисунке показано о состоянии веб-графа в момент $t(1)$. В условной колонке под $t(1)$ изображены четыре непересекающихся подмножества вершин *Nsol* (неизолированные вершины), *A* (максимальная КСС), *B* (еще одна невырожденная КСС) и *Sol* (изолированные вершины), объединение которых равно *S*. Под обозначением каждого множества стоит число, равное мощности этого множества. На 397 вершин веб-графа имеется всего-навсего 69 дуг.

Максимальная КСС веб-графа *A* в момент $t(1)$ состоит из следующих вершин (мы будем их именовать по названиям учреждений, так проще для восприятия, а в скобках писать доменные имена): Институт экономики КарНЦ РАН (*economy.krc.karelia.ru*), Институт леса КарНЦ РАН (*forestry.krc.karelia.ru*), Институт биологии КарНЦ РАН (i*b.krc.karelia.ru*), Институт геологии КарНЦ РАН (*ig.krc.karelia.ru*), Институт языка, литературы и истории КарНЦ РАН (*illhportal.krc.karelia.ru*), Институт прикладных математических исследований КарНЦ РАН (*mathem.krc.karelia.ru*), Институт водных проблем Севера КарНЦ РАН (*water.krc.karelia.ru*), Карельский научный центр РАН (*www.krc.karelia.ru*).

Вторая КСС *B* содержит две вершины, которым соответствуют Институт вычислительных технологий СО РАН (*www.ict.nsc.ru*) и Сибирское отделение РАН (*www-sbras.nsc.ru*).

Состояние веб-графа в момент *t*(*2*) на рисунке не отражено, поскольку за первый прошедший отрезок времени множество дуг возросло на 8, что не изменило компонент связности , хотя и увеличило подмножество вершин *Nsol* за счет уменьшения *Sol*. К моменту *t*(*3*) появилась еще одна КСС *C* с двумя вершинами: Геологический институт КолНЦ РАН (*www.kolasc.net.ru*) и Кольский научный центр РАН (*geoksc.apatity.ru*). К моменту *t*(*4*) изменений почти не произошло (он отсутствует на рисунке), а к моменту *t*(*5*) на две вершины возросла КСС *B*, добавленное подмножество обозначено *D*: Институт цитологии и генетики СО РАН (*www.bionet.nsc.ru*) и Институт химической биологии и фундаментальной медицины СО РАН (*www.niboch.nsc.ru*). Возникла еще одна КСС *E*: Российская академия наук (*www.ras.ru*), Институт научной информации по общественным наукам РАН (*www.inion.ru*), Уральское отделение РАН (*www.uran.ru*). Подмножество *Nsol* за это время потихоньку увеличивалось, а *Sol* убывало.

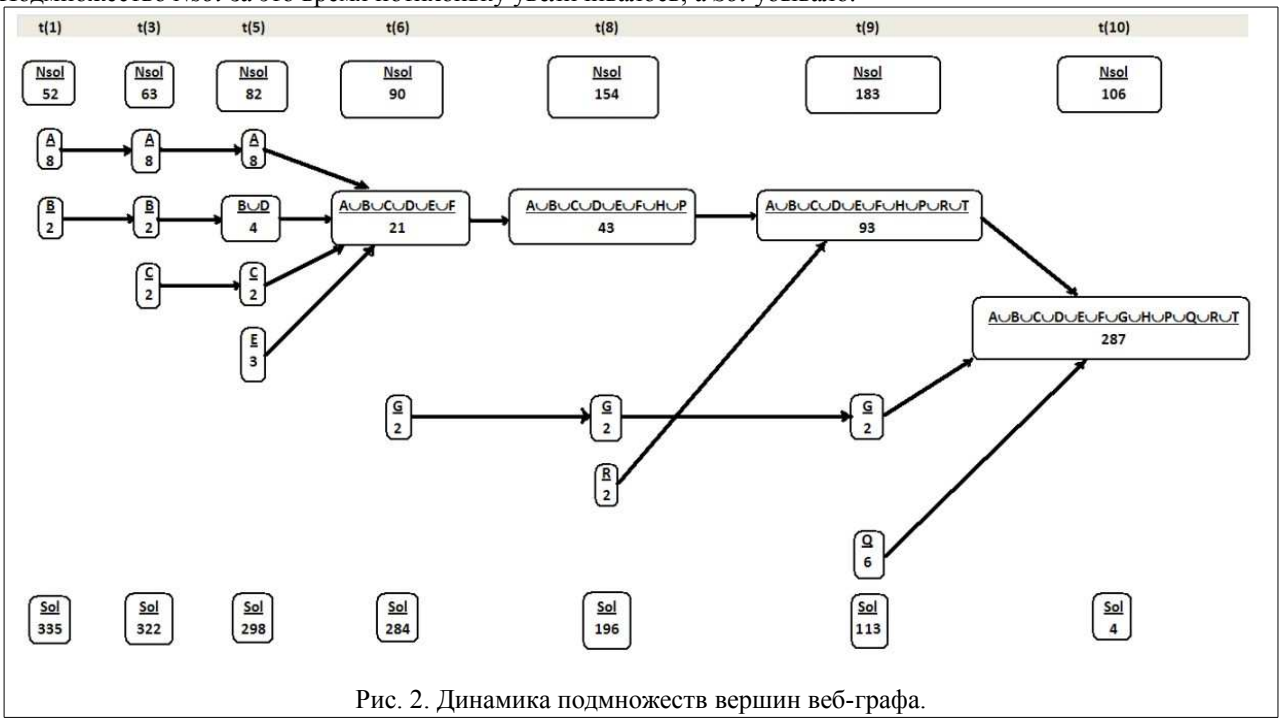

Рис. 2. Динамика подмножеств вершин веб-графа.

К моменту *t*(*6*) количество дуг в графе увеличилось уже до 113 (почти в 2 раза с начальным состоянием). При этом произошли существенные изменения в структуре подмножеств. Все КСС, существовавшие раздельно в предыдущий момент времени, объединились в одну, содержащую 21 вершину, с добавлением подмножества *F* из 4 вершин. В подмножество *F* входят Институт проблем химической физики РАН (*www.icp.ac.ru*), Институт ядерной физики им. Г.И. Будкера СО РАН (*www.inp.nsk.su*), Институт проблем развития науки РАН (*www.issras.ru*) и Институт вычислительной математики и математической геофизики СО РАН (*www.sscc.ru*). Возникла и еще одна маленькая КСС *G*: Нижегородский научный центр РАН (*www.nncras.ru*) и Институт прикладной физики РАН (*www.iapras.ru*).

Дальнейшее описание потребовало бы перечисления большого количества официальных сайтов учреждений и будет опущено. Главное, что мы уловили основные тенденции изменения модели веб-пространства официальных сайтов научных учреждений: за очередной временной шаг происходит экспоненциальный прирост количества дуг в веб-графе, что ведет к росту максимальной КСС путем поглощения меньших КСС, и возникновению новых КСС. Таким образом, к моменту *t(10)* сформировалась единственная КСС, поглотившая более 70% вершин, практически выродилось подмножество изолированных вершин, а оставшиеся неизолированные вершины инцидентны вершинам из КСС. Количество дуг в веб-графе достигло 2331.

Заключение

Заметим, что в начальной точке предложенная модель содержит две КСС, первая из которых состоит из сайтов учреждений Карельского научного центра РАН, а вторая – Сибирского отделения РАН. Затем к ним присоединяется КСС, содержащая сайты Кольского научного центра РАН. Далее немного подросла сибирская КСС, и появилась новая группа, (внимание!) включающая сайт РАН. После этого начался стремительный рост максимальной КСС.

В теоретическом плане ход процесса подтверждает такие идеи, используемые при моделировании веб-графов, как предпочтительное присоединение и начальная притягательность вершин. Основные модели веб-графов описаны, например, в [6]. В соответствии с моделью предпочтительного присоединения в каждый момент времени появляется новый сайт, и этот сайт ставит фиксированное количество ссылок на своих предшественников; причем вероятность, с которой новый сайт поставит ссылку на один из прежних сайтов,

пропорциональна числу уже имевшихся на тот сайт ссылок. Начальная притягательность может пониматься как некоторая предпочтительность для нового сайта в выставлении ссылок на уже имеющиеся сайты; например, скорее появится ссылка на институт, работающий в той же научной области, в которой работает учреждение-владелец нового сайта или входящий в состав того же научного центра, нежели на сайт института из другой области.

Этим же объясняется и тот факт, что количество дуг растет экспоненциально, а количество связных вершин – полиномиально.

В содержательном плане наличие «карельской» КСС на ранних этапах объясняется тем, что большинство веб-ресурсов (включая официальные сайты институтов) здесь создается одной группой веб-разработчиков из Института прикладных математических исследований начиная с 1997 года. Развитие «сибирской» КСС можно обосновать большой работой по программам развития телекоммуникационных и информационных ресурсов, координируемой Институтом вычислительных технологий СО РАН, в рамках базовых программ фундаментальных исследований СО РАН.

Реализация целевых программ по информатизации научных учреждений и Президиума РАН привела к появлению на сайте РАН такого раздела, как «Информационные системы научных учреждений РАН» (*http://www.ras.ru/sciencestructure/informationsystems.aspx*), содержащего ссылки на научные учреждения РАН, что и явилось мощнейшим толчком к усилению связности веб-пространства научных учреждений.

Немного жалко, что у нас нет накопленной базы данных по гиперссылкам научного веб-пространства по временным срезам. Можно было бы попробовать «привязать» абстрактные моменты времени $t(1)$, $t(2)$,...,$t(10)$ к реальным годам и спрогнозировать динамику веб-графа на ближайшие годы. Правда, можно смело утверждать, что вскоре исчезнут изолированные вершины, а затем все сайты из S войдут в КСС.

Возвращаясь к Распоряжению Правительства [1] можно заметить, что множество S следует увеличить по крайней мере в два раза; а значит нас ждёт много работы.

Работа выполнена при поддержке Программы стратегического развития Петрозаводского государственного университета на 2012–2016 годы и гранта РГНФ № 12-03-12001.

ЛИТЕРАТУРА:

1. Распоряжение Правительства Российской Федерации от 30 декабря 2013 года №2591-р. http://www.rg.ru/2014/01/09/fano-site-dok.html (дата обращения 29.04.2014).
2. Вебометрический рейтинг научных учреждений России. http://webometrics-net.ru (дата обращения 20.04.2014).
3. А.А. Печников, Д.И. Чернобровкин Адаптивный краулер для поиска и сбора внешних гиперссылок // Управление большими системами. Выпуск 36. М.: ИПУ РАН, 2012. С.301-315.
4. Gephi, an open source graph visualization and manipulation software. https://gephi.org (дата обращения 22.04.2014).
5. А.С. Головин, А.А. Печников База данных внешних гиперссылок для исследования фрагментов Веба // Информационная среда вуза XXI века: материалы VII Всероссийской научно-практической конференции (23-27 сентября 2013 г.). Петрозаводск, 2013. С. 55-57.
6. А.М. Райгородский Модели случайных графов и их применения // Труды Московского физико-технического института. 2010. Т. 2. № 4. С. 130-140.