

ОЛЬГА ЮРЬЕВНА ДЕРБЕНЕВА

начальник управления по инновационно-производственной деятельности, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
oder@petsu.ru

АЛЕКСЕЙ ГЕОРГИЕВИЧ МАРАХТАНОВ

заместитель директора Регионального центра новых информационных технологий, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
alikrpk@gmail.com

ОЛЬГА ЮРЬЕВНА НАСАДКИНА

кандидат технических наук, директор Регионального центра новых информационных технологий, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
onasad@petsu.ru

АНДРЕЙ АНАТОЛЬЕВИЧ ПЕЧНИКОВ

доктор технических наук, доцент, и. о. заведующего лабораторией телекоммуникационных систем, Институт прикладных математических исследований Карельского научного центра РАН (Петрозаводск, Российская Федерация)
pechnikov@krc.karelia.ru

К ВОПРОСУ О МОДУЛЯРНОСТИ ВЕБ-ПРОСТРАНСТВА ОБРАЗОВАТЕЛЬНОГО УЧРЕЖДЕНИЯ*

Исследование модулярности веб-пространства высшего учебного заведения на примере Петрозаводского государственного университета (ПетрГУ) позволяет получить результаты, описывающие укрупненную структуру веб-пространства вуза, и обнаружить самоорганизацию веб-пространства, происходящую без внешнего (административного) регулирования, что позволяет предложить рекомендации, направленные на улучшение связности веб-пространства, в том числе и с использованием административного воздействия.

Ключевые слова: веб-пространство, веб-сайт, гиперссылка, связность сайтов, веб-граф, модулярность

ВВЕДЕНИЕ

Веб-ресурсы крупных организаций, предприятий и учреждений относятся к регламентируемым веб-ресурсам [5]. Это означает, что существует (или должен существовать) официальный документ, в котором изложены цели и задачи веб-ресурса, основные структурные составляющие, правила изменения информации и т. д.

Под веб-сайтом понимается веб-ресурс (множество взаимосвязанных html-страниц и документов), идентифицируемый в Вебе посредством уникального доменного имени.

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.

Уточним используемое в статье понятие «гиперссылка». 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 с одинаковым адресом-приемником и контекстом, 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). Такая гиперссылка называется «уникальной внешней гиперссылкой» [5], и, поскольку далее рассматриваются только такие ссылки, мы будем называть их гиперссылками (или просто ссылками).

Математической моделью, успешно используемой для анализа веб-пространства организации, является веб-граф, построенный следующим образом: множество вершин соответствуют сайтам организации, а множество дуг – гиперссылкам, связывающим эти сайты. Понятно, что такой веб-граф является ориентированным графом без петель, имеющим кратные дуги.

Для исследования структуры веб-пространства используются как традиционные теоретико-гра-

фовые подходы, основанные на обнаружении компонент связности, кратчайших путей [5], так и подходы, развитые в последнее десятилетие применительно к Вебу. К ним относятся работы по изучению самоорганизации веб-сообществ [10] и исследования модулярности графа [13].

Неформально под веб-сообществом понимается некоторое подмножество вершин веб-графа, для которого количество дуг, связывающих вершины-участники веб-сообщества, больше, чем количество дуг, связывающих их с другими вершинами.

Модулярность, в свою очередь, это свойство графа и некоторого разбиения его на подграфы. Мера модулярности показывает, насколько данное разбиение качественно в том смысле, что существует много ребер, лежащих внутри подграфов, и мало ребер, лежащих вне подграфов (соединяющих подграфы между собой).

В данном случае речь идет о ребрах, а не о дугах, поскольку ориентированный граф достаточно тривиально преобразуется в неориентированный. На практике значения модулярности, лежащие в пределах от 0,3 до 0,7, говорят о том, что граф имеет структуру с различными подграфами с заданными свойствами. Для обозначения подграфов, на которые разбивается граф, в русскоязычной литературе используются различные термины, такие как «кластеры», «неявные сообщества», «модули»; далее мы будем использовать термин «модуль». Можно считать, что неориентированный граф с высоким значением меры модулярности является хорошим объектом для последующего выявления сообществ в его ориентированном прообразе.

Целью данной работы является исследование модулярности веб-пространства высшего учебного заведения (на примере ПетрГУ). Результаты, полученные в рамках исследования, позволяют описать структуру веб-пространства вуза, обнаружить (или, наоборот, не обнаружить) самоорганизацию веб-пространства, происходящую без внешнего (административного) регулирования, и предложить рекомендации, направленные на улучшение связности веб-пространства с использованием административного воздействия.

Полученные результаты и сделанные выводы, демонстрируемые на примере ПетрГУ, могут служить руководством к действию и легко переносимы на другие вузы, организации, предприятия и учреждения.

ВЕБ-ГРАФ ПЕТРГУ

Общее количество сайтов, составляющих веб-пространство ПетрГУ, в данном исследовании равно 147. Следуя работе [6], их можно разбить на 13 групп. Перечислим эти группы (в скобках в качестве примера указано по одному характерному представителю данной группы для ПетрГУ):

1. Официальный сайт университета (официальный сайт ПетрГУ – petrsu.ru);

2. Сайты факультетов (математический факультет – mf.petrsu.ru);

3. Сайты кафедр (кафедра математического анализа ПетрГУ – analysis.petrsu.ru);

4. Сайты научной библиотеки, ботанического сада, институтов, центров (Научная библиотека – library.petrsu.ru);

5. Сайты филиалов университета, университетских лицеев и т. д. (Кольский филиал ПетрГУ – www.arcticsu.ru);

6. Сайты издательства, научных журналов, медиа-ресурсов (журнал «Принципы экологии» – ecopri.ru);

7. Сайты структурных подразделений университета, не вошедшие в группы 2–6 (Региональный центр новых информационных технологий – rcnit.petrsu.ru);

8. Сайты научных конференций, организуемых университетом (конференция «Космос братьев Гримм» – grimms.petrsu.ru);

9. Сайты программ и проектов, выполняемых сотрудниками университета («Британцы в Карелии» – britons.karelia.ru);

10. Сайты учебных ресурсов, разработанные сотрудниками университета («Виртуальная педагогика» – praktika.karelia.ru);

11. Сайты информационно-справочных систем и ресурсов университета («Аспирантура ПетрГУ» – aspirant.petrsu.ru);

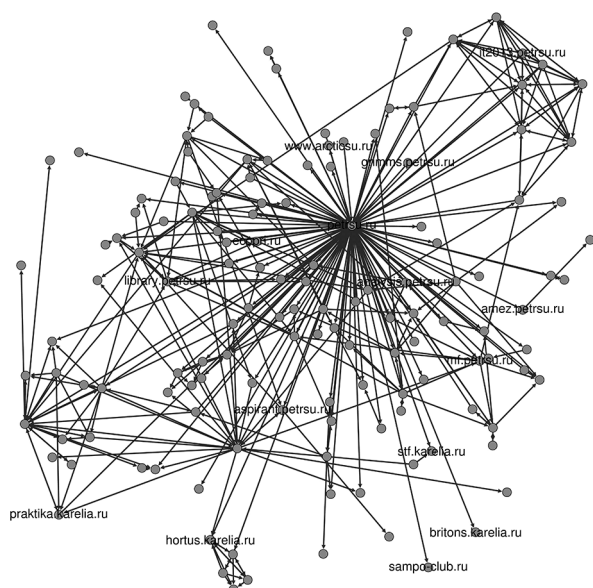
12. Персональные сайты сотрудников университета (сайт Андрея Мезенцева – amez.petrsu.ru);

13. Другие сайты: сайты творческих организаций, профкома и сайты, не вошедшие в группы 1–12 (Туристический клуб ПетрГУ «Сампо» – sampo-club.ru).

Сканирование сайтов веб-пространства ПетрГУ с целью сбора исходящих гиперссылок производилось программой BeeCrawler [14]. Для хранения, обработки и анализа гиперссылок использовалась специализированная база данных внешних гиперссылок [3]. На 147 сайтах веб-пространства ПетрГУ было отсканировано около 100 000 страниц и сформировано множество, содержащее 11 200 исходящих с этих сайтов гиперссылок.

Далее из 11 200 гиперссылок были отобраны 1352 гиперссылки, которые связывают сайты веб-пространства ПетрГУ, и построен веб-граф $G=G(V,E)$; здесь V (vertex) – множество вершин, соответствующих сайтам веб-пространства, E (edge) – множество дуг, соответствующих гиперссылкам, связывающим эти сайты, $|V|=147$, $|E|=1352$. Поскольку ряд сайтов связан гиперссылками в количестве большем, чем 1, то мы имеем $G(V,E)$ как ориентированный граф с кратными дугами без петель.

На рисунке приводится несколько упрощенное изображение веб-графа $G(V, E)$: во избежание загромождения рисунка кратные дуги не нарисованы, приведены названия только некоторых вершин и исключены 10 изолированных вершин. Головной сайт petrstu.ru представлен вершиной с наибольшей инцидентностью, расположенной почти в самом центре рисунка. Десять изолированных вершин соответствуют сайтам, которые не связаны гиперссылками с другими сайтами ПетрГУ. Отметим также, что 40 вершин являются «висячими», то есть имеют либо только исходящие, либо только входящие дуги, поскольку 26 сайтов не имеют входящих ссылок, а 14 – исходящих, связывающих их с другими сайтами ПетрГУ.



Веб-граф веб-пространства ПетрГУ

Относительно почти 36 % сайтов веб-пространства ПетрГУ сразу же можно сделать вывод о том, что они не являются участниками тематических сообществ [7], для которых наличие исходящих и входящих гиперссылок для каждого сайта в рамках сообщества является обязательным условием. Вместе с тем даже беглый взгляд на рисунок позволяет обнаружить, по крайней мере, две группы сайтов, имеющих очень высокую плотность связывающих их дуг.

ИССЛЕДОВАНИЕ МОДУЛЯРНОСТИ ВЕБ-ГРАФА ПЕТРГУ

Дадим достаточно строгое и вместе с тем не громоздкое определение функции модулярности, следуя [2], для чего введем следующие обозначения:

A – матрица инцидентности графа $G(V, E)$ [4],
 A_{ij} – количество дуг из вершины i в вершину j ,
 m – количество дуг в графе, $m = |E|$,
 $Pr()$ – вероятность некоторого события,

S – некоторое множество модулей, на которые разбит граф $G(V, E)$, s – обозначение одного из модулей $s \in S$. Тогда

$$Q = \frac{1}{2m} \sum_{s \in S} \sum_{i, j \in s} [A_{ij} - Pr(A_{ij} = 1)].$$

На первом этапе исследования рассматривается граф $G^1(V, E^1)$, который отличается от $G(V, E)$ тем, что в нем отсутствуют кратные ребра; тогда в нашем случае $|E^1| = 419$.

Для анализа веб-графов была использована открытая программная платформа Gephi [11]. В Gephi для построения модульной структуры графа используется эвристический алгоритм, предложенный в [9]. Количество модулей $K = |S|$, на которые разбивается граф, автоматически определяется алгоритмом из соображений максимального значения функции модулярности Q .

В случае графа $G^1(V, E^1)$ максимальное значение $Q = 0,486$ получено для $K = 9$.

Для пяти из девяти модулей достаточно четко отслеживается тематика сайтов.

В одном случае это 5 сайтов, посвященных теме ботанических садов (электронный журнал, ботанический сад ПетрГУ, ИПС «Ботанические коллекции» и др.). В другом случае тоже 5 сайтов, но уже филологической тематики (словарь-конкорданс публицистики Ф. М. Достоевского, проект «Digital Humanities», проект «Филолог.ру» и др.).

Еще один модуль из 9 вершин соответствует сайтам конференций по информационно-коммуникационным технологиям (серия из сайтов 8 конференций, посвященных информационной среде вуза, и сайт Всероссийской ассоциации вузов – пользователей Oracle).

Модуль из 5 вершин содержит вершины, сайты которых аффилированы с кафедрой информационно-измерительных систем и физической электроники ПетрГУ (собственно сайт кафедры, сайт научно-образовательного центра по фундаментальным проблемам приложений физики низкотемпературной плазмы и ряд других сайтов).

И, наконец, достаточно крупный модуль, содержащий 18 вершин, также можно объединить в рамках единой темы: это сайты проектов, электронных журналов и сайты разработчиков этих сайтов.

Для остальных модулей говорить о единой тематике не представляется возможным.

На втором этапе был исследован веб-граф с кратными дугами $G(V, E)$. Здесь максимальное значение $Q = 0,537$ получено для $K = 11$. В этом случае появляется возможность тематически идентифицировать 9 модулей. Пять из них практически совпадают с тематическими модулями, выявленными на первом этапе. В то же время обнаруживается, например, крупный модуль из 14 вершин, соответствующих сайтам Карель-

ской государственной педагогической академии (КГПА), вошедшей в состав ПетрГУ в 2013 году.

ОБ ИСПОЛЬЗОВАНИИ РЕЗУЛЬТАТОВ

Развитие веб-пространства для университета является одной из приоритетных задач. Именно состояние этого пространства в значительной степени характеризует вуз в сети Интернет, влияет на позиции сайтов вуза в поисковых системах и рейтингах, что, безусловно, является фактором, влияющим на конкурентоспособность вуза в целом.

Результаты приведенных в работе исследований имеют большую практическую ценность, поскольку служат основой для принятия различных решений по развитию веб-пространства университета, в том числе технических и управленческих.

Построенный граф наглядно показывает, что, несмотря на наличие значительного числа информационных ресурсов, развиваемых и поддерживаемых в вузе, далеко не все из них имеют гипертекстовые связи с другими сайтами веб-пространства университета, в том числе с сайтами схожих тематик. Учитывая, что такие параметры, как число внешних и внутренних гиперссылок, учитываются и поисковыми системами при ранжировании в результатах поиска [1], и системами рейтингования вузов (такими, как Webometrics [12]), очевидно, что повышение уровня связности ресурсов, составляющих веб-пространство организации, может привести к росту позиций сайтов вуза в результатах поиска и рейтингах. В то же время стоит отметить, что алгоритмы поисковых систем учитывают не только наличие ссылки, но и тематическую близость ресурсов, связанных гиперссылкой, повышение уровня связности должно осуществляться не хаотично или случайным образом,

а в рамках выделенных тематических модулей или кластеров.

Стоит отметить, что наличие ссылок – важный, но не единственный параметр, влияющий на позиции сайтов веб-пространства. Современные поисковые системы учитывают более 800 факторов, а также в ряде случаев понижают степень влияния ссылок на общий результат [8]. Большой вес приобретают другие параметры, такие как поведенческие факторы, уникальность текстовой информации и пр., которые также нужно учитывать при создании и развитии университетских интернет-ресурсов.

С организационной точки зрения развитию веб-пространства вуза будет способствовать принятие регламентов и рекомендаций, определяющих принципы функционирования и развития сайтов, составляющих веб-пространство университета, например разработка правил размещения гиперссылок на ресурсах. Данные решения должны быть подкреплены выводами, сделанными на базе приведенного в настоящей статье исследования. Кроме того, имеет смысл для ряда ресурсов, выделенных в рамках данного исследования (например, для ресурсов, не имеющих связей с основным сайтом), внести точечные правки в содержимое с целью повышения уровней связности в тематических сегментах (например, с целью интеграции тематических ресурсов КГПА в веб-пространство ПетрГУ).

Важно, чтобы обозначенная работа не проводилась формально, ради простого увеличения числовых показателей, а приводила к действительному улучшению качества ресурсов, составляющих пространство, повышению удобства пользователей, работающих с веб-сайтами вуза, улучшению полноты и достоверности представленной на сайтах информации.

* Работа выполнена при поддержке Программы стратегического развития ПетрГУ на 2012–2016 гг.

СПИСОК ЛИТЕРАТУРЫ

1. Ашманов И., Иванов А. Оптимизация и продвижение сайтов в поисковых системах. СПб.: Питер, 2013. 465 с.
2. Бузун Н., Коршунов А. Выявление пересекающихся сообществ в социальных сетях // Доклады всероссийской научно-практической конференции «Анализ Изображений, Сетей и Текстов» (Екатеринбург, 16–18 марта 2012 г.). М.: Национальный Открытый Университет «ИНТУИТ», 2012. С. 87–96.
3. Головин А. С., Печников А. А. База данных внешних гиперссылок для исследования фрагментов Веба // Информационная среда вуза XXI века: Материалы VII Всероссийской научно-практ. конф. (23–27 сентября 2013 г.). Петрозаводск, 2013. С. 55–57.
4. Зыков А. А. Основы теории графов. М.: Наука, 1987. 383 с.
5. Печников А. А. Модель университетского Веба // Вестник Нижегородского университета им. Н. И. Лобачевского. 2010. № 6. С. 208–214.
6. Печников А. А., Дербенева О. Ю. Анализ гиперссылок веб-пространства Петрозаводского государственного университета // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2013. № 2 (131). С. 100–106.
7. Печников А. А. Методы исследования регламентируемых тематических фрагментов Web // Труды Института системного анализа Российской академии наук. Сер. «Прикладные проблемы управления макросистемами». 2010. Т. 59. С. 134–145.
8. Ссылочные факторы в формуле ранжирования Яндекса: 3 месяца спустя после «отключения» [Электронный ресурс]. Режим доступа: <http://www.computerra.ru/100955/ssylochnyye-faktoryi-v-formule-ranzhirovaniya-yandeksa-3-mesyatsa-spustya-posle-otklyucheniya>
9. Blondel V. D., Guillaume J.-L., Lambiotte R. E. Lefebvre Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment, 1742–5468, P10008, 2008. DOI: 10.1088/1742-5468/2008/10/P10008.

10. Flake G. W., Lawrence S. R., Giles C. L., Coetzee F. M. Self-Organization and Identification of Web Communities // IEEE Computer. 2002. 35(3). P. 66–71.
11. Gephi – The Open Graph Viz Platform. URL: <https://gephi.org>
12. Methodology. Ranking Web of Universities. URL: <http://www.webometrics.info/en/Methodology>
13. Newman M. E. J., Girvan M. Finding and evaluating community structure in networks // Physical Review E: Statistical, Nonlinear, and Soft Matter Physics. Vol. 69. № 2. Article ID 026113. 2004.
14. 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.

Derbeneva O. Yu., Petrozavodsk State University (Petrozavodsk, Russian Federation)

Marakhtanov A. G., Petrozavodsk State University (Petrozavodsk, Russian Federation)

Nasadkina O. Yu., Petrozavodsk State University (Petrozavodsk, Russian Federation)

Pechnikov A. A., Institute of Applied Mathematical Research, Karelian Research Centre of RAS
(Petrozavodsk, Russian Federation)

REVISITING THE WEB SPACE MODULARITY OF AN EDUCATIONAL INSTITUTION

The study of the web space modularity of a higher educational institution based on the example of the Petrozavodsk State University yields results that describe the extended structure of the institution's web space, and allows the self-structured (not regulated by an outside authority) pattern of web space to be discovered. This, in turn, enables recommendations on how to improve the connectedness of the web space using, among other methods, an external administrative authority.

Key words: web space, web site, hyperlink, connectivity of sites, web graph, modularity

REFERENCES

1. Ashmanov I., Ivanov I. *Optimizatsiya i prodvizhenie saytov v poiskovykh sistemakh* [Optimization and promotion of sites in search engines]. St. Petersburg, Piter Publ., 2013. 465 p.
2. Buzun N., Korshunov A. Identifying overlapping communities in social networks [Vyyavlenie peresekeyuscchikh-sya soobschestv v sotsial'nykh setyakh]. *Doklady vserossiyskoy nauchno-prakticheskoy konferentsii "Analiz izobrazheniy, Setey i Tekstov"* [Repots of Conference "Analysis of Images, Social Networks, and Texts", April, 10–12th, Yekaterinburg]. Moscow, 2012. P. 87–96.
3. Golovin A. S., Pechnikov A. A. Database of external hyperlinks for Web fragments investigation [Baza dannykh vneshnikh giperssylok dlya issledovaniya fragmentov Web]. *Informatsionnaya sreda vuza XXI veka: Materialy VII Vserossiyskoy nauchno-prakticheskoy konferentsii (23–27 sentyabrya 2013 g.)* [University Infomedia XXI Century: Proceedings of VII All-Russian Scientific and Practical Conference (September 23–27, 2013)]. Petrozavodsk, 2013 P. 55–57.
4. Zykov A. A. *Osnovy teorii grafov* [Basics of the graph theory]. Moscow, Nauka Publ., 1987. 383 p.
5. Pechnikov A. A. Model of a university Web [Model' universitetskogo Web]. *Vestnik Nizhegorodskogo universiteta im. N. I. Lobachevskogo* [Vestnik of the Lobachevsky State University of Nizhni Novgorod]. 2010. № 6. P. 208–214.
6. Pechnikov A. A., Derbeneva O. Yu. Hyperlinks analysis of the web space of the Petrozavodsk State University [Analiz giperssylok web-prostranstva Petrozavodskogo gosudarstvennogo universiteta]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural and Engineering Sciences]. 2013. № 2 (131). P. 100–106.
7. Pechnikov A. A. Methods of investigation of regulated thematic fragments Web [Metody issledovaniya reglamentiruemyykh tematicheskikh fragmentov Web]. *Trudy Instituta systemnogo analiza Rossiyskoy akademii nauk. Ser. "Prikladnye problemy upravleniya makrosistemami"* [Proceedings of the Institute of Systems Analysis Russian Academy of Sciences. Series: Applied problems of control macro systems]. 2010. Vol. 59. P. 134–145.
8. *Ssylochnye faktory v formule ranzhirovaniya Yandeksa: 3 mesyatsa spustya posle "otklyucheniya"* [Reference factors in ranking formula Yandex: 3 months after the "off"]. URL: <http://www.computerra.ru/100955/ssylochnnye-factoryi-v-formule-ranzhirovaniya-yandeksa-3-mesyatsa-spustya-posle-otklyucheniya>
9. Blondel V. D., Guillaume J.-L., Lambiotte R. E. Lefebvre Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment, 1742–5468, P10008, 2008. DOI: 10.1088/1742-5468/2008/10/P10008.
10. Flake G. W., Lawrence S. R., Giles C. L., Coetzee F. M. Self-Organization and Identification of Web Communities // IEEE Computer. 2002. 35(3). P. 66–71.
11. Gephi – The Open Graph Viz Platform. URL: <https://gephi.org>
12. Methodology. Ranking Web of Universities. URL: <http://www.webometrics.info/en/Methodology>
13. Newman M. E. J., Girvan M. Finding and evaluating community structure in networks // Physical Review E: Statistical, Nonlinear, and Soft Matter Physics. Vol. 69. № 2. Article ID 026113. 2004.
14. 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.

Поступила в редакцию 02.09.2014