УДК 004.738.5

О СХОДСТВЕ СТРУКТУР ВЕБ-ПРОСТРАНСТВ С ОДИНАКОВОЙ ТЕМАТИКОЙ

[1]Печников А.А., [2]Павлов А.Г.

[1]ФБГУН «Институт прикладных математических исследований Карельского научного центра Российской академии наук», Петрозаводск, e-mail: pechnikov@krc.karelia.ru;
[2]ФГБОУ ВПО «Санкт-Петербургский государственный университет», Санкт-Петербург, e-mail: pavlolvandrei@gmail.com

В статье описан подход к кластеризации веб-пространств крупных организаций по задаваемым формальным характеристикам. Для исследования веб-пространства организации в качестве его математической модели строится веб-граф, вершинами которого являются веб-сайты, составляющие веб-пространство организации, а дугами – гиперссылки, связывающие эти сайты. Элементы веб-графа получены с помощью реализованной программы-краулера. На примере Санкт-Петербургского государственного университета показаны результаты работы программы-краулера и построение веб-графа веб-пространства вуза. Экспериментальный анализ проведён для пяти вузов России, пяти научных учреждений и пяти производственных предприятий. С помощью краулера собраны все исходные данные, далее вычислен ряд основных характеристик для каждого веб-пространства и по ним вычислены вторичные характеристики, соответствующие важности головного сайта, плотности и связности веб-графа, количеству «висячих» вершин. Результаты кластеризации по этим характеристикам позволяют сформировать четыре кластера экспериментального множества, три из которых практически соответствуют тематике входящих в них веб-пространств крупных организаций. Проведенное исследование показывает перспективность продолжения работы на пути решения задач классификации веб-пространств в зависимости от их внутренней организации (самоорганизации) и тематики.

Ключевые слова: гиперссылка, веб-сайт, веб-пространство, веб-граф, кластерный анализ

ON THE SIMILARITY OF THE STRUCTURES OF WEB SPACES WITH IDENTICAL THEMES

[1]Pechnikov A.A., [2]Pavlov A.G.

[1]Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Petrozavodsk, e-mail: pechnikov@krc.karelia.ru;
[2]Saint-Petersburg State University, Saint-Petersburg, e-mail: pavlolvandrei@gmail.com

The article describes an approach to clustering web spaces of large organizations according to a set of given formal characteristics. In order to explore the organization's web space, a web graph is constructed as its mathematical model, which's vertices are the web-sites, and edges are the hyperlinks linking these sites. The elements of the web graph are obtained with the help of the implemented program-crawler. The example of the St. Petersburg State University shows the results of the work of the crawler and the construction of a web graph of the university's web space. Experimental analysis was carried out for five universities in Russia, five scientific institutions and five manufacturing enterprises. With the help of the crawler all the initial data is collected, then a number of basic characteristics for each web space is computed, according to which the secondary characteristics corresponding to the importance of the head site, the density and connectivity of the web graph, and the number of «hanging» vertices are calculated. The results of clustering on these characteristics make it possible to form four clusters of the experimental set, three of which practically correspond to the topic of the web spaces of large organizations included in them. The conducted research shows the prospects of continuing work on the way of solving the problems of classification of web spaces depending on their internal organization (self-organization) and subject matter.

Keywords: hyperlink, website, web space, web graph, cluster analysis

Исследование веб-пространства организаций является актуальной проблемой в связи со стремительным развитием Веба и ресурсов, представленных в нем. Сайты крупных организаций, таких как Санкт-Петербургский госуниверситет или Газпром, имеют десятки и сотни сайтов и тысячи связывающих их гиперссылок. Эти исследования помогают определить, насколько организация следит за тенденцией развития своих сайтов и предоставляет результаты своей деятельности.

Веб-сайт – это совокупность html-страниц и веб-документов, связанных внутренними гиперссылками и обладающих единством содержания, идентифицируемая в Вебе по уникальному доменному имени [1]. Веб-пространство организации – это множество, состоящее из веб-сайтов организации, которые связаны между собой гиперссылками. У веб-пространства всегда можно выделить его «головной сайт», официальный сайт организации. Внутренние гиперссылки – это гиперссылки, которые ссылаются на html-страницы заданного веб-пространства, при этом URL-источник также является страницей этого веб-пространства.

Для описания веб-пространства можно использовать веб-граф. В общем случае веб-граф – это ориентированный граф, вершинами которого являются html-страницы, а ребрами – гиперссылки, связывающие данные вершины. Чтобы построить веб-граф сайта, необходимо получить сведения о его структуре: html-страницы и гиперссылки, связывающие их. Краулер – программа, предназначенная для перебора страниц Веба с целью сбора и/или занесения определённой информации в базу данных [2].

Структурные исследования характеристик веб-графов в настоящее время достаточно хорошо исследованная область прикладной математики [3]. Компоненты сильной связности, клики, значения Page Rank и другие характеристики позволяют лучше понять развитие и функционирование как веб-пространств организаций, так и взаимодействие между ними.

Основной вопрос данной статьи ставится так: можно ли сказать, что одинаковые по тематике сайты имеют подобную (в некотором заданном смысле) структуру?

В нашем случае ответ на этот вопрос формулируется на основе проведенных экспериментов для 15 крупных организаций (по 5 вузов, научных институтов и производственных предприятий).

При этом необходимо было решить несколько подзадач:

1. Разработать программу-краулер для сбора информации о веб-пространстве организации.

2. Определить основные характеристики веб-графа, построенного по данным, полученным краулером (PageRank, клики, компоненты связности).

3. Исследовать вопрос о кластеризации множества веб-пространств по ряду формальных характеристик их веб-графов.

Эксперименты, проведенные на примере 15 крупных организаций с определением ряда формальных характеристик, используемых в разбиении данного множества веб-пространств на подмножества с близкими тематиками и структурами, дают хорошие результаты и позволяют сделать вывод о перспективности данного направления исследований.

Краулер

Для сбора информации о веб-пространстве организации была реализована программа-краулер, основной задачей которой является сбор доменных имен веб-сайтов и гиперссылок, связывающих их. Теме краулеров посвящено много работ [4], однако в открытом доступе не удалось найти подходящий краулер, который бы решал поставленную задачу без дополнительных затрат на обработку входных/выходных данных и ввода дополнительных параметров. Поэтому было решено реализовать свой краулер, удовлетворяющий таким требованиям, как простота в использовании, скорость обработки сайтов заданного веб-пространства, посещение только веб-сайтов, доменное имя которых является поддоменом домена головного сайта, индексирование гиперссылок, у которых домен URL адреса является поддоменом домена головного сайта.

Архитектура реализованного краулера содержит в себе блок краулинга (при запросе URL страницы получает ответ от веб-сервера, если доступ к странице получен, делает синтаксический анализ), блок сканирования (собирает все внутренние гиперссылки со страницы) и блок записи (обновляет список с доменными именами веб-сайтов и список гиперссылок).

Ниже описаны основные свойства реализованного краулера:

1. В качестве исходных данных подаётся адрес начальной страницы головного сайта исследуемого веб-пространства организации и максимальная глубина сканирования каждого сайта веб-пространства. Уровень веб-страницы определяется так: начальная страница, определяемая по доменному имени сайта, имеет уровень 0. Уровень любой другой страницы – это минимальное количество внутренних гиперссылок, ведущих от начальной страницы к данной.

2. Обход каждого сайта, начиная с главной заданной страницы, осуществляется «в ширину» по внутренним гиперссылкам.

3. Объекты сканирования – html-страницы. Гиперссылки, указывающие на файлы с расширениями rar, docx, 7z и тому подобное, и гиперссылки типа «mailto:» не рассматриваются.

4. Гиперссылки извлекаются с html-страниц, из тегов <a> параметра <href>, доменное имя которых является поддоменом любого уровня доменного имени главной страницы.

5. Для гиперссылки сервер должен выдавать ответ с кодом состояния HTTP равным 200 (OK – запрос успешен) [5].

6. Сканирование осуществляется до тех пор, пока не будет достигнута заданная глубина сканирования, либо список страниц, которые необходимо посетить, будет пуст.

7. В качестве результата выдаётся два файла: список всех найденных сайтов, доменное имя которых является поддоменом любого уровня доменного имени главной страницы и официальное название сайта;

список всех полученных гиперссылок, связывающих сайты из первого файла.

Краулер реализован на языке Java в интегрированной среде Intellij Idea [6], для синтаксического анализа страниц была использована библиотека Jsoup [7].

Примечательным является высокая эффективность программы (не осуществляется индексирование веб-сайтов и гиперссылок, которые не принадлежат веб-пространству исследуемой организации). Например, время работы программы для полной обработки веб-пространства СПбГУ примерно 3 часа 20 минут. Посещено 24590 страниц, найден 151 веб-сайт и 99930 связывающих их гиперссылок.

Веб-граф организации и его основные характеристики

Веб-граф – это множество $G(V, E)$, состоящее из html-страниц и/или документов, являющихся вершинами V веб-графа G, и гиперссылок E, связывающих элементы из множества V. Рассмотрим построение веб-графа организации на примере СПбГУ.

При помощи реализованной программы-краулера были получены списки вершин и дуг веб-графа.

Ниже, в табл. 1 и 2, представлены некоторые данные, полученные краулером.

Далее была сформирована табл. 3 по данным полученным краулером, а именно – для каждой пары из табл. 1 было подсчитано количество дуг, исходящих из одной вершины в другую.

Представим на рисунке визуализацию веб-графа, для этого была использована библиотека Jgraph [8], простая в использовании, с помощью которой можно построить и вывести на экран нужный граф.

Наибольшее количество исходящих или входящих гиперссылок имеют официальный сайт СПбГУ, его английская и китайская версии, сайт виртуальной приемной комиссии СПбГУ и сайт архива открытого доступа СПбГУ.

Также хорошие (в смысле инцидентности дуг) показатели имеют несколько веб-сайтов факультетов СПбГУ (факультет психологии, юридический факультет), веб-сайт научной деятельности СПбГУ, веб-сайт студенческого совета СПбГУ, веб-сайт научного парка СПбГУ.

Для дальнейшего анализа было определено несколько характеристик веб-графа, таких как количество вершин, количество дуг, максимальная клика (размерность), количество клик размерности 3 и более и компонента сильной связности [9, 10].

Таблица 1

Некоторые веб-сайты веб-пространства СПбГУ

Доменное имя сайта	Официальное название сайта
spbu.ru	СПбГУ
chinese.spbu.ru	SPBU-圣彼得堡国立大学
dspace.spbu.ru	DSpace at Saint Petersburg State University

Таблица 2

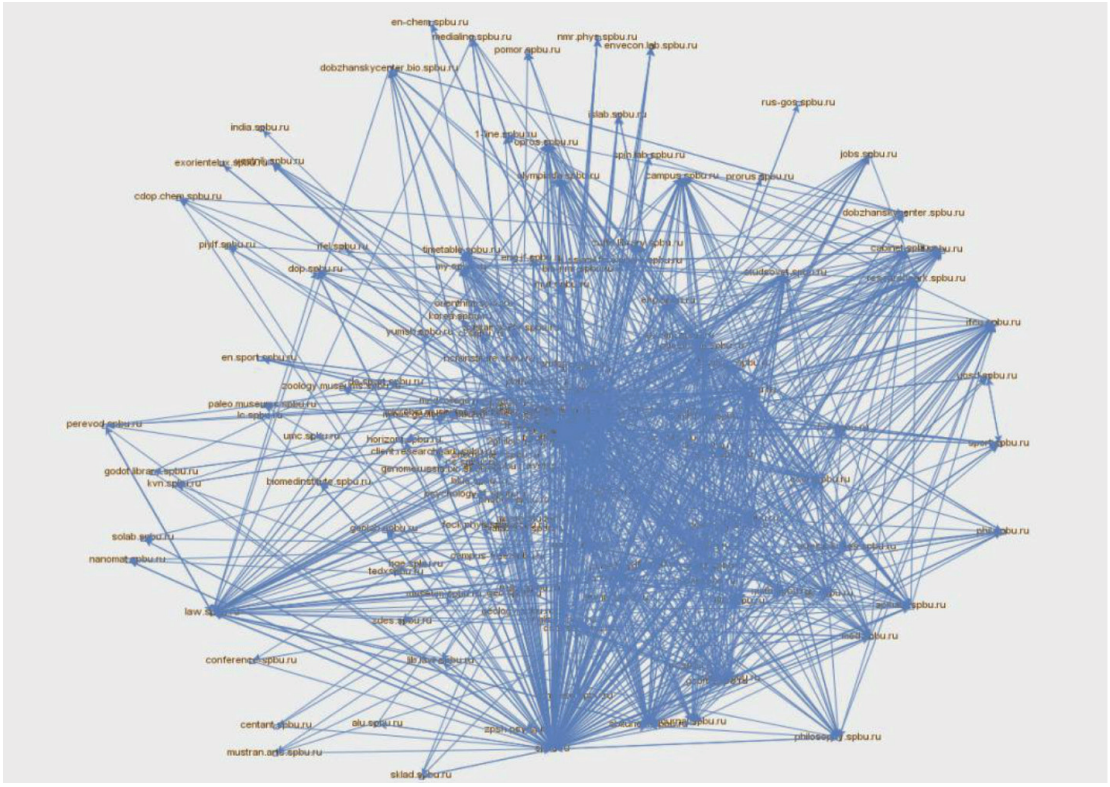
Некоторые дуги веб-пространства СПбГУ

URL-источник	URL-приемник
http://spbu.ru	http://chinese.spbu.ru
http://spbu.ru	https://dspace.spbu.ru
http://nauka.spbu.ru/megagrany-spbgu	https://ias.spbu.ru

Таблица 3

Представление веб-графа СПбГУ в виде списка дуг

Доменное имя источника	Доменное имя приемника	Количество дуг
guestbook.spbu.ru	spbu.ru	11664
spbu.ru	english.spbu.ru	7767
dspace.spbu.ru	spbu.ru	4432
dspace.spbu.ru	it.spbu.ru	4384
spbu.ru	chinese.spbu.ru	3883
nauka.spbu.ru	spbu.ru	2244
psy.spbu.ru	spbu.ru	2065



Визуализация веб-графа СПбГУ

Для веб-пространства СПбГУ (доменное имя головного сайта spbu.ru) был получен следующий ряд характеристик: количество вершин – 151, количество дуг – 99930, размерность максимальной клики – 6, количество клик размерности 3 и более – 75, размерность максимальной компоненты сильной связности – 123.

Для каждой вершины веб-графа было вычислено значение PageRank [11], в частности первые 5 вершин веб-пространства СПбГУ с наибольшими значениями имеют следующие показатели: spbu.ru – 0,1526; eng.spbu.ru – 0,0266; it.spbu.ru – 0,0251; abiturient.spbu.ru – 0,0238; guestbook.spbu.ru – 0,0204.

Сравнительное исследование веб-графов организаций

Для проведения экспериментов были взяты веб-пространства 15 организаций, информация о которых сведена в табл. 4. Первые пять организаций представляют российские вузы, следующие пять – производственные организации, и последние пять – научные учреждения России. В последних трех колонках в качестве примера приведены по три характеристики каждого веб-пространства.

В табл. 5 приводятся вторичные характеристики, полученные из основных, и используемые далее для разбиения исследуемого множества на непересекающиеся подмножества, состоящие из схожих объектов.

В колонке PR0/PR1 вторичная характеристика вычисляется как частное, полученное от деления PR головного сайта на PR наибольшего значения из всех остальных значений.

Кластеризация по методу k-средних при задаваемом разбиении на 4 кластера [12] дает следующие четыре подмножества: cl1: {ict.nsc, iki.rssi, ras}, cl2: {spbu, msu, mipt, petrsu, severstal, kunstkamera}, cl3: {urfu, krc.karelia}, cl4: {gazprom, rosneft, baltika, evraz}.

Кластер cl1 содержит только научные учреждения России, cl2 – в основном российские вузы, а cl4 – производственные организации. В cl3 оказался вуз и научное учреждение. Можно сказать, что введенные вторичные формальные характеристики веб-графов дают довольно хороший результат в смысле разбиения заданного множества веб-пространств на подмножества почти одной тематики.

В табл. 6 приводятся средние значения вторичных характеристик для каждого из кластеров.

Таблица 4

Сведения об исследуемых организациях

№ п/п	Организация	Условное обозначение	URL головного сайта	Кол-во вершин	Кол-во дуг	PR головного сайта
1	СПбГУ	spbu	spbu.ru	151	99930	0,0148
2	МГУ	msu	www.msu.ru	291	80154	0,0161
3	МФТИ	mipt	mipt.ru	85	26106	0,0228
4	УрФУ	urfu	urfu.ru	126	81777	0,0264
5	ПетрГУ	petrsu	petrsu.ru	53	87964	0,0882
6	ПАО «Газпром»	gazprom	www.gazprom.ru	80	1278255	0,0297
7	ПАО «Северсталь»	severstal	www.severstal.com	27	80028	0,0318
8	ПАО «НК «Роснефть»	rosneft	www.rosneft.ru	69	26719	0,0205
9	«Балтика»	baltika	www.baltika.ru	3	3647	0,0503
10	«ЕвразХолдинг»	evraz	www.evraz.com	10	280	0,043
11	Кунсткамера	kunstkamera	kunstkamera.ru	11	479	0,273
12	ИВТ СО РАН	ict.nsc	www.ict.nsc.ru	10	4673	0,0234
13	ИКИ РАН	iki.rssi	iki.rssi.ru	6	284	0,197
14	КарНЦ РАН	krc.karelia	www.krc.karelia.ru	42	25641	0,1029
15	РАН	ras	ras.ru	59	724	0,0405

Таблица 5

Характеристики веб-пространств, используемые в кластеризации

№ п/п	Организация	PR0/PR1	кол-во вершин / кол-во дуг	макс. клика / кол-во вершин	макс. КСС / кол-во вершин
1	spbu	5,7356	0,0015	0,0331	0,8145
2	msu	4,0318	0,0036	0,0137	0,6288
3	mipt	4,5390	0,0032	0,0353	0,7058
4	urfu	1,5400	0,0015	0,0317	0,9126
5	petrsu	5,2440	0,0006	0,0566	0,7547
6	gazprom	1,2042	0,0001	0,9125	0,9625
7	severstal	4,0847	0,0003	0,5929	0,9629
8	rosneft	1,2122	0,0025	0,6811	1,0000
9	baltika	1,0000	0,0008	1,0000	1,0000
10	evraz	1,0000	0,0357	0,5000	0,8000
11	kunstkamera	6,4019	0,0229	0,0000	0,7272
12	ict.nsc	2,2309	0,0021	0,3000	0,8000
13	iki.rssi	2,3831	0,0211	0,0000	0,8333
14	krc.karelia	1,5123	0,0016	0,0952	0,7380
15	ras	1,9546	0,0814	0,0508	0,4237

Таблица 6

Средние значения вторичных характеристик

№ п/п	Вторичные характеристики	cl1	cl2	cl3	cl4
1	PR0/PR1	2,1896	5,0062	1,5262	1,1041
2	кол-во верш / кол-во дуг	0,0349	0,0054	0,0016	0,0098
3	макс. клика / кол-во верш	0,1169	0,1219	0,0635	0,7734
4	макс. КСС / кол-во верш	0,6857	0,7657	0,8253	0,9406

Наиболее характерным является кластер cl4, содержащий производственные организации, с которого начнем анализ. Он выделяется сильной связностью и малым количеством «висячих» вершин (не имеющих исходящих дуг), – об этом говорят характеристики 3 и 4. Значимости головного сайта внимание не уделяется (характеристика 1).

Элементы кластера научных учреждений cl1 обладают низкой плотностью дуг (характеристика 2), невысокой связностью (характеристики 3 и 4) и большим количеством «висячих» вершин (характеристика 4).

У элементов кластера cl2 (в основном это вузы) явно выделяется головной сайт (характеристика 1). Плотность и связность также достаточно высоки (характеристики 2 и 3).

Элементы «смешанного» кластера cl3 (кластер, содержащий вуз и научное учреждение) имеют очень высокую плотность и очень слабую максимальную клику.

Понятно, что столь малое количество экспериментов не позволяет делать какие-либо глобальные выводы, однако дает возможность определить дальнейшие направления исследований.

Работа выполнена при частичной поддержке гранта РФФИ 15-01-06105А, проект «Разработка вебометрических и эргономических моделей и методов анализа эффективности присутствия в Вебе информационных веб-пространств крупных организаций».

Список литературы

1. Печников А.А. Применение вебометрических методов для исследования информационного веб-пространства научной организации (на примере Карельского научного центра РАН) // Труды Карельского научного центра Российской академии наук. Серия «Математическое моделирование и информационные технологии». – 2013. – № 1. – С. 86–95.

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

3. D. Easley and J. Kleinberg Networks, Crowds, and Markets: Reasoning about a Highly Connected World / Cambridge University Press. – 2010. – 744 p.

4. Web crawler [Электронный ресурс]. – URL: https://en.wikipedia.org/wiki/Web_crawler#Open-source_crawlers.

5. Status codes in HTTP [Электронный ресурс]. – URL: https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html.

6. Intellij IDEA the Java IDE – JetBrains [Электронный ресурс]. – URL: https://www.jetbrains.com/idea/.

7. Jsoup Java HTML Parser 1.10.2 API [Электронный ресурс]. – URL: https://jsoup.org/apidocs/org/jsoup/nodes/Document.html.

8. JGraph mxgraph [Электронный ресурс]. – URL: https://github.com/jgraph/mxgraph.

9. Харари Ф. Теория графов. – М.: Мир, 1973. – 301 с.

10. Кристофидес Н. Теория графов. Алгоритмический подход. – М.: Мир, 1978. – 429 с.

11. Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine // Computer Networks and ISDN Systems. – 1998. – № 30. – P. 107–117.

12. Халафян А.А. STATISTICA 6. Статистический анализ данных / Бином-Пресс. – 2007. – 512 с.