УДК 004.738.5

ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ ВЕБ-ГРАФА ИНФОРМАЦИОННОГО ВЕБ-ПРОСТРАНСТВА САНКТ-ПЕТЕРБУРГСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Печников А.А.

ФГБУН «Институт прикладных математических исследований» Карельского научного центра Российской академии наук, Петрозаводск, e-mail: pechnikov@krc.karelia.ru

Веб-пространство организации (предприятия, учреждения) – это множество веб-сайтов организации, связанных посредством гиперссылок. Как правило, в таком множестве выделяется так называемый «головной сайт» (официальный сайт организации), сайты подразделений, проектов, различных мероприятий, форумы, вики-ресурсы. Веб-граф – это граф, вершинами которого служат веб-сайты, а ребра соединяют те вершины, между которыми имеются гиперссылки. В том случае, когда известны веб-сайты, составляющие веб-пространство вуза, и связывающие их гиперссылки, можно построить веб-граф как модель веб-пространства для исследования основных теоретико-графовых свойств с целью их содержательной интерпретации и выработки управленческих решений, направленных на улучшение характеристик присутствия в Вебе. В статье излагаются основные подходы и методы построения и исследования веб-пространства крупного вуза на примере Санкт-Петербургского государственного университета. Показано, что наиболее значимыми сайтами веб-пространства (кроме официального сайта вуза) являются сайты, относящиеся к так называемым веб-коммуникаторам. Методы исследования веб-ресурсов СПбГУ и полученные результаты могут иметь универсальное значение и быть полезными не только для СПбГУ, но и для других вузов России.

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

CONSTRUCTION AND STUDY OF THE WEB GRAPH OF THE INFORMATION WEB SPACE OF ST. PETERSBURG UNIVERSITY

Pechnikov A.A.

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences, Petrozavodsk, e-mail: pechnikov@krc.karelia.ru

Web space of the organization (company, institution) is the set of the organization's web sites connected by hyperlinks. As a rule, in this set stands out the so-called «parent website» (official website of the organization), sites of units and departments, projects, events, forums, wikis. The web graph is a graph whose vertices are the web sites, and edges connect the vertices between which there are hyperlinks. In the case where the web sites that make up the web space of the university and connecting them hyperlinks are known , you can build a web graph as a model of web space for the study of basic graph-theoretic properties with a view to meaningful interpretation and development of administrative decisions aimed at improving the characteristics of presence on the Web. The article presents the main approaches and methods of construction and research of web space a large institution on the example of Saint-Petersburg state University. It is shown that the most significant sites of the web space (in addition to the official website of the University) are sites that are related to the so-called web communicators. Research methods resources on the web Saint-Petersburg state University and the results can have a universal value and be useful not only for this University, but also for other Russian universities.

Keywords: website, hyperlink, web space, web graph

Веб-пространство организации (предприятия, учреждения) – это множество веб-сайтов организации, связанных посредством гиперссылок. Как правило, в таком множестве выделяется так называемый «головной сайт» (официальный сайт организации), сайты подразделений, проектов, различных мероприятий, форумы, вики-ресурсы. Веб-ресурсы вузов относятся к так называемым регламентируемым (администрируемым) веб-ресурсам [2].

Следуя работе [5] определим веб-граф как граф, вершинами которого служат веб-сайты, а ребра соединяют те вершины, между которыми имеются гиперссылки. Под гиперссылками в данном случае мы понимаем не все ссылки между сайтами. На различных страницах одного сайта могут встречаться гиперссылки на один и тот же внешний адрес, имеющие одинаковый контекст (в частном случае – анкор), и количество таких «одинаковых» гиперссылок может быть равно количеству страниц на сайте (например – ссылка на сайт вышестоящей организации). Из такого множества гиперссылок, имеющих одинаковый адрес-приёмник и контекст, сделанных с данного сайта, в нашем исследовании мы рассматриваем только одну – ту, которая находится на странице, имеющей максимальный уровень (наивысшим считается уровень начальной страницы сайта). Веб-граф является моделью веб-пространства для исследования его основных теоретико-графовых свойств

с целью их содержательной интерпретации и выработки управленческих решений, направленных на улучшение характеристик присутствия в Вебе.

В статье излагаются основные подходы и методы построения и исследования веб-пространства крупного вуза на примере Санкт-Петербургского государственного университета (СПбГУ). К более ранним работам, в которых исследуются веб-ресурсы СПбГУ, относится, например, работа [6]. Для СПбГУ формирование единого информационного пространства университета и развитие его веб-пространства является одной из приоритетных задач [4]. Отсюда следует, что методы исследования веб-ресурсов СПбГУ и полученные результаты могут иметь универсальное значение и быть полезными не только для СПбГУ, но и для других вузов России.

Для сбора и анализа гиперссылок с целью построения веб-графа, а также для исследования веб-графа использовались следующие вебометрические инструменты:

– программа для поиска и сбора внешних гиперссылок BeeCrawler [8],

– база данных внешних гиперссылок (БДВГ, сайт http://grid.krc.karelia.ru/webometrics2) [1],

– открытая платформа для визуализации графов Gephi (см. https://gephi.github.io).

Построение веб-графа информационного веб-пространства университета

В работе был применен следующий подход к построению множества веб-сайтов информационного пространства университета. Вначале сканируется головной сайт вуза и все полученные в результате сканирования внешние гиперссылки после соответствующей обработки, а именно, – очистки от ошибок и нормализации (см. https://ru.wikipedia.org/wiki/Нормализация_URL), – вносятся в БДВГ. Головному сайту присваивается номер уровня, равный 0.

Далее с использованием БДВГ проводится анализ внешних гиперссылок, сделанных с головного сайта (в данном случае это официальный сайт СПбГУ spbu.ru), на предмет поиска:

а) доменных имен, аффилированных с доменом головного сайта (для spbu.ru, например, it.spbu.ru);

б) доменных имен, не аффилированных с доменом головного сайта, но содержащих вхождения подстрок домена головного сайта (например, для spbu.ru это подстроки «spbu» или ‘spb’ и найденный сайт www.spbumag.nw.ru) с дальнейшей содержательной проверкой сайтов;

в) доменных имен, не аффилированных с доменом головного сайта и не содержащих вхождения соответствующих подстрок, на которые имеются ссылки с головного сайта, и обнаруживаемые прямым просмотром анкоров гиперссылок (например, igor-krylov.ru, на который имеется гиперссылка с анкором «личная страница преподавателя»);

В нашем случае для СПбГУ мы получили множество из 97 доменных имен сайтов, которым присвоен номер уровня, равный 1. Теперь с использованием БДВГ проводится анализ внешних гиперссылок, сделанных с сайтов уровня 1, по тем же признакам (а)–(в). Для СПбГУ получаем множество из 162 доменных имен сайтов уровня 2. Аналогично строится множество из 25 доменных имен сайтов уровня 3 и множество из одного доменного имени сайта уровня 4. На уровне 5 сайты отсутствуют. Полученное множество доменных имен веб-сайтов принимается в качестве множества вершин, а соединяющие эти вершины гиперссылки принимаются в качестве дуг веб-графа. Построенный веб-граф схематично изображен на рис. 1.

Веб-граф университета представляет собой $G = G(V, E)$, где V – множество вершин, а E – множество дуг, то есть пар вершин $(v_i, v_j) \in V$. По построению граф $G = G(V, E)$ является ориентированным графом с кратными ребрами без петель. Отметим, что множество вершин V является объединением непересекающихся подмножеств вершин $V = V_0 \cup V_1 \cup V_2 \cup \ldots \cup V_k$, где любое подмножество V_i является множеством вершин, имеющих уровень i ($i = 0...k$). Реальный веб-граф информационного веб-пространства СПбГУ содержит 286 вершин и 9729 дуг. Часть сайтов веб-пространства СПбГУ, соответствующих вершинам веб-графа, приведены в табл. 1. В табл. 2 приведены примеры гиперссылок, которым соответствуют дуги веб-графа.

Укрупненное поуровневое изображение веб-графа СПбГУ приведено на рис. 2.

Геометрическими фигурами изображены подмножества вершин V_i ($i = 0...4$), а линиями со стрелками – подмножества дуг из E, соединяющих подмножества соответствующих уровней. К примеру, подмножество вершин 1-го уровня состоит из 97 вершин, на которые имеется в сумме 689 дуг с головного сайта (что эквивалентно дугам с V_0), 923 дуги с V_3 и 31 дуга с V_3. В свою очередь, с V_1 имеется 1941 дуга на V_0, 573 дуги на V_2 и 4731 дуга, связывающих вершины самого подмножества V_1, что обозначено на рисунке петлёй.

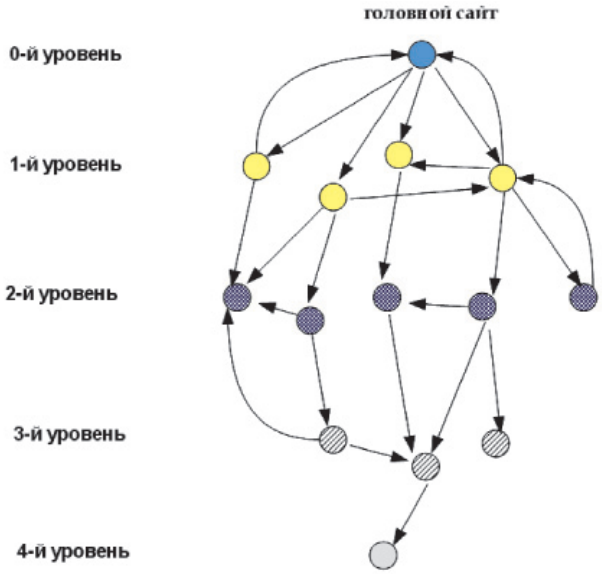

Рис. 1. Схематичное изображение веб-графа

Таблица 1

Сайты информационного веб-пространства СПбГУ

Уровень	Название сайта	Доменное имя
0	Санкт-Петербургский государственный университет	spbu.ru
1	Информационно-образовательный портал	1-line.spbu.ru
1	Биолого-почвенный факультет СПбГУ	bio.spbu.ru
2	Второе высшее филологическое образование	2philology.spbu.ru
2	Кафедра эстетики и философии культуры СПбГУ	aesthetics.philosophy.spbu.ru
3	Филиал № 2 «Полилог» курсов иностранных языков	6linya.spbu.ru
3	Школьная астрономия Петербурга	almucantarat.astro.spbu.ru
4	Проект «Контрреформация и схоластика»	creform.spbu.ru

Таблица 2

Гиперссылки информационного веб-пространства СПбГУ

Доменное имя-источник	Доменное имя-приемник	Анкор
spbu.ru	artesliberales.spbu.ru	Свободные искусства и науки
artesliberales.spbu.ru	abiturient.spbu.ru	Информационный центр Приемной комиссии
philosophy.spbu.ru	aesthetics.philosophy.spbu.ru	Здесь
testing.spbu.ru	ege.testing.spbu.ru	Система дистанционного тестирования

Исследование веб-графа с использованием стандартных возможностей Gephi

Открытая платформа для визуализации графов Gephi содержит ряд стандартных возможностей, удобных для исследования веб-графа, таких как нахождение компонент связности [9] и оценок значимости вершин по алгоритму ссылочного ранжирования PageRank [7].

Максимальная компонента сильной связности (КСС) веб-графа СПбГУ содержит 220 вершин, что свидетельствует о сильной связности веб-пространства СПбГУ. В максимальную КСС входит головной сайт, 86 из 97 сайтов 1-го уровня, 120 из 162 сайтов 2-го уровня и 13 из 25 сайтов 3-го уровня. Обнаруживаются еще две маленьких КСС, содержащих по 2 вершины.

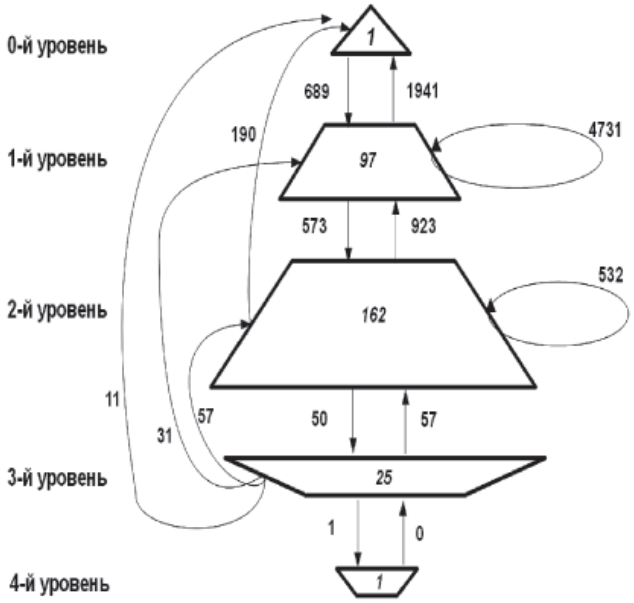

Рис. 2. Поуровневое укрупненное изображение веб-графа СПбГУ

Таблица 3

Первые тридцать сайтов с наибольшими значениями PageRank

№ п/п	Название сайта	Доменное имя	PageRank	Тип
1	2	3	4	5
1	Санкт-Петербургский государственный университет	spbu.ru	0,1007	посредник
2	Saint Petersburg State University (eng)	eng.spbu.ru	0,0280	посредник
3	Виртуальная приемная СПбГУ	guestbook.spbu.ru	0,0244	посредник
4	Приемная комиссия	abiturient.spbu.ru	0,0208	посредник
5	Физический факультет СПбГУ	www.phys.spbu.ru	0,0200	посредник
6	Управление научных исследований СПбГУ	csr.spbu.ru	0,0178	посредник
7	Журнал «Санкт-Петербургский университет»	journal.spbu.ru	0,0176	посредник
8	Юридический факультет СПбГУ	law.spbu.ru	0,0172	посредник
9	Научная библиотека им. М. Горького СПбГУ	library.spbu.ru	0,0166	посредник
10	Управление по работе с молодежью СПбГУ	students.spbu.ru	0,0147	посредник
11	Международная деятельность	ifea.spbu.ru	0,0124	посредник
12	Институт философии СПбГУ	philosophy.spbu.ru	0,0120	посредник
13	Факультет психологии СПбГУ	psy.spbu.ru	0,0113	посредник
14	Математико-механический факультет СПбГУ	www.math.spbu.ru	0,0112	посредник
15	Электронные журналы	cufts.library.spbu.ru	0,0111	коллектор
16	Биолого-почвенный факультет СПбГУ	bio.spbu.ru	0,0108	посредник
17	Учебная деятельность СПбГУ	edu.spbu.ru	0,0106	–
18	Общеуниверситетская кафедра физической культуры и спорта СПбГУ	sport.spbu.ru	0,0104	–
19	Студгородок СПбГУ	campus.spbu.ru	0,0098	посредник
20	Высшая школа менеджмента СПбГУ	www.gsom.spbu.ru	0,0097	индуктор
21	Институт химии СПбГУ	chem.spbu.ru	0,0080	посредник
22	Студсовет СПбГУ	studsovet.spbu.ru	0,0078	посредник
23	Институт истории СПбГУ	history.spbu.ru	0,0077	посредник
24	Геологический факультет СПбГУ	geology.spbu.ru	0,0076	индуктор

Окончание табл. 3

1	2	3	4	5
25	Расписание СПбГУ	timetable.spbu.ru	0,0076	–
26	Управление-служба информационных технологий	it.spbu.ru	0,0074	–
27	Институт «Высшая школа журналистики и массовых коммуникаций» СПбГУ	jf.spbu.ru	0,0072	индуктор
28	Восточный факультет СПбГУ	orient.spbu.ru	0,0069	индуктор
29	Эндаумент-фонд СПбГУ	fund.spbu.ru	0,0068	индуктор
30	Научный парк СПбГУ	researchpark.spbu.ru	0,0067	индуктор

Первые тридцать сайтов с наибольшим значением PageRank, полученных с помощью Gephi, приведены в табл. 3 (последняя колонка «тип» понадобится в дальнейшем изложении и будет объяснена ниже).

Достаточно ожидаемо наивысшее значение PageRank с большим отрывом имеет официальный сайт СПбГУ. На втором месте находится англоязычная версия официального сайта, далее следуют два сайта приемной комиссии СПбГУ. Из 22 учебно-научных структурных подразделений СПбГУ (институты и факультеты) половина имеет сайты в первой тридцатке.

Исследование веб-графа с использованием концептуальной модели фрагмента Веба

Концептуальная модель фрагмента Веба описана, например, в работе [3]. Одной из важных составляющих концептуальной модели является множество так называемых сайтов-коммуникаторов. Неформально сайт-коммуникатор – это сайт, который имеет входящие ссылки с «достаточно большого» количества сайтов некоторого целевого множества T и/или исходящие ссылки на «достаточно большое» количество сайтов из T. В нашем случае целевое множество T – это множество сайтов информационного веб-пространства университета.

Определим две функции: *insiteco<<pechnik01.wmf>>unt*(A,s) задает количество сайтов из некоторого множества *A*, имеющих гиперссылки на заданный сайт *s*, а *outsitecount*(*s*, *A*) – количество сайтов из *A*, на которые существуют гиперссылки с сайта *s*. Обозначим нижнее и верхнее пороговые значения, как λ и μ; λ, μ – целые и $\lambda \le \mu$.

В качестве нижнего порогового значения по аналогии с [3] принимается

$$\lambda = round\left(\frac{\sum_{t\in T} insitecount\ (T,t)}{|T|}\right).$$

В некотором смысле λ характеризует «среднюю степень» интереса к сайту целевого множества, проявляемую со стороны его сайтов-коллег.

По аналогии с [3] определим верхнее пороговое значение следующим образом:

$$\mu = round\left(\frac{\sum_{insitecount(T,u)\ge\lambda} insitecount\ (T,u)}{\left|\{u \mid insitecount\ (T,u)\ge\lambda\ \}\right|}\right).$$

Сайтом-посредником называется сайт *u*, для которого выполняется условие

$$insitecount(T, u) \ge \mu\ \&\ outsitecount(u, T) \ge \lambda.$$

Сайтом-коллектором называется сайт *u*, для которого выполняется условие

$$insitecount(T, u) \ge \mu\ \&\ \lambda > outsitecount(u, T) \ge 1.$$

Сайтом-индуктором называется сайт *u*, для которого выполняется условие

$$\mu > insitecount(T, u) \ge \lambda\ \&\ outsitecount(u, T) \ge \lambda.$$

Для веб-пространства СПбГУ были получены значения $\lambda = 6$ и $\mu = 19$ и множества следующей мощности: 19 посредников, 2 коллектора и 27 индукторов. Все сайты-коммуникаторы находятся на 0-м и 1-м уровнях.

Наибольший интерес в структуре веб-графа представляют сайты-посредники. В последней колонке табл. 3 «тип» указан тип для каждого сайта из первых тридцати по PageRank. Можно увидеть, что в первую тридцатку попали все 18 сайтов-коммуникаторов. Кроме того, сюда же попали один из двух коллекторов и шесть из двадцати семи индукторов (причем индукторы находятся в конце таблицы).

Заключение

В статье описана методика построения веб-графа как модели информационного веб-пространства крупного университета и два способа определения наиболее значимых вершин в веб-графе. Первый способ является стандартным вычислением PageRank

для вершин построенного веб-графа. Второй способ основан на предложенном ранее автором способе определения сайтов-коммуникаторов концептуальной модели фрагмента Веба. Предложенная методика была использована для моделирования веб-пространства на примере СПбГУ. Показано, что наибольшей значимостью в смысле PageRank в нем обладают вершины, соответствующие сайтам-коммуникаторам.

Работа выполнена при 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. Головин А.С., Печников А.А. 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.

2. Печников А.А. Модель университетского Веба // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2010. – № 6. – С. 208–214.

3. Печников А.А. Методы исследования регламентируемых тематических фрагментов Web // Труды Института системного анализа Российской академии наук. Серия: Прикладные проблемы управления макросистемами. – 2010. – Т. 59. – С. 134–145.

4. Программа развития федерального государственного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» до 2020 года (в ред. распоряжения Правительства РФ от 26.06.2014 № 1156-р). http://spbu.ru/files/upload/2014.06.26_1156-p.pdf.

5. Райгородский А.М. Модели случайных графов и их применения // Труды МФТИ. – 2010. – Т. 2, № 4. – С. 130–140.

6. Blekanov I.S., Sergeev S.L., Maksimov A.I. Analysis of the topology of large Web segments using Broder's bow-tie model // Life Science Journal. – Т. 11. – № 6 Spec. – Iss. 2014. – P. 258–261.

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

8. Pechnikov A.A., Chernobrovkin D.I. Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. – 2014. – Vol. 75, № 3. – P. 587–593.

9. Tarjan R. E. Depth-first search and linear graph algorithms // SIAM Journal on Computing. – 1972. – Vol. 1, № 2. – P. 146–160. – DOI:10.1137/0201010.

References

1. Golovin A.S., Pechnikov A.A. Baza dannyh vneshnih giperssylok dlja issledovanija fragmentov Weba // Informacionnaja sreda vuza XXI veka: materially VII Vserossiiskoi nauchno-prakticheskoi konferencii (23–25 sentjabrja 2013 g.). Petrozavodsk, 2013. pp. 55–57.

2. Pechnikov A.A. Model' universitetskogo Weba // Vestnik Nizheegorodskogo universitete im. N.I. Lobachevskogo [Vestnik of Lobachevsky State University of Nizhni Novgorod], 2010. no. 6. pp. 208–214.

3. Pechnikov A.A. Metody issledovanija reglamentiruemyh tematicheskih fragmentov Web // Trudy Instituta sistemnogo analiza Rossiiskoi akademii nauk. Serija: Prikladnye problemy upravlenija makrosistemami. Tom 59. 2010. pp. 134–145.

4. Programma razvitija federal'nogo gosudarstvennogo obrazovatelnogo uchrezhdenija vysshego professionalnogo obrazovanija "Sankt-Peterburgskii gosudarstvennyi universitet do 2020 goda (v red. rasporjajenija Pravitelstva RF ot от 26.06.2014 no. 1156-p). http://spbu.ru/files/upload/2014.06.26_1156-p.pdf.

5. Raigorodskii A.M. Modeli sluchainyh grafov i ih primenenija // Trudy MFTI. 2010. Tom 2, no. 4. pp. 130–140.

6. Blekanov I.S., Sergeev S.L., Maksimov A.I. Analysis of the topology of large Web segments using Broder's bow-tie model // Life Science Journal. T. 11. no. 6 Spec. Iss. 2014. pp. 258–261.

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

8. Pechnikov A.A., Chernobrovkin D.I. Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. 2014, Vol. 75, no. 3. pp. 587–593.

9. Tarjan R. E. Depth-first search and linear graph algorithms // SIAM Journal on Computing, 1972, Vol. 1, no. 2. pp. 146–160. – DOI:10.1137/0201010.

Рецензенты:

Кириллов А.Н., д.ф.-м.н., доцент, ведущий научный сотрудник лаборатории информационных компьютерных технологий, Институт прикладных математических исследований Карельского научного центра Российской академии наук, г. Петрозаводск;

Рогов А.А., д.т.н., профессор, заведующий кафедрой «Теория вероятностей и анализ данных», ФГБОУ ВПО «Петрозаводский государственный университет», г. Петрозаводск.